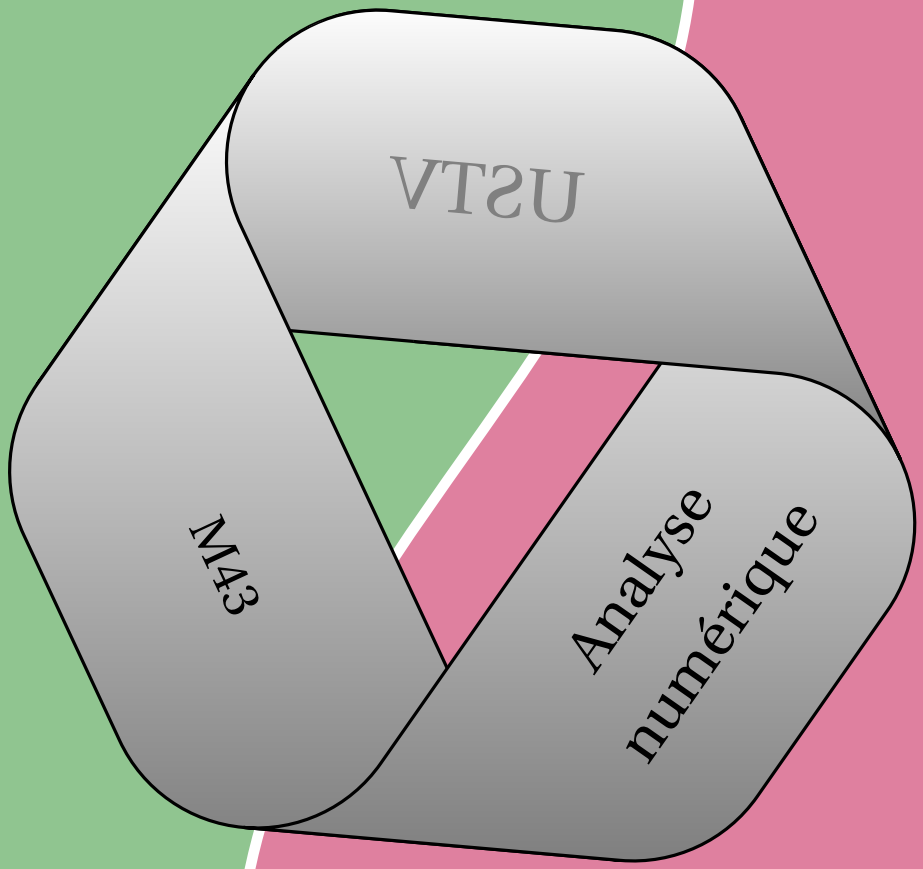
L2

2010/2011

Recueil d'exercices et aide-mémoire



G. FACCANONI

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

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 $\exists \alpha \in]a,b[$ tel que $f(\alpha) = 0$.

Dichotomie. En partant de $I_0 = [a,b]$, la méthode de dichotomie 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 $f(a_k)f(b_k) < 0$.

Plus précisément :

Require: a, b, ε
  $k \leftarrow 0$
  $a_k \leftarrow a$
  $b_k \leftarrow b$
  $x_k \leftarrow \frac{a_k + b_k}{2}$
  while $|b_k - a_k| > \varepsilon$ **do**
    $k \leftarrow k + 1$
    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
    $x_{k+1} \leftarrow \frac{a_{k+1} + b_{k+1}}{2}$
  end while

Point fixe. Il est toujours possible, pour $f\colon [a,b] \to \mathbb{R}$, de transformer le problème $f(x) = 0$ en un problème équivalent $x - \varphi(x) = 0$, où la fonction auxiliaire $\varphi\colon [a,b] \to \mathbb{R}$ 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 utilisant l'algorithme itératif suivant :

Require: x_0, ε
  $k \leftarrow 0$
  while $|x_{k+1} - x_k| > \varepsilon$ **do**
    $k \leftarrow k + 1$
    $x_{k+1} \leftarrow \varphi(x_k)$
  end while

Cas particuliers :

Méthode de la Corde	$\varphi(x_k) = x_k - \dfrac{b-a}{f(b)-f(a)} f(x_k).$
Méthode de Newton	$\varphi(x_k) = x_k - \dfrac{f(x_k)}{f'(x_k)}.$

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

1. $\varphi(x) \in [a,b]$ pour tout $x \in [a,b]$ (stabilité)
2. $\varphi \in \mathscr{C}^1([a,b])$ (régularité)
3. il existe $K < 1$ tel que $|\varphi'(x)| \leq K$ pour tout $x \in [a,b]$ (contraction)

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 $[a,b]$. Plus particulièrement,

▷ 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 $|\varphi(\alpha)|$ 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.

De plus, si $\varphi \in \mathscr{C}^{p+1}([a,b])$ pour un certain $p \geq 1$ et 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 $p+1$.

Exercice 1.1. 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 $g(x) = \frac{2}{3}x + \frac{1}{3}\frac{a}{x^2}$ ($a > 0$ fixé).

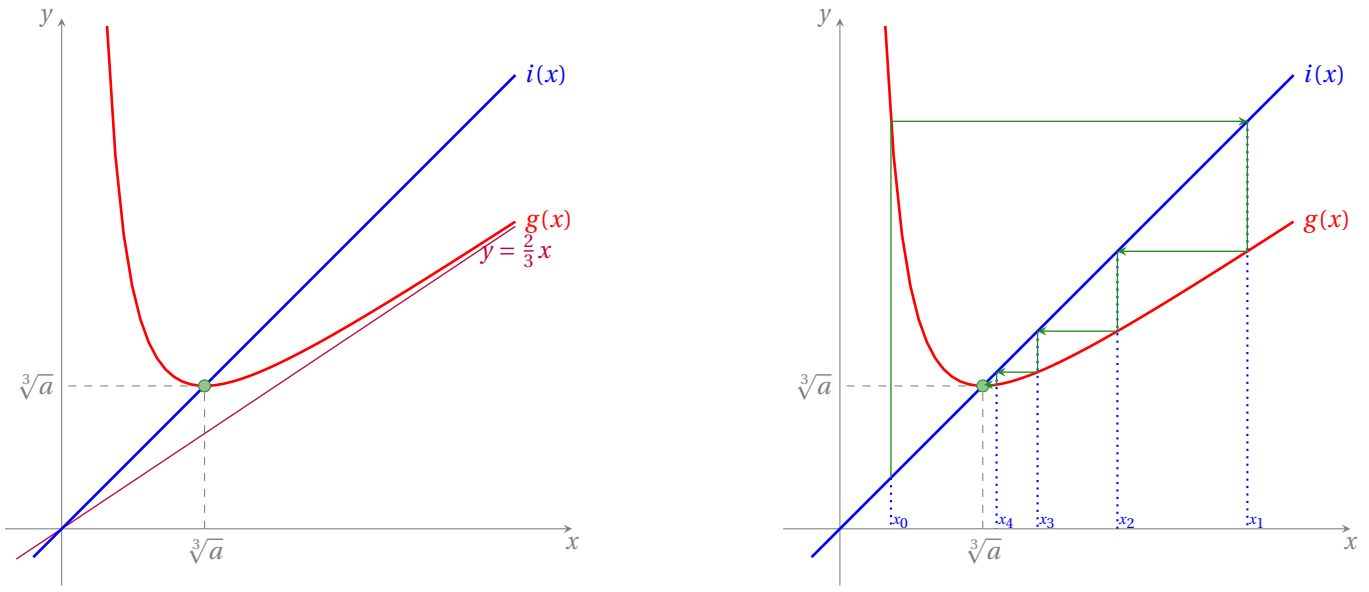
1. Faire l'étude complète de la fonction g.
2. Comparer g à l'identité.
3. Soit la suite $(x_n)_{n\in\mathbb{N}}$ définie par
$$x_{n+1} = g(x_n), \qquad x_0 > 0.$$
À l'aide des graphe de g et de l'identité sur $\mathbb{R}_+^*$, dessiner la suite $(x_n)_{n\in\mathbb{N}}$ sur l'axe des abscisses. Observer 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 $g : \mathbb{R}_+^* \to \mathbb{R}$ définie par $g(x) = \frac{2}{3}x + \frac{1}{3}\frac{a}{x^2}$:
   - ⋆ $g(x) > 0$ pour tout $x \in \mathbb{R}_+^*$;
   - ⋆ $\lim\limits_{x\to 0^+} g(x) = \lim\limits_{x\to+\infty} g(x) = +\infty$;
     $\lim\limits_{x\to+\infty} \frac{g(x)}{x} = \frac{2}{3}$ et $\lim\limits_{x\to+\infty} g(x) - \frac{2}{3}x = 0$ 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		$\sqrt[3]{a}$		$+\infty$
$g'(x)$		$-$		$+$	
$g(x)$	$+\infty$	↘	$\sqrt[3]{a}$	↗	$+\infty$

2. Graphe de g comparé au graphe de $i(x) = x$: voir la figure 1.1a. 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 1.1a.
4. On en déduit que pour tout $x > 0$ on a $g(x) \geq \sqrt[3]{a}$. Donc, pour tout $k > 0$, $x_k = g(x_{k-1}) \geq \sqrt[3]{a}$. Pour étudier la convergence de la méthode on procède alors par deux étapes :
   4.1. on vérifie d'abord la CN pour tout $\alpha \in [\sqrt[3]{a}, +\infty[: \alpha = g(\alpha) \iff \alpha = \sqrt[3]{a}$;
   4.2. vérifions maintenant les CS :
      4.2.1. pour tout x dans $[\sqrt[3]{a}, +\infty[$ on a $g(x) > \sqrt[3]{a}$ donc $g : [\sqrt[3]{a}, +\infty[\to [\sqrt[3]{a}, +\infty[$;
      4.2.2. $g \in \mathscr{C}^1([\sqrt[3]{a}, +\infty[)$;



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

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

FIGURE 1.1: Exercice 1.1

4.2.3. 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 (et 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.

6. Algorithme de point fixe :

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

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 α et x_k. Donc

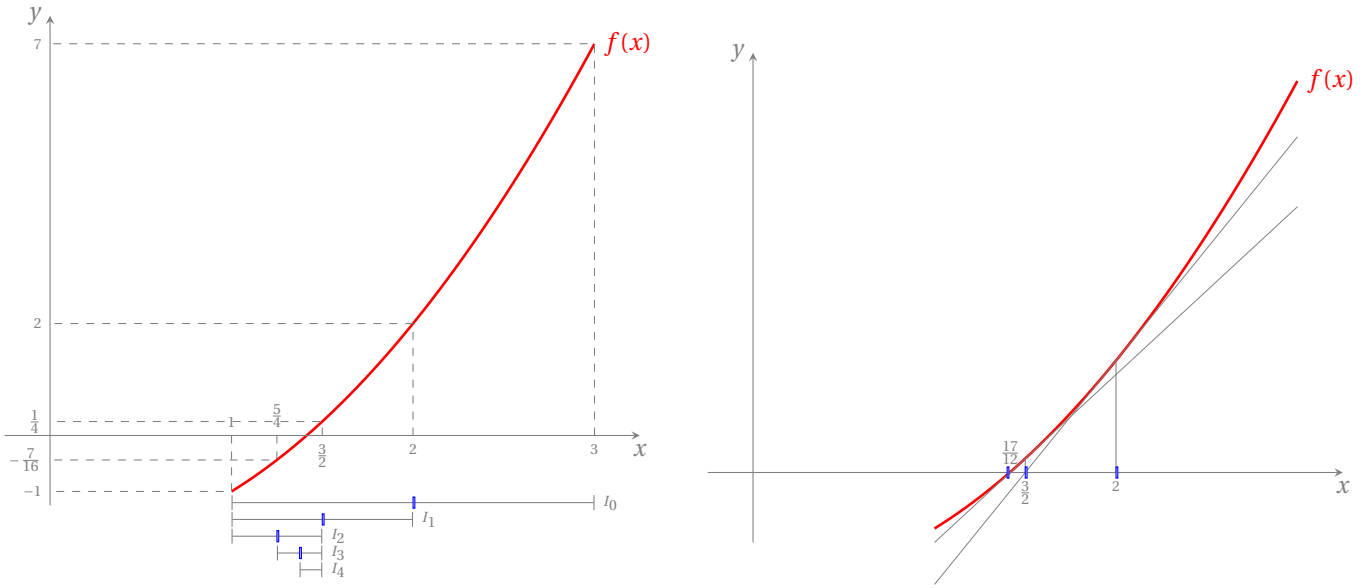
$$|x_{k+1} - x_k| = |e_{k+1} - e_k| = |g'(z_k) - 1|e_k \simeq |g'(\alpha) - 1|e_k.$$

Puisque $g'(\alpha) = 0$, on a bien $|x_{k+1} - x_k| \simeq e_k$.

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

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^3 - a}{3x_k^2} = x_k - \frac{1}{3}x_k + \frac{a}{3x_k^2} = \frac{2}{3}x_k + \frac{a}{3x_k^2}$$

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).



(a) Méthode de la dichotomie.

(b) Méthode de Newton.

FIGURE 1.2: Approximation du zéro de la fonction $f(x) = x^2 - 2$.

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 on doit effectuer pour améliorer d'un ordre de grandeur la précision de l'approximation de la racine ?

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 1.2a.
- ▷ 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 1.2b.

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	$\frac{3}{2} = 1{,}5$	$\frac{17}{12} = 1{,}41\bar{6}$	$\frac{17}{24} + \frac{12}{17} \simeq 1{,}4142156$

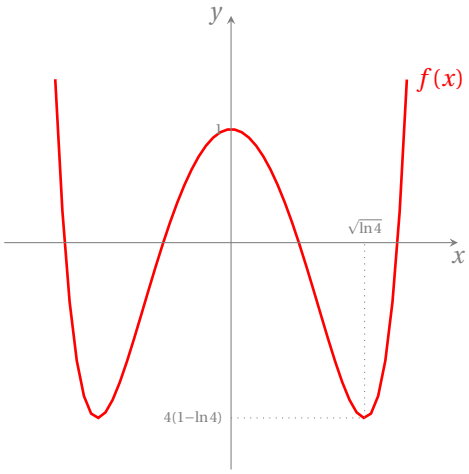
On rappelle qu'avec la méthode de la dichotomie, les itération s'achèvent à la m-ème étape quand $|x_m - \alpha| \leq |I_m| < \varepsilon$, où ε est une tolérance fixée et $|I_m|$ désigne la longueur de l'intervalle I_m. Clairement $I_k = \frac{b-a}{2^k}$, donc pour avoir $|x_m - \alpha| < \varepsilon$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

$$m \geq \log_2 \frac{b-a}{\varepsilon} - 1.$$

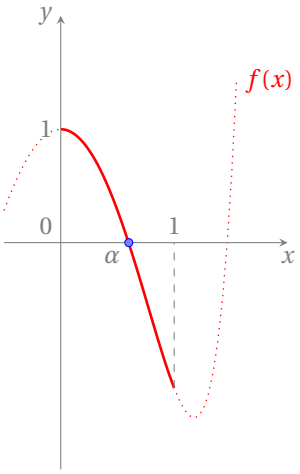
Améliorer d'un ordre de grandeur la précision de l'approximation de la racine signifie avoir

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

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



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



(b) Zoom.

FIGURE 1.3: Exercice 1.3

Exercice 1.3. 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
$$\begin{cases} x_{k+1} = \phi(x_k), \\ x_0 \in]0,1[, \end{cases} \tag{1.1}$$
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 (1.1), 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 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 1.3a 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 1.3b.
3. Étude de la convergence de la méthode (1.1) :
   3.1. on vérifie d'abord la CN pour tout $\alpha \in]0,1[$:
   $$\alpha = \phi(\alpha) \iff 2\alpha = \sqrt{\exp(\alpha^2)} \iff 4\alpha^2 = \exp(\alpha^2) \iff f(\alpha) = 0;$$
   3.2. vérifions maintenant les CS :
   3.2.1. pour tout x dans $]0,1[$ on a
   $$0 < \sqrt{\frac{\exp(x^2)}{4}} < \sqrt{\frac{e}{4}} < 1$$
   donc $\phi :]0,1[\to]0,1[$;

3.2.2. $\phi \in \mathscr{C}^1(]0,1[)$;
3.2.3. pour tout x dans $]0,1[$ on a

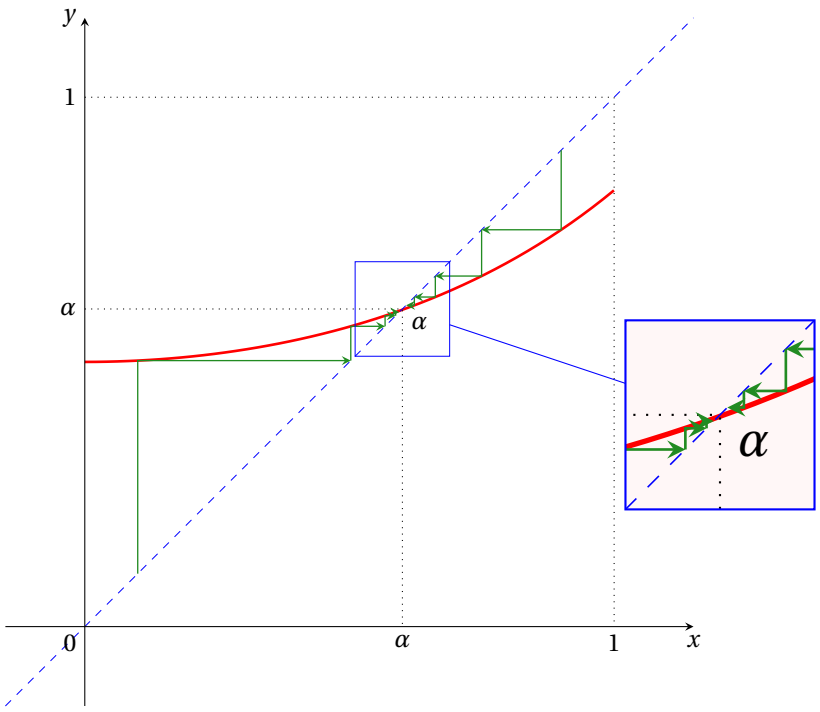
$$|\phi'(x)| = \left|\frac{x\sqrt{\exp(x^2)}}{2}\right| = |x\phi(x)| < |x| < 1$$

donc ϕ est contractante.

Alors la méthode (1.1) converge vers α point fixe de ϕ et zéro de f.
De plus, étant donné que

$$\phi'(\alpha) = \alpha\phi(\alpha) = \alpha^2 \neq 0$$

la méthode de point fixe (1.1) 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 (1.1) converge seulement à l'ordre 1 : la méthode de Newton est donc plus efficace.

Exercice 1.4. 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 $\lim_{x\to-\infty} f(x) = +\infty$ et $\lim_{x\to+\infty} f(x) = -\infty$.
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 :

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

2. Puisque $\lim_{x\to-\infty} f(x) = +\infty > 0$ et $\lim_{x\to+\infty} f(x) = -\infty < 0$, 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 $g : \mathbb{R} \to \mathbb{R}$;
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 $\frac{1}{2} < \alpha < 1$ 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 $x_0 \in R$.

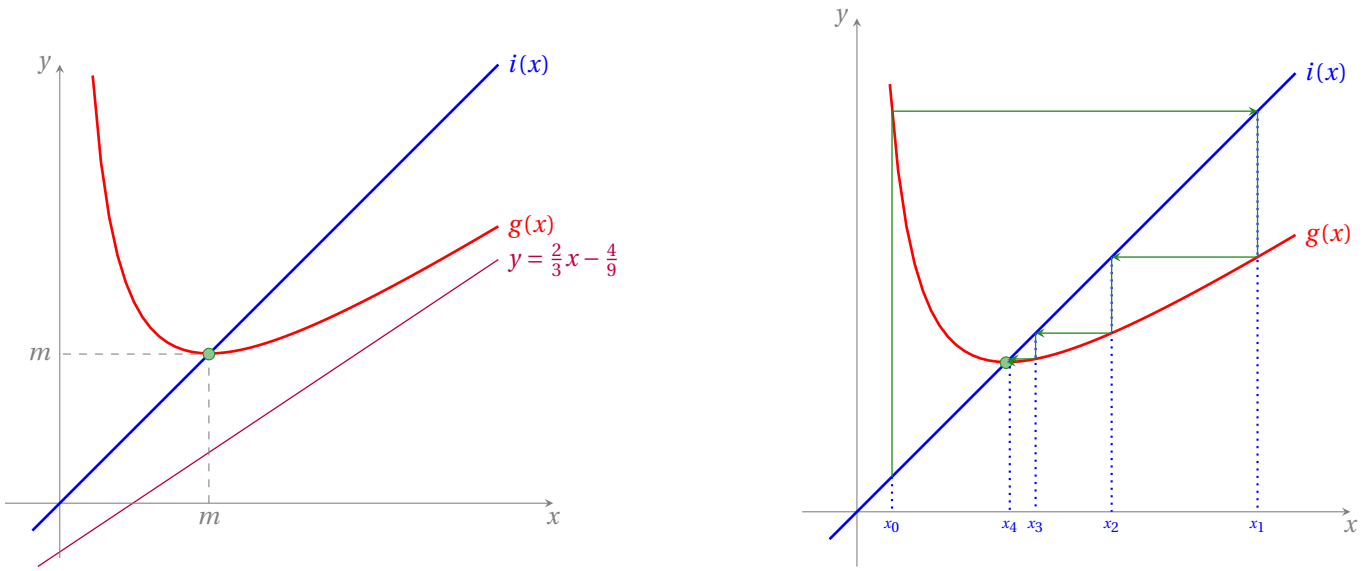
Exercice 1.5. Soit g la fonction définie sur $\mathbb{R}_+^*$ par
$$g(x) = \frac{2x^3+4x^2+10}{3x^2+8x}.$$
1. Faire l'étude complète de la fonction g. (On admettra que $x^3+4x^2-10=0$ admet comme unique solution $m \approx 1{,}36$ et que $g(m) = m$.)
2. Comparer g à l'identité.
3. Soit la suite $(x_n)_{n\in\mathbb{N}}$ 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		$+\infty$
$g'(x)$		$-$		$+$	
$g(x)$	$+\infty$	↘	m	↗	$+\infty$

2. Graphe de g comparé au graphe de $i(x) = x$: voir la figure 1.4a. 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 1.4b.
4. On en déduit que pour tout $x > 0$ on a $g(x) \geq m$. Donc, pour tout $k > 0$, $x_k = g(x_{k-1}) \geq m$. Pour étudier la convergence de la méthode on procède alors par deux étapes :
   4.1. on vérifie d'abord la CN pour tout $\alpha \in [m,+\infty[$: $\alpha = g(\alpha) \iff \alpha = m$;
   4.2. vérifions maintenant les CS :
      4.2.1. pour tout x dans $[m,+\infty[$ on a $g(x) > m$ donc $g\colon [m,+\infty[\to [m,+\infty[$;
      4.2.2. $g \in \mathscr{C}^1([m,+\infty[)$;
      4.2.3. pour tout x dans $[m,+\infty[$, $|g'(x)| = \left|\frac{(6x^2+8x)-g(x)(6x+8)}{3x^2+8x}\right| < 1$ alors g est contractante.

   Si les conditions précédentes sont vérifiées alors la méthode converge vers m point fixe de g (et racine de f).
5. Algorithme de point fixe :



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

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

FIGURE 1.4

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

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 $|x_{k+1} - x_k| \simeq e_k$.

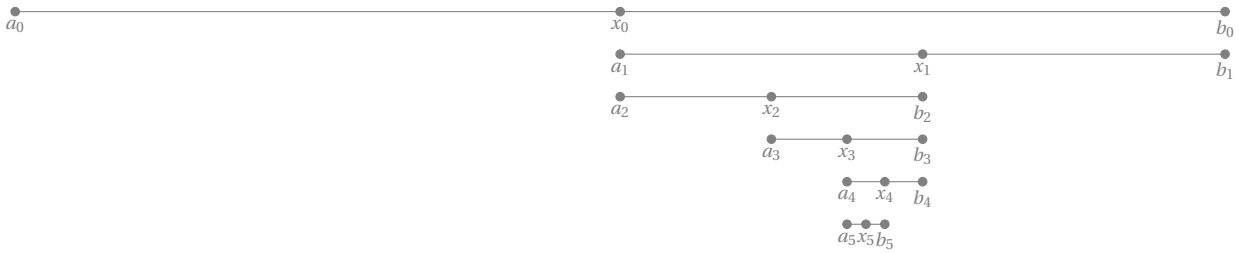
Exercice 1.6. Décrire la méthode de la dichotomie et l'utiliser pour calculer le zéro de la fonction $f(x) = x^3 - 4x - 8.95$ dans l'intervalle $[2;3]$ avec une précision de 10^{-2} en remplissant le tableau suivant :

k	a_k	x_k	b_k	signe de $f(a_k)$	signe de $f(x_k)$	signe de $f(b_k)$
0	2	2.5	3	-	-	+
1	2.5		3			
2						
3						
4						
5						
6						

SOLUTION. En partant de $I_0 = [a, b]$, la méthode de dichotomie 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 $f(a_k)f(b_k) < 0$. Plus précisément :

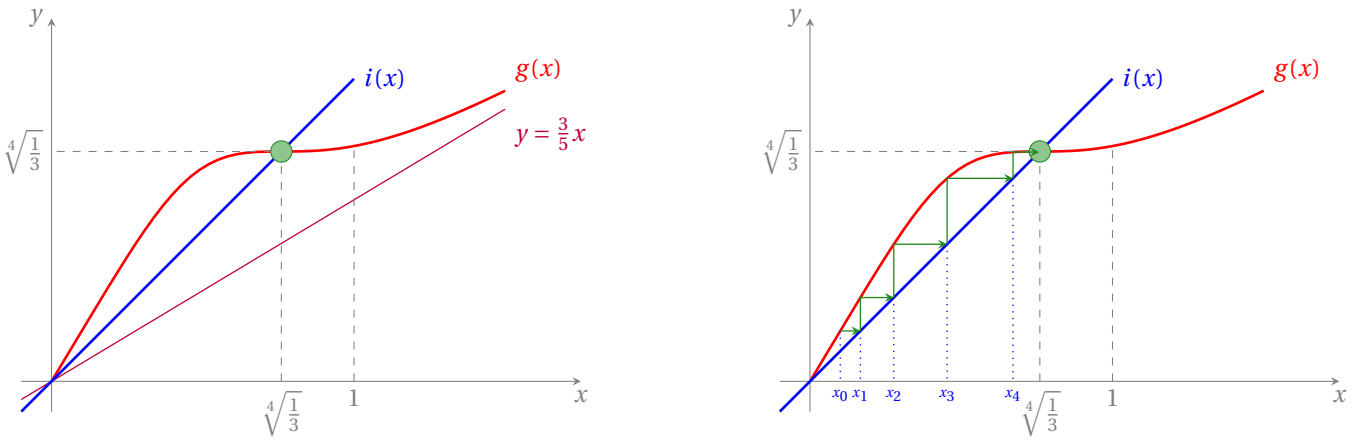
Require: a, b, ε
  $k \leftarrow 0$
  $a_k \leftarrow a$
  $b_k \leftarrow b$
  $x_k \leftarrow \frac{a_k+b_k}{2}$
  while $|b_k - a_k| > \varepsilon$ **do**
    $k \leftarrow k+1$
    if $f(a_k)f(x_k) < 0$ **then**
      $a_{k+1} \leftarrow a_k$
      $b_{k+1} \leftarrow x_k$
    else
      $a_{k+1}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

k	a_k	x_k	b_k	signe de $f(a_k)$	signe de $f(x_k)$	signe de $f(b_k)$
0	2	2.5	3	-	-	+
1	2.5	2.75	3	-	+	+
2	2.5	2.625	2.75	-	-	+
3	2.625	2.6875	2.75	-	-	+
4	2.6875	2.71875	2.75	-	+	+
5	2.6875	2.703125	2.71875	-	-	+
6	2.703125	2.7109375	2.71875	-	+	+



Exercice 1.7. 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.



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

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

FIGURE 1.5

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.

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 $x \geq 0$.

2.1. ▷ $g(x) \geq 0$ pour tout $x \geq 0$ 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 $g'(x) \geq 0$ 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 1.5a.

▷ 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 1.5b.

2.3. Étudions la convergence de la méthode [1]. On remarque que

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

1. Remarquons 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

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

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 $\ell \leq \sqrt[4]{\frac{1}{3}}$. 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 $\ell \leq \sqrt[4]{\frac{1}{3}}$. 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).

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: $g : x \mapsto g(x)$
  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 $n+1$ 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. 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). [1]

Erreur. Si $y_i = f(x_i)$ pour $i = 0, 1, \dots, n$, $f : I \to \mathbb{R}$ 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

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

1. 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

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

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 $\Pi_m(x) = a_0 + a_1 x + \dots a_m x^m \in \mathbb{P}_m$ tel quel

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

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

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

qu'on peut réécrire comme

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

Cette relation est appelée formule d'interpolation de Hermite. [2]

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

$$\begin{aligned} Q(x) = {} & y_0 \left(1 - 2(x-x_0)\left(\frac{1}{x_0-x_1} + \frac{1}{x_0-x_2}\right)\right)\left(\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}\right)^2 + y'_0(x-x_0)\left(\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}\right)^2 \\ & + y_1 \left(1 - 2(x-x_1)\left(\frac{1}{x_1-x_0} + \frac{1}{x_1-x_2}\right)\right)\left(\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}\right)^2 + y'_1(x-x_1)\left(\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}\right)^2 \\ & + y_2 \left(1 - 2(x-x_2)\left(\frac{1}{x_2-x_0} + \frac{1}{x_2-x_1}\right)\right)\left(\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\right)^2 + y'_2(x-x_2)\left(\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\right)^2, \end{aligned}$$

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

$$\begin{cases} a_0 + a_1 x_0 + \dots a_N x_0^N = y_0 \\ a_0 + a_1 x_1 + \dots a_N x_1^N = y_1 \\ \dots \\ a_n + a_1 x_n + \dots a_N x_n^N = y_n \\ a_1 + a_2 x_0 + \dots N a_N x_0^{N-1} = y'_0 \\ a_1 + a_2 x_0 + \dots N a_N x_1^{N-1} = y'_1 \\ \dots \\ a_n + a_1 x_n + \dots N a_N x_n^{N-1} = y'_n \end{cases} \quad \text{i.e.} \quad \begin{pmatrix} 1 & x_0 & \dots & x_0^N \\ 1 & x_1 & \dots & x_1^N \\ \vdots & & \vdots & \vdots \\ 1 & x_n & \dots & x_n^N \\ 0 & x_0 & \dots & N x_0^{N-1} \\ 0 & x_1 & \dots & N x_1^{N-1} \\ \vdots & & \vdots & \vdots \\ 0 & x_n & \dots & N x_n^{N-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ \vdots \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \\ y'_0 \\ y'_1 \\ \vdots \\ y'_n \end{pmatrix}$$

qu'on peut réécrire comme

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

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 $a=x_0<x_1<\cdots<x_n=b$, la fonction $s_k(x)\colon [a;b]\to\mathbb{R}$ est une spline de degré k relative aux nœuds $\{x_i\}$ si

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

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 $a=x_0<x_1<\cdots<x_n=b$, la fonction $\ell(x)\colon [a;b]\to\mathbb{R}$ est une spline linéaire relative aux nœuds $\{x_i\}$ si

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

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

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

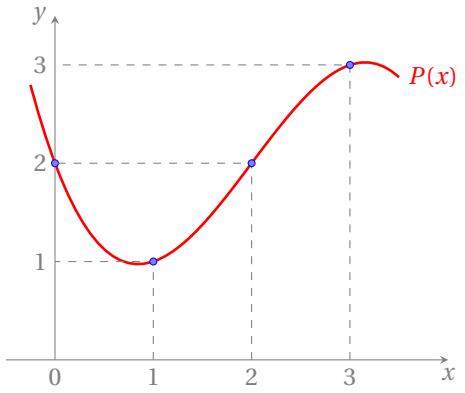
Exercice 2.1. Construire le polynôme de Lagrange P qui interpole les points $(0,2)$, $(1,1)$, $(2,2)$ et $(3,3)$.

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

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



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

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

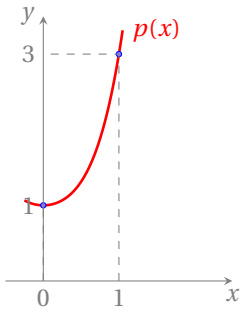
Exercice 2.2. Trouver le polynôme de l'espace vectoriel $\mathrm{Vec}\{1+x^2, x^4\}$ qui interpole les points $(0,1)$ et $(1,3)$.

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 $(1,3)$:

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

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 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 $\mathrm{Vec}\{1, x, x^2\}$ qui interpole les trois points $(-1,-1)$, $(0,0)$ et $(1,-1)$.

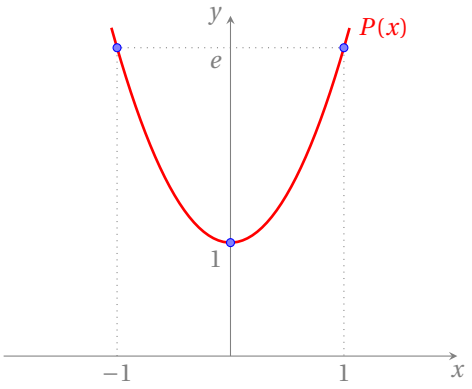
SOLUTION.

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

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

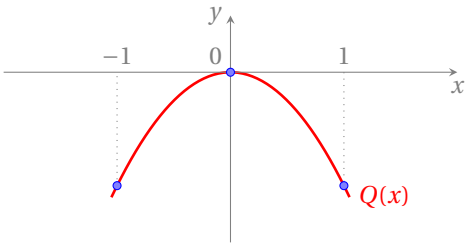
Ici $n = 2$ donc on a

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



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 $(1,-1)$:

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

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

Exercice 2.4.

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

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

2. Par construction

$$Q(-1) = P(-1),$$
$$Q(0) = P(0),$$

$$Q(1) = P(1),$$

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

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

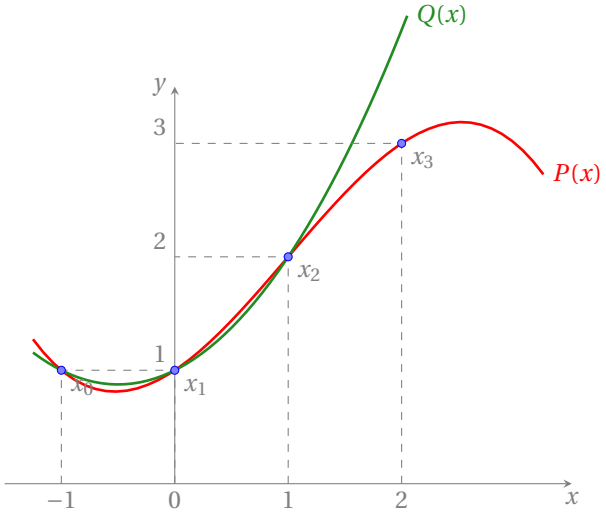
Ainsi

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

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 2.5.

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 2.6. 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

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

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

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

Exercice 2.7. 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

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

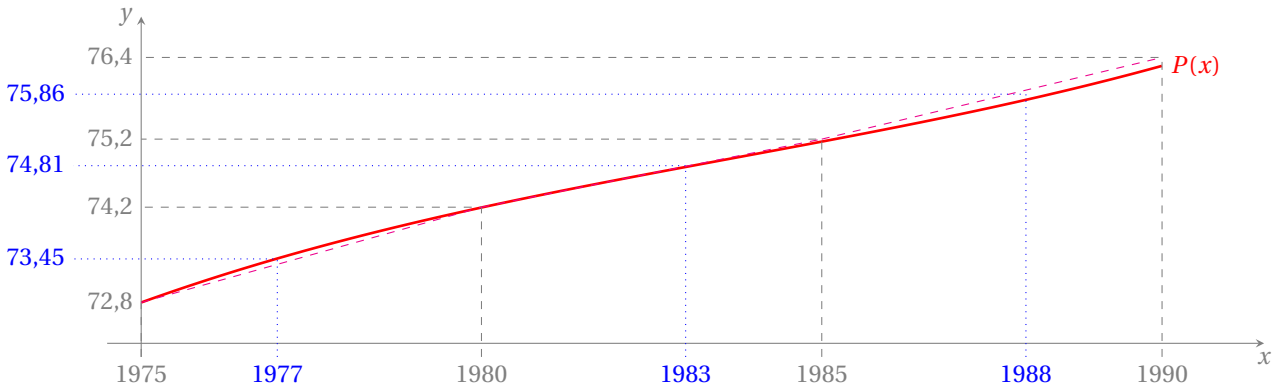
Ici $n = 3$ et si on choisit de poser $x_0 = 0$ pour l'année 1975, $x_1 = 5$ pour l'année 1980 etc., on a

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

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 2.8. La production de citron d'un Pays a évoluée comme suit

Année	1965	1970	1980	1985	1990	1991
Production ($\times 10^5$ kg)	17769	24001	25961	34336	29036	33417

Utiliser une interpolation polynomiale pour estimer la production des années 1962, 1977 et 1992. Comparer les résultats avec les valeurs réels : 12380, 27403 et 32059. Comparer ensuite avec les valeurs estimés par une interpolation linéaire par morceaux (splines de degré 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

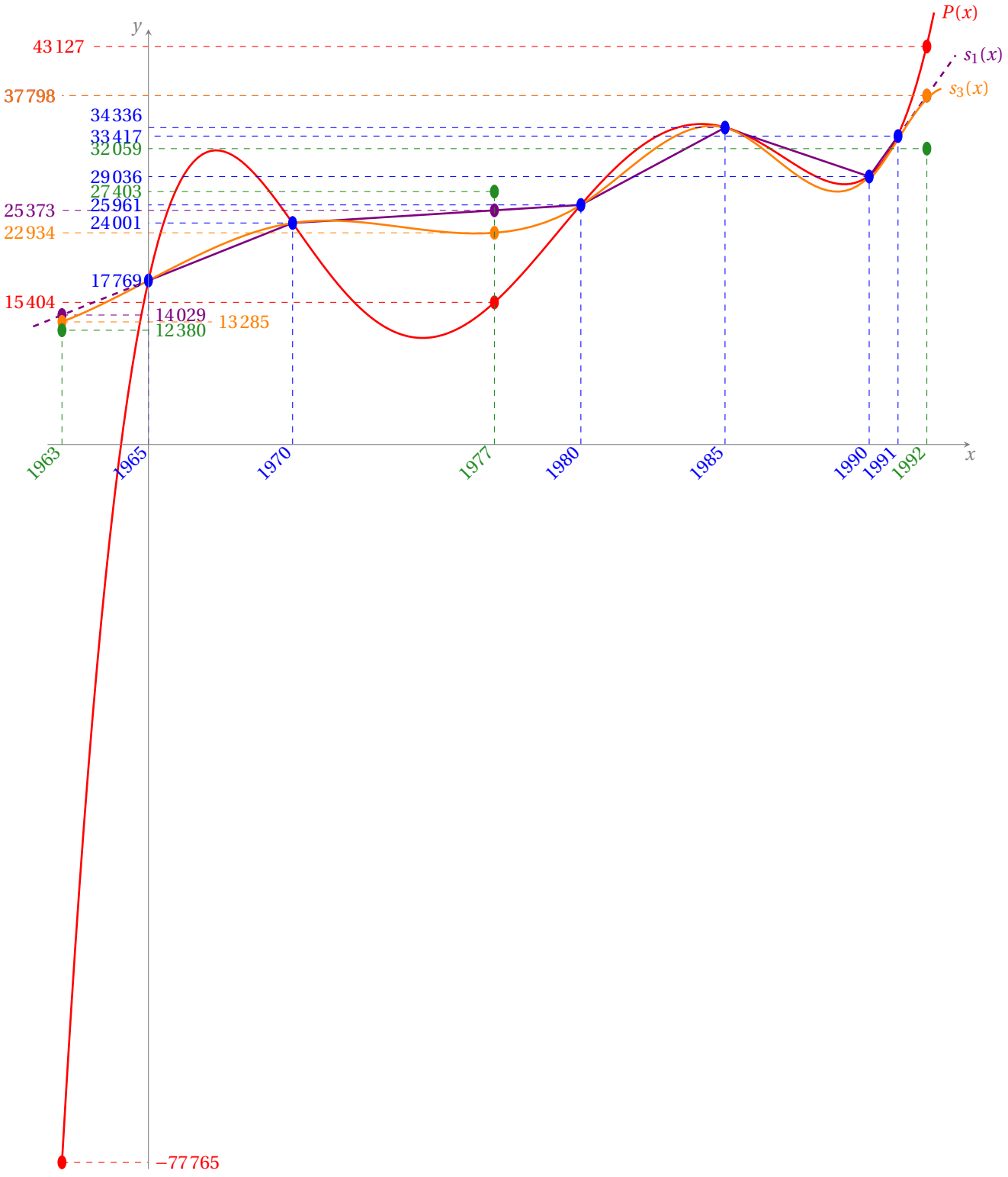
$$p_n(x) = \sum_{i=0}^{n} \left(y_i \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} \right).$$

Ici $n = 5$ et si on choisit de poser $x_0 = 0$ pour l'année 1965, $x_1 = 5$ pour l'année 1970, $x_2 = 15$ pour l'année 1980 etc., on a

$$\begin{aligned}
P(x) = {} & y_0 \frac{(x-x_1)(x-x_2)(x-x_3)(x-x_4)(x-x_5)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)(x_0-x_4)(x_0-x_5)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)(x-x_4)(x-x_5)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)(x_1-x_4)(x_1-x_5)} \\
& + y_2 \frac{(x-x_0)(x-x_1)(x-x_3)(x-x_4)(x-x_5)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)(x_2-x_4)(x_2-x_5)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_4)(x-x_5)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)(x_3-x_4)(x_3-x_5)} \\
& + y_4 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_3)(x-x_5)}{(x_4-x_0)(x_4-x_1)(x_4-x_2)(x_4-x_3)(x_4-x_5)} + y_5 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_3)(x-x_4)}{(x_5-x_0)(x_5-x_1)(x_5-x_2)(x_5-x_3)(x_5-x_4)} = \\
= {} & 17769 \frac{(x-5)(x-15)(x-20)(x-25)(x-26)}{(0-5)(0-15)(0-20)(0-25)(0-26)} + 24001 \frac{(x-0)(x-15)(x-20)(x-25)(x-26)}{(5-0)(5-15)(5-20)(5-25)(5-26)} \\
& + 25961 \frac{(x-0)(x-5)(x-20)(x-25)(x-26)}{(15-0)(15-5)(15-20)(15-25)(15-26)} + 34336 \frac{(x-0)(x-5)(x-15)(x-25)(x-26)}{(20-0)(20-5)(20-15)(20-25)(20-26)} \\
& + 29036 \frac{(x-0)(x-5)(x-15)(x-20)(x-26)}{(25-0)(25-5)(25-15)(25-20)(25-26)} + 33417 \frac{(x-0)(x-5)(x-15)(x-20)(x-25)}{(26-0)(26-5)(26-15)(26-20)(26-25)} = \\
= {} & -5923 \frac{(x-5)(x-15)(x-20)(x-25)(x-26)}{325000} + 24001 \frac{x(x-15)(x-20)(x-25)(x-26)}{31500} \\
& - 25961 \frac{x(x-5)(x-20)(x-25)(x-26)}{82500} + 4292 \frac{x(x-5)(x-15)(x-25)(x-26)}{5625} \\
& - 7259 \frac{x(x-5)(x-15)(x-20)(x-26)}{6250} + 3713 \frac{x(x-5)(x-15)(x-20)(x-25)}{4004} =
\end{aligned}$$

$$= 17769 + \frac{1407243973}{100100}x - \frac{39576120413}{9009000}x^2 + \frac{10206685933}{22522500}x^3 - \frac{325261939}{17325000}x^4 + \frac{7663144}{28153125}x^5$$

On a alors que
- ▷ la production de l'année 1962 estimée par cette interpolation est $P(-3) = -77765{,}36380$,
- ▷ la production de l'année 1977 estimée par cette interpolation est $P(12) = 15404{,}96367$,
- ▷ la production de l'année 1992 estimée par cette interpolation est $P(27) = 43127{,}20660$.



Exercice 2.9. 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	$\frac{2-e}{e} - \frac{2}{e^{\frac{2}{e}}} \approx 0.6940422999$

3 Quadrature

Soit f une fonction réelle intégrable sur l'intervalle $[a;b]$. Le calcul explicite de l'intégrale définie $I(f) = \int_a^b f(x)\mathrm{d}x$ 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

$$I_n(f) = \int_a^b f_n(x)\mathrm{d}x, \qquad n \geq 0.$$

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

$$E_n(f) \leq \int_a^b |f(x) - f_n(x)|\mathrm{d}x \leq (b-a)\|f - f_n\|_\infty.$$

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

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

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

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

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 $I_n(f) = I(f)$ pour tout polynôme $f \in \mathbb{P}^r$.

Théorème. Toute formule de quadrature interpolatoire utilisant $n+1$ nœuds distincts a un degré d'exactitude au moins égale à n. *En effet, si $f \in \mathbb{P}^n$, 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 est dite stable s'il existe $M \in \mathbb{R}_+^*$ tel que $\sum_{i=0}^n |\alpha_i| \leq M$.

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 $j = 0, 1, \ldots, m$. 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 $y = f(x)$ qui envoie l'intervalle $[a;b]$ dans l'intervalle $[c;d]$. Comme on veut une transformation affine, on cherche $f(x)$ sous la forme $mx + q$. On doit alors résoudre le système linéaire

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

On obtient

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

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

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

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

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

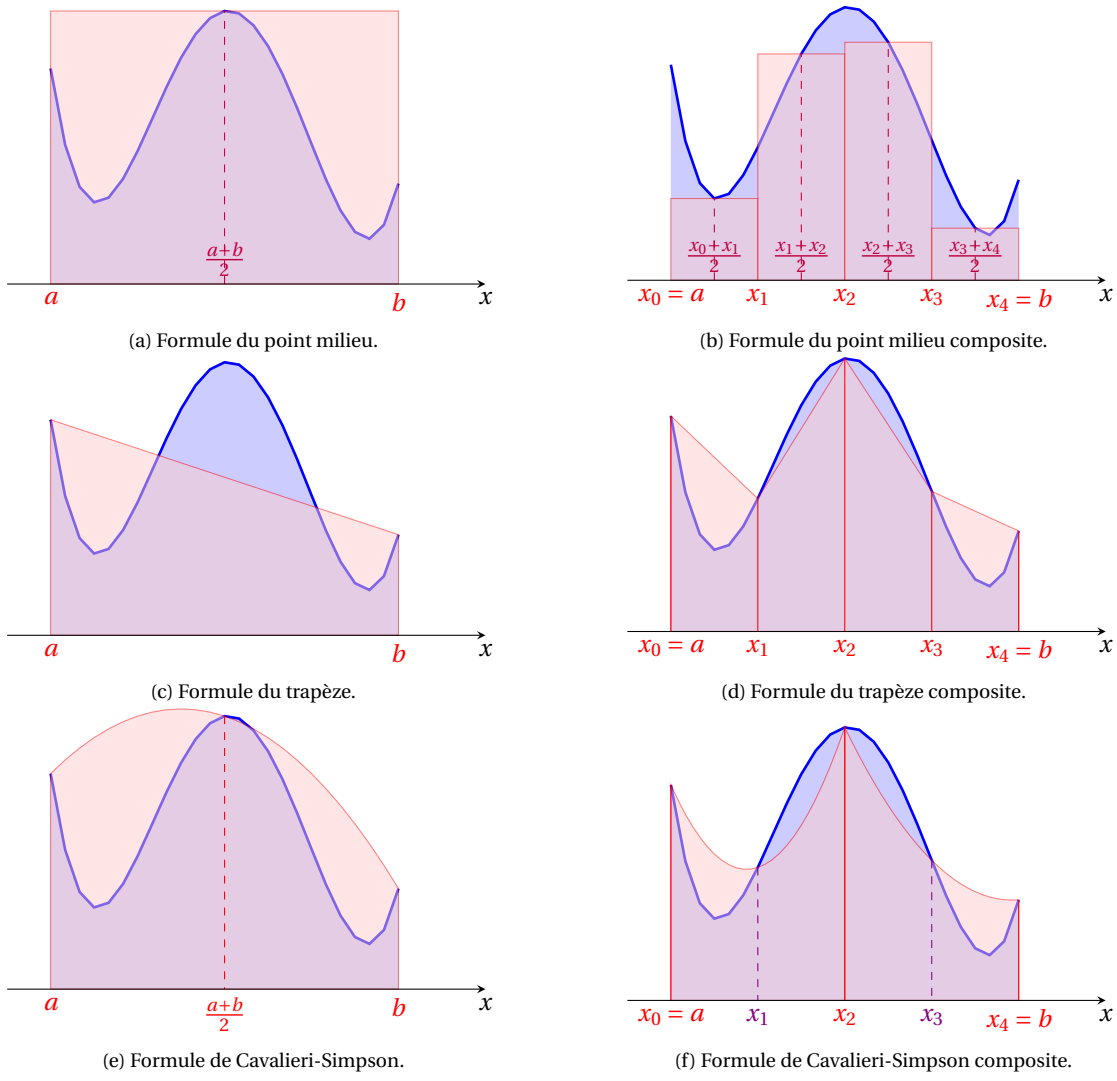
Exemples de formules de quadrature interpolatoires

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

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



(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 $n = 0, 1, 2$.

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 $m \geq 1$ En introduisant les noeds de quadrature $x_k = a + \frac{2k+1}{2} H$ pour $k = 0, 1, \ldots, m-1$ 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

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

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 $m\geq 1$. En introduisant les nœuds de quadrature $x_k=a+kH$ pour $k=0,1,\ldots,m-1$ 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 $H=\frac{b-a}{m}$ avec $m\geq 1$. En introduisant les nœuds de quadrature $x_k=a+kH/2$ pour $k=0,1,\ldots,2m$ on obtient

$$I_{2,m}(f)=\frac{H}{6}\left(f(a)+2\sum_{r=1}^{m-1}f(x_{2r})+4\sum_{s=0}^{s-1}f(x_{2s+1})+f(b)\right)=\frac{H}{6}\left(f(a)+2\sum_{r=1}^{m-1}f(a+rH)+4\sum_{s=0}^{m-1}f\left(a+\frac{2s+1}{2}H\right)+f(b)\right).$$

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

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

Exercice 3.1. 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. Soit $p(x) = a + bx + cx^2 + dx^3 + ex^4$, alors

$$\int_{-1}^{1} p(x)\,\mathrm{d}x = \left[ax + b\frac{x^2}{2} + c\frac{x^3}{3} + d\frac{x^4}{4} + e\frac{x^5}{5}\right]_{-1}^{1} = 2a + 2\frac{c}{3} + 2\frac{e}{5}$$

et

$$\frac{2}{3}\left(2p(-\tfrac{1}{2}) - p(0) + 2p(\tfrac{1}{2})\right) = 2a + 2\frac{c}{3} + \frac{e}{6}$$

La formule est donc exacte de degré 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 (exacte sur l'espace des polynôme de degré au plus 3)

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

Soit $h = x_{i+1} - x_i = \frac{b-a}{n}$. La formule précédente se réécrit

$$\int_{x_i}^{x_{i+1}} f(y)\,\mathrm{d}y \approx \frac{h}{3}\left[2f\left(x_i + \frac{h}{4}\right) - f\left(x_i + \frac{h}{2}\right) + 2f\left(x_i + \frac{3h}{4}\right)\right].$$

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

Algorithm 4 Calcul de $\int_a^b f(x)\,\mathrm{d}x$

Require: f
Require: a
Require: $b > a$
Require: $n > 0$
  $h \leftarrow \frac{b-a}{n}$
  $s \leftarrow 0$
  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 $I \leftarrow \frac{h}{3}s$

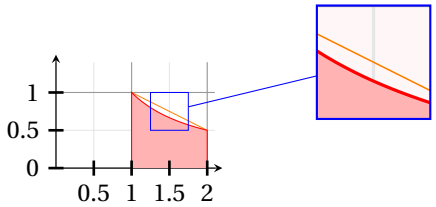
Exercice 3.2. On considère l'intégrale

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

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
$$|E|\leq\frac{(b-a)^4}{12m^2}\sup_{\xi\in]a;b[}|f''(\xi)|.$$
Donc ici on a
$$|E|\leq\frac{1}{12m^2}\max_{\xi\in]1;2[}\frac{2}{\xi^3}=\frac{1}{6m^2}.$$
Pour que $|E|<10^{-4}$ il faut 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 3.3. 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
$$\int_a^b f(x)\,\mathrm{d}x. \tag{3.1}$$
On propose dans un premier temps (question 1 à 5) de construire la formule de quadrature à 3 points de Simpson :
$$\int_{-1}^{1} g(x)\,\mathrm{d}x\approx\alpha g(-1)+\beta g(0)+\alpha g(1), \tag{3.2}$$
où les réels α et β sont à déterminer.

1. Sous quelle condition (portant sur α et β) la formule de quadrature (3.2) est exacte pour une fonction g constante ?
2. Sous quelle condition (portant sur α et β) la formule de quadrature (3.2) est exacte pour une fonction g polynomiale de degré au plus 2 ?
3. En déduire le choix de α et β rendant la formule de quadrature (3.2) exacte pour une fonction g polynomiale de degré au plus 2.
4. La formule de quadrature est-elle exacte pour tout polynôme de degré 3 ? La formule de quadrature est-elle exacte pour tout polynôme de degré 4 ?
5. 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 :
$$\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x.$$
6. 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 ?

7. Écrire l'algorithme du calcul de F.
8. 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.
9. En déduire une estimation d'erreur entre (3.1) et F.

SOLUTION.

1. Soit $g(x) = c$ alors on a

$$\begin{array}{ccc} \int_{-1}^{1} g(x)\,\mathrm{d}x & = & \alpha g(-1) + \beta g(0) + \alpha g(1) \\ \| & & \| \\ c\,[x]_{-1}^{1} & & \alpha c + \beta c + \alpha c \\ \| & & \| \\ 2c & & (2\alpha + \beta)c \end{array}$$

   d'où la relation

$$2 = 2\alpha + \beta.$$

2. Soit $g(x) = c + dx + ex^2$ alors on a

$$\begin{array}{ccc} \int_{-1}^{1} g(x)\,\mathrm{d}x & = & \alpha g(-1) + \beta g(0) + \alpha g(1) \\ \| & & \| \\ \left[cx + d\frac{x^2}{2} + e\frac{x^3}{3}\right]_{-1}^{1} & & \alpha(c - d + e) + \beta(c) + \alpha(c + d + e) \\ \| & & \| \\ 2c + \frac{2}{3}e & & (2\alpha + \beta)c + 2\alpha e \end{array}$$

   d'où les relations

$$\begin{cases} 2 = 2\alpha + \beta, \\ \frac{2}{3} = 2\alpha. \end{cases}$$

3. On dispose de 2 équations et 2 inconnues. Si on résout le système linéaire

$$\begin{cases} 2 = 2\alpha + \beta, \\ \frac{2}{3} = 2\alpha, \end{cases}$$

   on obtient l'unique solution

$$\alpha = \frac{1}{3}, \qquad \beta = \frac{4}{3}.$$

   On a retrouvé la formule de Simpson :

$$\int_{-1}^{1} g(x)\,\mathrm{d}x \approx \frac{1}{3}\left[g(-1) + 4g(0) + g(1)\right].$$

   On sait (cf. cours) que cette formule est exacte pour les polynômes de degré au plus 3 comme on va vérifier ci-dessous.
4. Soit $g(x) = c + dx + ex^2 + fx^3$ alors on a

$$\int_{-1}^{1} g(x)\,\mathrm{d}x = \left[cx + d\frac{x^2}{2} + e\frac{x^3}{3} + f\frac{x^4}{4}\right]_{-1}^{1} = 2c + \frac{2}{3}e$$

   et

$$\frac{1}{3}\left[g(-1) + 4g(0) + g(1)\right] = \frac{1}{3}\left[(c - d + e - f) + 4(c) + (c + d + e + f)\right] = 2c + \frac{2}{3}e$$

   donc la formule de quadrature est exacte pour tout polynôme de degré 3.
   On vérifie qu'elle n'est pas exacte pour les polynômes de degré 4 : soit $g(x) = x^4$, alors

$$\int_{-1}^{1} g(x)\,\mathrm{d}x = \left[\frac{x^5}{5}\right]_{-1}^{1} = \frac{2}{5}$$

   et

$$\frac{1}{3}\left[g(-1) + 4g(0) + g(1)\right] = \frac{1}{3}\left[1 + 1\right] = \frac{2}{3}$$

   donc la formule de quadrature n'est pas exacte pour les polynômes de degré 4.

5. Par le changement de variable $y = x_i + (x+1)\frac{x_{i+1}-x_i}{2}$ on déduit la formule de quadrature (exacte sur l'espace des polynôme de degré au plus 3)

$$\int_{x_i}^{x_{i+1}} f(y)\,\mathrm{d}y = \frac{x_{i+1}-x_i}{2}\int_{-1}^{1} f\left(x_i + (x+1)\frac{x_{i+1}-x_i}{2}\right)\mathrm{d}x \approx \frac{x_{i+1}-x_i}{6}\left[f(x_i) + 4f\left(\frac{x_{i+1}+x_i}{2}\right) + f(x_{i+1})\right].$$

6. On trouve ainsi la formule de quadrature composite (*i.e.* sur n sous-intervalles)

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n-1}\frac{x_{i+1}-x_i}{6}\left[f(x_i) + 4f\left(\frac{x_{i+1}+x_i}{2}\right) + f(x_{i+1})\right].$$

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 a

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &\approx \frac{H}{6}\sum_{i=0}^{n-1}\left[f(x_i) + 4f\left(x_i + \frac{H}{2}\right) + f(x_{i+1})\right]\\ &= \frac{H}{6}\left[f(a) + f(b) + 2\sum_{i=1}^{n-1} f(x_i) + 4\sum_{i=0}^{n-1} f\left(x_i + \frac{H}{2}\right)\right]\\ &= \frac{H}{6}\left[f(a) + f(b) + 2\sum_{i=1}^{n-1} f(a+iH) + 4\sum_{i=0}^{n-1} f\left(a + \frac{(i+1)}{2}H\right)\right].\end{aligned}$$

On peut changer de variables et réécrire la formule de quadrature sous la forme

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{h}{3}\left[f(a) + f(b) + 2\sum_{k=1}^{2n-1} f(a+2kh) + 4\sum_{k=0}^{2n-1} f(a+(k+1)h)\right] \qquad \text{avec } h = \frac{b-a}{2n}.$$

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

$$\begin{aligned}\frac{H}{6}\left[1 + 1 + 2\sum_{k=1}^{n-1} 1 + 4\sum_{k=0}^{n-1} 1\right] &= \frac{b-a}{6n}\left[2 + 2(n-1) + 4n\right]\\ &= \frac{b-a}{6n}6n = (b-a).\end{aligned}$$

7. Algorithme du calcul de F :

Algorithm 5 Calcul de $\int_a^b f(x)\,\mathrm{d}x$

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

8. Soit x un élément de $[x_i; x_{i+1}]$. Une formule de Taylor à l'ordre 3 pour f en x s'écrit

$$f(x) = P_i(x) + R_i(x),$$

avec

$$P_i(x) = f(x_i) + (x-x_i)f'(x_i) + (x-x_i)^2\frac{f''(x_i)}{2} + (x-x_i)^3\frac{f'''(x_i)}{6} \qquad \in \mathbb{P}_3$$

et le reste de Lagrange

$$R_i(x) = (x-x_i)^4\frac{f^{IV}(\xi)}{24} \qquad \text{avec} \quad \xi \in]x_i; x_{i+1}[.$$

On peut majorer R_i sur $[x_i; x_{i+1}]$ en fonction de $H = x_{i+1} - x_i$:

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

9. On en déduit l'estimation d'erreur entre (3.1) et F suivante [1]

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

Exercice 3.4. 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

$$\int_a^b f(x)\,\mathrm{d}x. \tag{3.3}$$

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

$$\int_{-1}^{1} g(x)\,\mathrm{d}x \approx \frac{4}{3}g(-\frac{w}{2}) + \frac{2}{3}g(w), \tag{3.4}$$

où $0 < w \le 1$ est à déterminer.

1. Montrer que cette méthode est toujours exacte pour toute fonction g polynomiale de degré 1.
2. Déterminer w pour que la formule de quadrature (3.4) soit exacte pour toute fonction g polynomiale de degré $m > 1$ et donner la plus grande valeur de m.
3. À l'aide d'un changement de variable affine, en déduire une formule de quadrature pour l'intégrale suivante :

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

4. En déduire une formule de quadrature à $2n$ points, notée F, pour le calcul approché de (3.3). Cette formule de quadrature est-elle stable ?
5. Écrire l'algorithme du calcul de F.

SOLUTION.

1. Soit $g(x) = c + dx$ alors on a

$$\begin{array}{ccc} \int_{-1}^{1} g(x)\,\mathrm{d}x & = & \frac{2}{3}g(w) + \frac{4}{3}g(-\frac{w}{2}) \\ \| & & \| \\ \left[cx + d\frac{x^2}{2}\right]_{-1}^{1} & & (\frac{2}{3}+\frac{4}{3})c + (\frac{2}{3} - \frac{4}{3}\frac{1}{2})dw \\ \| & & \| \\ 2c & & 2c \end{array}$$

   donc la méthode est exacte pour tout polynôme de degré au moins 1.
2. Soit $g(x) = c + dx + ex^2$ alors on a

$$\begin{array}{ccc} \int_{-1}^{1} g(x)\,\mathrm{d}x & = & \frac{2}{3}g(w) + \frac{4}{3}g(-\frac{w}{2}) \\ \| & & \| \\ \left[cx + d\frac{x^2}{2} + e\frac{x^3}{3}\right]_{-1}^{1} & & (\frac{2}{3}+\frac{4}{3})c + (\frac{2}{3} - \frac{4}{3}\frac{1}{2})dw + (\frac{2}{3}+\frac{1}{3})ew^2 \\ \| & & \| \\ 2c + \frac{2}{3}e & & 2c + ew^2 \end{array}$$

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

Pour que la méthode soit exacte pour tout polynôme de degré au moins 2 il faut choisir w tel que

$$\frac{2}{3} = w^2.$$

On obtient les deux solutions

$$w = -\sqrt{\frac{2}{3}}, \qquad w = \sqrt{\frac{2}{3}}.$$

L'hypothèse $0 < w \leq 1$ impose alors le choix

$$w = \sqrt{\frac{2}{3}}.$$

Soit maintenant $g(x) = c + dx + ex^2 + fx^3$. On a

$$\int_{-1}^{1} g(x)\,\mathrm{d}x = \left[cx + d\frac{x^2}{2} + e\frac{x^3}{3} + f\frac{x^4}{4}\right]_{-1}^{1} = 2c + \frac{2}{3}e$$

et, si $w^2 = \frac{2}{3}$, alors

$$\frac{2}{3}\left[g(w) + 2g(-w/2)\right] = 2c + \frac{2}{3}e + \frac{1}{2}f\left(\sqrt{\frac{2}{3}}\right)^3$$

donc la formule de quadrature est exacte pour tout polynôme de degré 2 mais n'est pas exacte pour les polynômes de degré 3.

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

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

4. 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)

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

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

5. Algorithme du calcul de F :

Algorithm 6 Calcul de $\int_a^b f(x)\,\mathrm{d}x$

Require: f
Require: a
Require: $b > a$
Require: $n > 0$
  $H \leftarrow \frac{b-a}{n}$
  $\alpha_1 \leftarrow a + H\left(1+\sqrt{\frac{2}{3}}\right)$
  $\alpha_2 \leftarrow a + H\left(1-\sqrt{\frac{1}{6}}\right)$
  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 3.5. 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
$$\begin{aligned} p(x) &= \frac{1}{4}\left[f(-1)(x+2)(x-1)^2+f'(-1)(x+1)(x-1)^2+f(1)(2-x)(x+1)^2+f'(1)(x-1)(x+1)^2\right]\\ &= \frac{1}{4}\Big[f(-1)(x^3-3x+2)+f'(-1)(x^3-x^2-x+1)+f(1)(-x^3+3x+2)+f'(1)(x^3+x^2-x-1)\Big]\\ &= \frac{2f(-1)+f'(-1)+2f(1)-f'(1)}{4}+\frac{3f(1)-3f(-1)-f'(-1)-f'(1)}{4}x\\ &+\frac{f'(1)-f'(-1)}{4}x^2+\frac{f(-1)+f'(-1)-f(1)+f'(1)}{4}x^3. \end{aligned}$$
   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

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

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

$$\begin{aligned} \int_{-1}^{1} p(x)\,\mathrm{d}x &= \left[\alpha x+\frac{\beta}{2}x^2+\frac{\gamma}{3}x^3+\frac{\delta}{4}x^4\right]_{-1}^{1} \\ &= 2\alpha+\frac{2}{3}\gamma \\ &= \frac{2f(-1)+2f(1)+f'(-1)-f'(1)}{2}+\frac{-f'(-1)+f'(1)}{6} \\ &= \frac{6f(-1)+6f(1)+3f'(-1)-3f'(1)-f'(-1)+f'(1)}{6} \\ &= f(-1)+f(1)+\frac{1}{3}(f'(-1)-f'(1)). \end{aligned}$$

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

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

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 3.6. 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

$$\int_0^1 f(x)\,\mathrm{d}x. \tag{3.5}$$

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

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 3.5 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 $y_1 < y_2 < \cdots < y_k$, on a l'estimation d'erreur

$$\forall x \in [y_1; y_k], \qquad |f(x) - p(x)| \leq \frac{\sup_{\xi \in [y_1; y_k]} |f^{(k)}(\xi)|}{k!} \prod_{j=1}^{k} (x - y_j).$$

En déduire une estimation de l'erreur de quadrature entre (3.5) 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 $Q(f)$.

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 $p(x) = d_0 + d_1 x + d_2 x^2 + d_3 x^3$) on a

$$\begin{array}{ccc} \int_0^1 p(x)\,\mathrm{d}x & = & Ap(0) + Bp(M) + Cp(1) \\ \| & & \| \\ \left[d_0 x + \frac{d_1}{2}x^2 + \frac{d_2}{3}x^3 + \frac{d_3}{4}x^4\right]_0^1 & & Ad_0 + Bd_0 + Cd_0 \\ \| & & \| \\ d_0 + \frac{d_1}{2} + \frac{d_2}{3} + \frac{d_3}{4} & & (A+B+C)d_0 + (BM+C)d_1 + (BM^2+C)d_2 + (BM^3+C)d_3 \end{array}$$

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 (3.5) 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

Algorithm 7 Calcul de $\int_0^1 f(x)\,\mathrm{d}x$

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

Un système linéaire de m équations à n inconnues est un ensemble de relations algébriques de la forme

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, \ldots, n$$

où les x_j sont les inconnues, les a_{ij} les coefficients du système et les b_i les composantes du second membre. Il est commode d'écrire ce système sous la forme matricielle $\mathbb{A}\mathbf{x} = \mathbf{b}$, où on a noté $\mathbb{A} = (a_{ij}) \in \mathbb{C}^{m\times n}$ la matrice des coefficients, $\mathbf{b} = (b_i) \in \mathbb{C}^m$ le vecteur du second membre et $\mathbf{x} = (x_i) \in \mathbb{C}^n$ le vecteur inconnu. Dans ces notes, nous ne traiterons que des systèmes carrés d'ordre n à coefficients réels, c'est-à-dire, lorsque $\mathbb{A} = (a_{i,j}) \in \mathbb{R}^{n\times n}$ 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 ;
2. $\mathrm{rg}(\mathbb{A}) = n$; [1]
3. le système homogène $\mathbb{A}\mathbf{x} = \mathbf{0}$ admet seulement la solution nulle.

La solution du système est donné — d'un point de vue théorique — par les formules de Cramer

$$x_j = \frac{\Delta_j}{\mathrm{Dét}(\mathbb{A})}, \qquad j = 1, \ldots, n$$

où Δ_j est le déterminant de la matrice obtenue en remplaçant la j-ième colonne de $\mathbb{A}$ par le second membre $\mathbf{b}$. Cette formule est cependant 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.

Conditionnement d'une matrice

Définition. Le conditionnement d'une matrice $\mathbb{A} \in \mathbb{C}^{n\times n}$ 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 norma ; ceci est signalé en introduisant un indice dans la notation.

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.

Cas particulier. Si $\mathbb{A}$ est symétrique et définie positive [2],

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

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

1. Le rang de $\mathbb{A}$, noté $\mathrm{rg}(\mathbb{A})$, est l'ordre maximum des déterminants extraits non nuls de $\mathbb{A}$. Une matrice est de rang maximum si $\mathrm{rg}(\mathbb{A}) = \min(m, n)$.
2. $\mathbb{A} \in \mathbb{R}^{n\times n}$ est
   - ▷ symétrique si $a_{ij} = a_{ji}$ pour tout $i, j = 1, \ldots, n$,
   - ▷ définie positive si pour tout vecteurs $\mathbf{x} \in \mathbb{R}^n$ avec $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^T\mathbb{A}\mathbf{x} > 0$.

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

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

Factorisation On factorise la matrice $\mathbb{A} \in \mathbb{R}^{n\times n}$ sous la forme d'un produit de deux matrices $\mathbb{LU}$ ainsi calculées

for $k = 1$ to $n-1$ **do**
  for $i = k+1$ to n **do**
    $\ell_{ik} \leftarrow \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$
    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

À la fin de la procédure les éléments de la matrice triangulaire supérieure $\mathbb{U}$ sont donné par $u_{ij} = a_{ij}$ pour $i = 1,\dots,n$ et $j = i,\dots,n$. Dans l'algorithme on n'a pas calculé les éléments diagonaux de $\mathbb{L}$ car ils sont tous égaux à 1 (souvent on stocke les éléments ℓ_{ij} pour $i = 1,\dots,n$ et $j = 1,\dots,i-1$ encore dans $\mathbb{A}$). Le coût de cette factorisation est de $\frac{2}{3}n^3$.

Triangulaire Une fois calculées les matrices $\mathbb{L}$ et $\mathbb{U}$, résoudre le système linéaire consiste simplement à résoudre successivement

1. le système triangulaire inférieur $\mathbb{L}\mathbf{y} = \mathbf{b}$ par l'algorithme

$$y_1 = \frac{b_1}{\ell_{11}}, \qquad y_i = \frac{1}{\ell_{11}}\left(b_i - \sum_{j=1}^{i-1} \ell_{ij} y_j\right), \qquad i = 2,\dots,n$$

2. le système triangulaire supérieure $\mathbb{U}\mathbf{x} = \mathbf{y}$ par l'algorithme

$$x_n = \frac{y_n}{u_{nn}}, \qquad x_i = \frac{1}{u_{nn}}\left(y_i - \sum_{j=i+1}^{n} u_{ij} x_j\right), \qquad j = n-1,\dots,1$$

Pivot. Dans cet algorithme les éléments $a_{kk}^{(k)}$, appelé *éléments pivotales*, doivent être différents de zéro. Si la matrice est inversible mais un élément pivotale 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 $\bar{r}$
    $\ell_{ik} \leftarrow \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$
    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. $\mathbb{PA} = \mathbb{LU}$), 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 $\mathbb{U}\mathbf{x} = \mathbf{y}$.

Si la matrice $\mathbb{A}$ est symétrique et définie positive ou si est à diagonale dominante [3] alors la technique du pivot n'est pas nécessaire.

Déterminant. La factorisation $\mathbb{LU}$ permet de calculer le déterminant de $\mathbb{A}$ en $O(n^3)$ car

$$\text{dét}(\mathbb{A}) = \text{dét}(\mathbb{L})\text{dét}(\mathbb{U}) = \prod_{k=1}^{n} u_{kk}.$$

3. $\mathbb{A} \in \mathbb{R}^{n\times n}$ 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),

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

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

1. Résoudre les systèmes linéaires par la méthode d'élimination 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 $\mathbb{A}^{-1}$.

SOLUTION.

1. Résolution par la méthode d'élimination de Gauss du premier système

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

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

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

donc

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

Résolution par la méthode d'élimination de Gauss du second système

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

$$\xrightarrow[L_4\leftarrow L_4-7L_2]{L_3\leftarrow L_3-2L_2} \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 10 \\ 0 & -1 & -2 & -7 & -10 \\ 0 & 0 & -4 & 4 & 0 \\ 0 & 0 & 4 & 36 & 40 \end{array}\right)$$

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

donc

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

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

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

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 $\mathbb{L}\mathbf{y}=\mathbf{b}$

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

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

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

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

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

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

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

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

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

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

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

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

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 construit une suite de vecteurs $\mathbf{x}^{(k)} \in \mathbb{R}^n$ qui converge vers la solution exacte $\mathbf{x}$ pour tout vecteur initiale $\mathbf{x}^{(0)} \in \mathbb{R}^n$.

Méthode de Jacobi.

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

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

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 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 la solution du système $(0,1,2)$.

Méthode de Gauss-Sidel. C'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. On améliore ainsi la vitesse de convergence.

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

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.

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 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 la solution du système $(0,1,2)$.

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 :

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

   ainsi

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

2. Méthode de Gauss-Seidel :

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

   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

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

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

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

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. 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.

1. Vérifier que la factorisation $\mathbb{LU}$ 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 $\mathbb{LU}$ 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 LU 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 $\mathbb{LU}$ il suffit de transformer la dernière ligne, ce qui requiert le calcul de l'unique multiplicateur $m = -\alpha/\beta$ 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 est calculé par la suite récurrente

$$\begin{cases} \delta_n = \alpha\delta_{n-1} - \alpha^{n-2}\beta^2, \\ \delta_1 = \alpha, \end{cases}$$

qui se réécrit directement

$$\delta_n = \alpha^n - (n-1)\alpha^{n-2}\beta^2.$$

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

$$\lambda_{1,2} = \alpha \pm \sqrt{(n-1)\beta}, \qquad \lambda_3 = \cdots = \lambda_n = \alpha.$$

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 4.3. 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 4.4. 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

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} + m_{i,k} a_{k,j}^{(k)}, \qquad i,j = k+1,$$
$$m_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}, \qquad i = k+1.$$

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

Exercice 4.5. 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 $\mathbb{A}\mathbb{B} = \mathbb{B}\mathbb{A} = \mathbb{I}$. Comme

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

   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 4.6. 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 $\mathbb{LU}$ 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 $\mathbb{LU}$:

$$\begin{pmatrix} \varepsilon & 1 \\ 2 & 1 \end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - \frac{2}{\varepsilon} L_1} \begin{pmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{2}{\varepsilon} \end{pmatrix}$$

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

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

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

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

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

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

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

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

Second système :

$$\begin{pmatrix} 2 & 1 \\ \varepsilon & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Factorisation $\mathbb{LU}$:

$$\begin{pmatrix} 2 & 1 \\ \varepsilon & 1 \end{pmatrix} \xrightarrow{L_2 \leftarrow L_2 - \frac{\varepsilon}{2} L_1} \begin{pmatrix} 2 & 1 \\ 0 & 1 - \frac{\varepsilon}{2} \end{pmatrix}$$

donc

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

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

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

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

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

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

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

$$\begin{pmatrix} 1 & 0 \\ \frac{\varepsilon}{2} & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \Longrightarrow \quad y_1 = 0, \quad y_2 = 1;$$
$$\begin{pmatrix} 2 & 1 \\ 0 & -\frac{2}{\varepsilon} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \Longrightarrow \quad b = -\frac{\varepsilon}{2}, \quad a = \frac{\varepsilon}{4}.$$

Exercice 4.7. Rappeler l'algorithme vu en cours pour calculer la décomposition $\mathbb{LU}$ 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 $\mathbb{LU}$ 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 $\mathbb{LU}$ 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

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

Deuxième système :

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

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

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

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

Troisième système :

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

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

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

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

Exercice 4.8. 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 :

$$\left(\begin{array}{cc|c} 10&1&11\\ 2&10&12 \end{array}\right) \xrightarrow{L_2\leftarrow L_2-\frac{2}{10}L_1} \left(\begin{array}{cc|c} 10&1&11\\ 0&\frac{49}{5}&\frac{49}{5} \end{array}\right) \Longrightarrow \begin{cases} 10a+b=11 \\ \frac{49}{5}b=\frac{49}{5} \end{cases} \Longrightarrow \begin{cases} a=1 \\ b=1. \end{cases}$$

▷ Second système :

$$\left(\begin{array}{cc|c} 2&10&12\\ 10&1&11 \end{array}\right) \xrightarrow{L_2\leftarrow L_2-\frac{10}{2}L_1} \left(\begin{array}{cc|c} 2&10&12\\ 0&-49&-49 \end{array}\right) \Longrightarrow \begin{cases} 2a+10b=12 \\ -49b=-49 \end{cases} \Longrightarrow \begin{cases} a=1 \\ b=1. \end{cases}$$

Jacobi ▷ Premier système :

$$\begin{cases} 10a+b=11 \\ 2a+10b=12 \end{cases} \Longleftrightarrow \begin{cases} a=\frac{11-b}{10} \\ b=\frac{12-2a}{10} \end{cases}$$

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 4.9. 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

$$\mathbb{L}=\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix} \qquad \mathbb{U}=\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}$$

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

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

$$\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}3\\-3\\6\end{pmatrix} \quad \Longrightarrow \quad x_3=6,\quad x_2=-7,\quad x_1=10.$$

2. Pour le seconde système on a

$$\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}6\\0\\-3\end{pmatrix} \quad\Longrightarrow\quad y_1=6,\quad y_2=-12,\quad y_3=27;$$
$$\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}6\\-12\\27\end{pmatrix} \quad\Longrightarrow\quad x_3=27,\quad x_2=-32,\quad x_1=35.$$

3. Pour le dernier système on a

$$\begin{pmatrix}1&0&0\\2&1&0\\3&4&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}0\\3\\6\end{pmatrix} \quad\Longrightarrow\quad y_1=0,\quad y_2=3,\quad y_3=-6;$$
$$\begin{pmatrix}1&-5&-7\\0&-3&-4\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}6\\-12\\27\end{pmatrix} \quad\Longrightarrow\quad x_3=-6,\quad x_2=7,\quad x_1=-7.$$

Exercice 4.10. Calculer, lorsqu'il est possible, la factorisation $\mathbb{LU}$ des matrices suivantes :

$$\mathbb{A}_1=\begin{pmatrix}1&2&3\\2&4&5\\7&8&9\end{pmatrix},\qquad \mathbb{A}_2=\begin{pmatrix}1&2&3\\7&8&9\\2&4&5\end{pmatrix}.$$

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

SOLUTION.

Matrice $\mathbb{A}_1$:

$$\mathbb{A}_1=\begin{pmatrix}1&2&3\\2&4&5\\7&8&9\end{pmatrix}\xrightarrow[L_3\leftarrow L_3-\frac{7}{1}L_1]{L_2\leftarrow L_2-\frac{2}{1}L_1}\begin{pmatrix}1&2&3\\0&0&-1\\0&-6&-12\end{pmatrix}$$

La factorisation $\mathbb{LU}$ 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 $\mathbb{A}_2$:

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

La factorisation $\mathbb{LU}$ de la matrice $\mathbb{A}_2$ 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 $\mathbb{LU}$ 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 $\mathbb{A}_1$ cela aurait donné

$$\mathbb{A}_1=\begin{pmatrix}1&2&3\\2&4&5\\7&8&9\end{pmatrix}\xrightarrow{L_2\leftrightarrow L_3}\begin{pmatrix}1&2&3\\7&8&9\\2&4&5\end{pmatrix}\xrightarrow[L_3\leftarrow L_3-\frac{2}{1}L_1]{L_2\leftarrow L_2-\frac{7}{1}L_1}\begin{pmatrix}1&2&3\\0&-6&-12\\0&0&-1\end{pmatrix}.$$

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 $\mathbb{L}\mathbf{x} = \mathbf{y}$. Dans notre exemple cela donne

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

Exercice 4.11. 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 $\mathbb{L}\mathbb{U}$ avec

- ▷ $\ell_{ii} = 1$ pour $i = 1, \ldots, n$,
- ▷ $\ell_{ij} = 0$ 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 $\mathbb{L}\mathbb{U}$ de la matrice $\mathbb{A}$:

$$\begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 1 & 1 & 1 & \dots & 1 & 1 \end{pmatrix} \xrightarrow{L_n \leftarrow L_n - \frac{1}{1} L_1} \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 2 & 2 & \dots & 2 & 2 \end{pmatrix}$$

$$\xrightarrow{L_n \leftarrow L_n - \frac{2}{1} L_2} \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 0 & 4 & \dots & 4 & 4 \end{pmatrix}$$

$$[\ldots]$$

$$\xrightarrow{L_n \leftarrow L_n - \frac{2^{n-2}}{1} L_{n-1}} \begin{pmatrix} 1 & -1 & \dots & \dots & \dots & -1 \\ 0 & 1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 2^{n-1} \end{pmatrix}$$

par conséquent on obtient les matrices

$$\mathbb{L} = \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \ddots & & \vdots \\ 0 & 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 & 0 \\ 1 & 2 & 4 & \dots & 2^{n-2} & 1 \end{pmatrix} \quad \text{et} \quad \mathbb{U} = \begin{pmatrix} 1 & -1 & -1 & \dots & \dots & -1 \\ 0 & 1 & -1 & \ddots & & \vdots \\ 0 & 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & -1 & -1 \\ 0 & 0 & \dots & 0 & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 2^{n-1} \end{pmatrix}.$$

Exercice 4.12. Soit les systèmes linéaires

$$\begin{cases} 4x_1 + 3x_2 + 3x_3 = 10 \\ 3x_1 + 4x_2 + 3x_3 = 10 \\ 3x_1 + 3x_2 + 4x_3 = 10 \end{cases} \tag{4.1}$$

$$\begin{cases} 4x_1 + x_2 + x_3 = 6 \\ x_1 + 4x_2 + x_3 = 6 \\ x_1 + x_2 + 4x_3 = 6 \end{cases} \tag{4.2}$$

4.12.1 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 ?

4.12.2 Écrire les méthodes de Jacobi et de Gauss-Seidel pour ces deux systèmes linéaires en remplissant le tableau suivant :

	$\mathbb{A}_1\mathbf{x} = \mathbf{b}$	$\mathbb{A}_2\mathbf{x} = \mathbf{b}$
Jacobi	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} \\ \\ \end{pmatrix}$	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} \\ \\ \end{pmatrix}$
Gauss-Seidel	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} \\ \\ \end{pmatrix}$	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} \\ \\ \end{pmatrix}$

4.12.3 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) :

4.12.3.1 avec la méthode de Jacobi pour le système (4.1),

4.12.3.2 avec la méthode de Gauss-Seidel pour le système (4.1),

4.12.3.3 avec la méthode de Jacobi pour le système (4.2),

4.12.3.4 avec la méthode de Gauss-Seidel pour le système (4.2).

4.12.4 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 $\mathbb{A}\mathbf{x} = \mathbf{b}$:

$$\underbrace{\begin{pmatrix} 4 & 3 & 3 \\ 3 & 4 & 3 \\ 3 & 3 & 4 \end{pmatrix}}_{\mathbb{A}_1} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 10 \end{pmatrix} \qquad \text{et} \qquad \underbrace{\begin{pmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{pmatrix}}_{\mathbb{A}_2} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 6 \\ 6 \end{pmatrix}$$

4.12.1 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 [4] sont $\lambda_1 = \lambda_2 = 1$ et $\lambda_3 = 10$) : la méthode de Gauss-Seidel converge.

4. $\det \mathbb{A}_1(\lambda) = (4-\lambda)^3 + 27 + 27 - 9(4-\lambda) - 9(4-\lambda) - 9(4-\lambda) = 64 - 48\lambda + 12\lambda^2 - \lambda^3 + 54 - 108 + 27\lambda = -\lambda^3 + 12\lambda^2 - 21\lambda + 10$. Une racine évidente est $\lambda = 1$ et on obtient $\det \mathbb{A}_1(\lambda) = (\lambda-1)(-\lambda^2 + 11\lambda - 10) = (\lambda-1)^2(\lambda-10)$.

4.12.2 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	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 10-3x_2^{(k)}-3x_3^{(k)} \\ 10-3x_1^{(k+1)}-3x_3^{(k)} \\ 10-3x_1^{(k+1)}-3x_2^{(k+1)} \end{pmatrix}$	$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ x_3^{(k+1)} \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 6-x_2^{(k)}-x_3^{(k)} \\ 6-x_1^{(k+1)}-x_3^{(k)} \\ 6-x_1^{(k+1)}-x_2^{(k+1)} \end{pmatrix}$

4.12.3 On obtient les suites suivantes

4.12.3.1 Jacobi pour le système (4.1) :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4}\begin{pmatrix} 10-3\times 0-3\times 0 \\ 10-3\times 0-3\times 0 \\ 10-3\times 0-3\times 0 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{2} \\ \frac{5}{2} \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4}\begin{pmatrix} 10-3\times \frac{5}{2}-3\times \frac{5}{2} \\ 10-3\times \frac{5}{2}-3\times \frac{5}{2} \\ 10-3\times \frac{5}{2}-3\times \frac{5}{2} \end{pmatrix} = \begin{pmatrix} -\frac{5}{4} \\ -\frac{5}{4} \\ -\frac{5}{4} \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \frac{1}{4}\begin{pmatrix} 10-3\times \frac{-5}{4}-3\times \frac{-5}{4} \\ 10-3\times \frac{-5}{4}-3\times \frac{-5}{4} \\ 10-3\times \frac{-5}{4}-3\times \frac{-5}{4} \end{pmatrix} = \begin{pmatrix} \frac{35}{8} \\ \frac{35}{8} \\ \frac{35}{8} \end{pmatrix}$$

4.12.3.2 Gauss-Seidel pour le système (4.1) :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4}\begin{pmatrix} 10-3\times 0-3\times 0 \\ 10-3\times \frac{5}{2}-3\times 0 \\ 10-3\times \frac{5}{2}-3\times \frac{5}{8} \end{pmatrix} = \begin{pmatrix} \frac{5}{2} \\ \frac{5}{8} \\ \frac{5}{32} \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4}\begin{pmatrix} 10-3\times \frac{5}{8}-3\times \frac{5}{32} \\ 10-3\times \frac{245}{128}-3\times \frac{5}{32} \\ 10-3\times \frac{245}{128}-3\times \frac{485}{512} \end{pmatrix} = \begin{pmatrix} \frac{245}{128} \\ \frac{485}{512} \\ \frac{725}{2048} \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \begin{pmatrix} \frac{12485}{8192} \\ \frac{35765}{32768} \\ \frac{70565}{131072} \end{pmatrix}$$

4.12.3.3 Jacobi pour le système (4.2) :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4}\begin{pmatrix} 6-1\times 0-1\times 0 \\ 6-1\times 0-1\times 0 \\ 6-1\times 0-1\times 0 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ \frac{3}{2} \\ \frac{3}{2} \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4}\begin{pmatrix} 6-1\times \frac{3}{2}-1\times \frac{3}{2} \\ 6-1\times \frac{3}{2}-1\times \frac{3}{2} \\ 6-1\times \frac{3}{2}-1\times \frac{3}{2} \end{pmatrix} = \begin{pmatrix} \frac{3}{4} \\ \frac{3}{4} \\ \frac{3}{4} \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \frac{1}{4}\begin{pmatrix} 6-1\times \frac{3}{4}-1\times \frac{3}{4} \\ 6-1\times \frac{3}{4}-1\times \frac{3}{4} \\ 6-1\times \frac{3}{4}-1\times \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{9}{8} \\ \frac{9}{8} \\ \frac{9}{8} \end{pmatrix}$$

4.12.3.4 Gauss-Seidel pour le système (4.2) :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(1)} = \frac{1}{4}\begin{pmatrix} 6-1\times 0-1\times 0 \\ 6-1\times \frac{3}{2}-1\times 0 \\ 6-1\times \frac{3}{2}-1\times \frac{9}{8} \end{pmatrix}\begin{pmatrix} \frac{3}{2} \\ \frac{9}{8} \\ \frac{27}{32} \end{pmatrix}$$

$$\Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(2)} = \frac{1}{4}\begin{pmatrix} 6-1\times \frac{9}{8}-1\times \frac{27}{32} \\ 6-1\times \frac{129}{128}-1\times \frac{27}{32} \\ 6-1\times \frac{129}{128}-1\times \frac{531}{512} \end{pmatrix} = \begin{pmatrix} \frac{129}{128} \\ \frac{531}{512} \\ \frac{2025}{2048} \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(3)} = \frac{1}{4}\begin{pmatrix} 6-1\times \frac{531}{512}-1\times \frac{2025}{2048} \\ 6-1\times \frac{8139}{8192}-1\times \frac{2025}{2048} \\ 6-1\times \frac{8139}{8192}-1\times \frac{32913}{32768} \end{pmatrix} = \begin{pmatrix} \frac{8139}{8192} \\ \frac{32913}{32768} \\ \frac{131139}{131072} \end{pmatrix}$$

4.12.4 Calcul de la solution exacte à l'aide de la méthode d'élimination de Gauss :

▷ Système (4.1) :

$$\left(\begin{array}{ccc|c} 4 & 3 & 3 & 10 \\ 3 & 4 & 3 & 10 \\ 3 & 3 & 4 & 10 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 - \frac{3}{4} L_1]{L_2 \leftarrow L_2 - \frac{3}{4} L_1} \left(\begin{array}{ccc|c} 4 & 3 & 3 & 10 \\ 0 & 7/4 & 3/4 & 5/2 \\ 0 & 3/4 & 7/4 & 5/2 \end{array}\right) \xrightarrow{L_3 \leftarrow L_3 - \frac{3/4}{7/4} L_2} \left(\begin{array}{ccc|c} 4 & 3 & 3 & 10 \\ 0 & 7/4 & 3/4 & 5/2 \\ 0 & 0 & 10/7 & 10/7 \end{array}\right) \Longrightarrow \mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

▷ Système (4.2) :

$$\left(\begin{array}{ccc|c} 4 & 1 & 1 & 6 \\ 1 & 4 & 1 & 6 \\ 1 & 1 & 4 & 6 \end{array}\right) \xrightarrow[L_3 \leftarrow L_3 - \frac{1}{4} L_1]{L_2 \leftarrow L_2 - \frac{1}{4} L_1} \left(\begin{array}{ccc|c} 4 & 1 & 1 & 6 \\ 0 & 15/4 & 3/4 & 9/2 \\ 0 & 3/4 & 15/4 & 9/2 \end{array}\right) \xrightarrow{L_3 \leftarrow L_3 - \frac{3/4}{15/4} L_2} \left(\begin{array}{ccc|c} 4 & 1 & 1 & 6 \\ 0 & 15/4 & 3/4 & 9/2 \\ 0 & 0 & 18/5 & 18/5 \end{array}\right) \Longrightarrow \mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

5 Équations différentielles ordinaires

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.

Le problème de Cauchy

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, de la forme suivante :

Problème de Cauchy. Trouver $y\colon I \subset \mathbb{R} \to \mathbb{R}$ tel que

$$\begin{cases} y'(t) = f(t, y(t)), & \forall\, t \in I, \\ y(t_0) = y_0, \end{cases}$$

où $f\colon I \times \mathbb{R} \to \mathbb{R}$ est une fonction donnée et y est la dérivée de y par rapport à t. Enfin, t_0 est un point de I et y_0 une valeur appelée donnée initiale.

On rappelle dans la proposition suivante un résultat classique d'analyse.

Proposition. On suppose que 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
$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|, \qquad \forall\, t \in I, \quad \forall\, y_1, y_2 \in \mathbb{R}.$$

Alors la solution $y = y(t)$ du problème de Cauchy existe, est unique et appartient à $\mathscr{C}^1(I)$.

Malheureusement, 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.

Le principe de toutes ces méthodes est de subdiviser l'intervalle $I = [t_0, T]$, avec $T < +\infty$, en N_h intervalles de longueur $h = (T - t_0)/N_h$; h est appelé le pas de discrétisation. Alors, pour chaque nœud $t_n = t_0 + nh$ $(1 \leq n \leq N_h)$ on cherche la valeur inconnue u_n qui approche $y_n = y(t_n)$. L'ensemble des valeurs $\{u_0 = y_0, u_1, \ldots, u_{N_h}\}$ représente la solution numérique.

Exemples

Rappelons ici trois schémas numériques : la méthode d'Euler explicite, la méthode de Heun (ou schéma de Runge-Kutta d'ordre 2) et la méthode de Runge-Kutta classique, d'ordre 4.

Soit $h = t_{i+1} - t_i$; alors on a

Euler explicite

$$\begin{cases} y_0 = \eta, \\ y_{i+1} = y_i + hf(t_i, y_i), & i = 0, \ldots, n-1; \end{cases} \tag{5.1}$$

Euler implicite

$$\begin{cases} y_0 = \eta, \\ y_{i+1} = y_i + hf(t_i + h, y_{i+1}), \quad i = 0, \dots, n-1; \end{cases} \tag{5.2}$$

Trapèze ou Crank-Nicholson

$$\begin{cases} y_0 = \eta, \\ y_{i+1} = y_i + \frac{h}{2}\left(f(t_i, y_i) + f(t_{i+1}, y_{i+1})\right), \quad i = 0, \dots, n-1; \end{cases} \tag{5.3}$$

Heun

$$\begin{cases} y_0 = \eta, \\ y_{i+1} = y_i + \frac{h}{2}\left(f(t_i, y_i) + f(t_{i+1}, y_i + hf(t_i, y_i))\right), \quad i = 0, \dots, n-1; \end{cases} \tag{5.4}$$

RK4

$$\begin{cases} y_0 = \eta, \\ y_{i+1} = y_i + \frac{h}{6}\left(f(t_i, y_{i,1}) + 2f(t_i + h/2, y_{i,2}) + 2f(t_i + h/2, y_{i,3}) + f(t_{i+1}, y_{i,4})\right), \quad i = 0, \dots, n-1; \end{cases} \tag{5.5}$$

avec

$$\begin{aligned} y_{i,1} &= y_i \\ y_{i,2} &= y_i + \frac{h}{2} f(t_i, y_{i,1}) \\ y_{i,3} &= y_i + \frac{h}{2} f(t_i + h/2, y_{i,2}) \\ y_{i,4} &= y_i + hf(t_i + h/2, y_{i,3}) \end{aligned}$$

Stabilité absolue (A-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». La propriété

$$\lim_{n \to +\infty} u_n = 0$$

est appelée stabilité absolue.

Exercice 5.1. On considère le problème de Cauchy

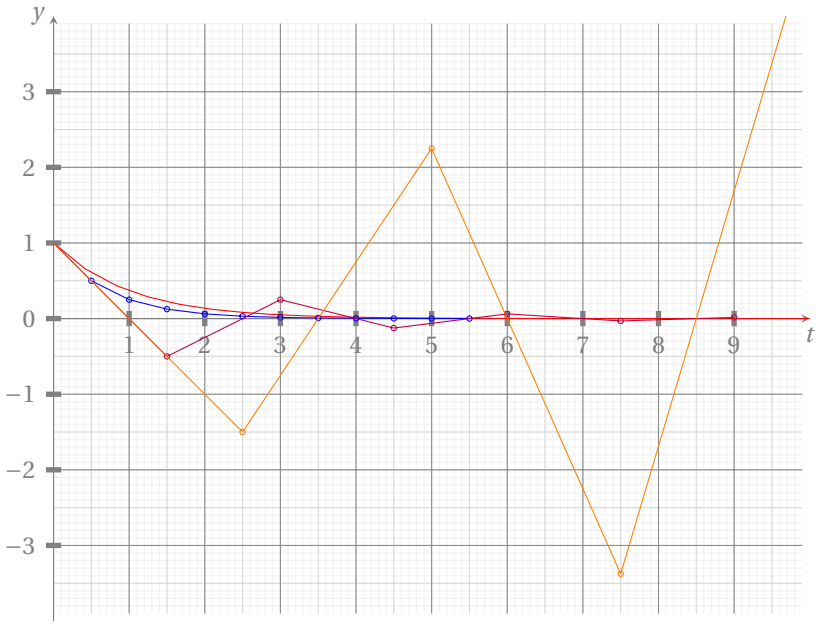
$$\begin{cases} y'(t) = -y(t), \\ y(0) = 1, \end{cases}$$

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 dessiner sur le plan suivant 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 est $y(t) \equiv 0$, toutes les autres solutions sont du type $y(t) = Ce^{-t}$. Donc l'unique solution du problème de Cauchy est $y(t) = e^{-t}$ 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 $t + \Delta t$ de 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 $y_k = \left(\frac{1}{2}\right)^k$.



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 reste 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 que une telle suite
   - ▷ diverge si $|q| > 1$ ou $q = -1$,
   - ▷ est stationnaire si $q = 1$,
   - ▷ converge vers 0 $|q| < 1$.

Exercice 5.2. Soit le problème de Cauchy :

$$\begin{cases} u'(t) + 10u(t) = 0, \quad \forall t \in \mathbb{R}, \\ u(0) = u_0 > 0. \end{cases} \tag{5.6}$$

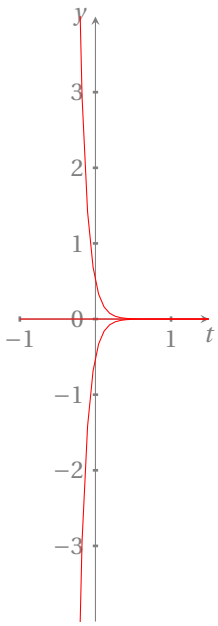
1. Montrer qu'il existe une unique solution globale $u \in \mathscr{C}^\infty(\mathbb{R},\mathbb{R})$ que vous préciserez explicitement.
2. Soit le schéma numérique de Cranck-Nicholson défini par la suite $\{u_n\}_{n\in\mathbb{N}}$ vérifiant
$$\frac{u_{n+1}-u_n}{\Delta t} + 5(u_{n+1}+u_n) = 0, \qquad \forall n \in \mathbb{N},$$
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 pour tout $\Delta t > 0$
$$|r| < 1.$$
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 $T - \Delta t < n^*\Delta t \leq T$. Montrer que
$$\lim_{\Delta t \to 0} u_{n^*} = u_0 e^{-10T}.$$
7. Soit $\{v_n\}_{n\in\mathbb{N}}$ la suite définissant le schéma d'Euler explicite pour l'équation différentielle (5.6). Montrer que
$$\lim_{\Delta t \to 0} v_{n^*} = u_0 e^{-10T}.$$
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

$$\begin{cases} u'(t) = f(t,u(t)), \quad \forall t \in \mathbb{R}, \\ u(0) = u_0 > 0. \end{cases} \tag{5.7}$$

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 et $u \in \mathscr{C}^\infty([-T,T],\mathbb{R})$. 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 $u(t) = u_0 e^{-10t}$ 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

$$\lim_{x\to0}\frac{\ln(1+\alpha x)}{\alpha x}=1$$

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

$$\lim_{\Delta t\to0}u_{n^*}=u_0\lim_{\Delta t\to0}\left(\frac{1-5\Delta t}{1+5\Delta t}\right)^{n^*}=u_0e^{-10T}.$$

7. La suite définissant le schéma d'Euler explicite pour l'EDO assignée s'écrit

$$\frac{v_{n+1}-v_n}{\Delta t}=f(t_n,u_n)$$

i.e.

$$v_{n+1}=v_n-10\Delta t\,v_n=(1-10\Delta t)\,v_n=(1-10\Delta t)^{n+1}v_0.$$

Il s'agit à nouveau d'une suite géométrique de raison

$$r_e=1-10\Delta t$$

qui converge ssi $|r_e|<1$, *i.e.* ssi $\Delta t<0{,}2$ (le schéma d'Euler pour cette EDO est conditionnellement stable).

Soit $T>0$ fixé et considérons $n^*=n^*(\Delta t)$ tel que $T-\Delta t<n^*\Delta t\leq T$. Alors

$$\begin{array}{ccccc}
(1-10\Delta t)^{\frac{T}{\Delta t}-1} & \leq & (1-10\Delta t)^{n^*} & \leq & (1-10\Delta t)^{\frac{T}{\Delta t}} \\
\| & & & & \| \\
e^{(T-\Delta t)\frac{\ln(1-10\Delta t)}{\Delta t}} & & & & e^{T\frac{\ln(1-10\Delta t)}{\Delta t}} \\
\| & & & & \| \\
e^{-10(T-\Delta t)\frac{\ln(1-10\Delta t)}{-10\Delta t}} & & & & e^{-10T\frac{\ln(1-10\Delta t)}{-10\Delta t}} \\
\downarrow & & & & \downarrow \\
e^{-10T} & & & & e^{-10T}
\end{array}$$

d'où

$$\lim_{\Delta t\to 0} v_{n^*}=u_0\lim_{\Delta t\to 0}(1-10\Delta t)^{\frac{T}{\Delta t}}=u_0e^{-10T}.$$

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 5.3. Soit f la fonction définie par $f(x)=x\ln(1+x)$ et considérons l'équation différentielle ordinaire

$$\begin{cases} u'(t)=-f(u(t)), & \forall\, t\geq 0,\\ u(0)=u_0>0. \end{cases}$$

1. Vérifier que $f\in C^\infty(\mathbb{R}^+,\mathbb{R}^+)$ et est convexe ($f''\geq 0$).
2. Montrer qu'il existe $T>0$ et une unique solution $u\in C^\infty([0,T],\mathbb{R})$.
3. Montrer que pour tout $t\in[0,T]$, $u(t)>0$. En déduire que la solution u admet un prolongement sur $\mathbb{R}^+$ solution de l'équation différentielle et est bornée.
4. Écrire le schéma d'Euler implicite pour cette équation différentielle. On appellera u_n une approximation de $u(t_n)$, $t_n=n\Delta t$ pour $\Delta t>0$ fixé.
5. En se servant des résultats de la première partie, montrer que le schéma est bien défini, c'est à dire que si u_n est connu, u_{n+1} est bien défini et de façon unique. Quelle méthode numérique suggère la première partie pour déterminer u_{n+1} en fonction de u_n ? L'algorithme de détermination de u_{n+1} est-il convergent pour tout Δt ?
6. Montrer que pour cette équation, le schéma d'Euler implicite est A-stable pour tout Δt.
7. Calculer l'ordre de ce schéma.
8. Écrire l'algorithme complet de la méthode introduite pour approcher la solution de l'équation différentielle.
9. Donner sans le justifier un schéma du même ordre, inconditionnellement A-stable et plus simple à implémenter que le schéma d'Euler implicite.

SOLUTION.

1. On vérifie que $f\in C^\infty(\mathbb{R}^+,\mathbb{R}^+)$ (f positive car $\ln(1+x)\geq\ln(1)=0$ pour $x\geq 0$). Calcul de f'' trivial et on a $f''\geq 0$. De plus $f(0)=0$.
2. Comme f est $C^1(\mathbb{R}^+)$, d'après Cauchy-Lipschitz, il existe $T>0$ et une unique solution $u\in C^1([0,T],\mathbb{R}^+)$. En relisant l'équation différentielle, la composée de f et u est donc $C^1(\mathbb{R}^+)$ et ainsi $u\in C^2([0,T],\mathbb{R}^+)$. Par récurrence, en exploitant l'équation et la régularité de f, on grimpe en régularité sur u et $u\in C^\infty([0,T],\mathbb{R}^+)$.
3. La fonction nulle est solution de l'équation différentielle ($f(0)=0$), par l'unicité de la solution du problème de Cauchy, pour tout $t\in[0,T]$, $u(t)>0$ si $u_0>0$ (autrement dit : 2 trajectoires ne peuvent pas se croiser). De plus u est décroissante car $u'\leq 0$, 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'équation différentielle et est bornée.
4. Schéma d'Euler implicite pour cette équation différentielle. On appellera u_n une approximation de $u(t_n)$, $t_n=n\Delta t$ pour $\Delta t>0$ fixé.
$$\frac{u_{n+1}-u_n}{\Delta t}=-f(u_{n+1}).$$
5. Le schéma est bien défini, pour chaque $u_n\geq 0$, il existe un unique $u_{n+1}\in[0,u_n]$ qui est le zéro de la fonction g définie ci-dessus avec $u=u_n$ et $\delta=\Delta t$. La méthode numérique suggérée par la première partie pour déterminer u_{n+1} en fonction de u_n est l'algorithme de Newton qu'on a montré convergent indépendamment de $\delta=\Delta t$.
6. Montrer que pour cette équation, le schéma d'Euler implicite est A-stable pour tout Δt.
   $u_{n+1}\in[0,u_n]$, donc la suite est décroissante et positive, elle est donc bornée.
7. Calculer l'ordre de ce schéma. Cf cours et TD, le schéma est d'ordre 1.

8. Algorithme complet de la méthode introduite pour approcher la solution de l'équation différentielle : il faut écrire la boucle des itérations en n et à chaque itération il faut appeler l'algorithme de Newton pour déterminer u_{n+1} en fonction de u_n. Il s'agit là d'un algo itératif (encore une boucle) dont le critère d'arrêt portera par exemple sur l'écart de 2 itérés successifs. Voir TP pour ses 2 algos.
9. On prend un schéma semi-implicite qui ne nécessite pas l'emploi de la méthode de Newton :

$$\frac{u_{n+1}-u_n}{\Delta t} = -u_{n+1}\ln(1+u_n).$$

Ainsi

$$u_{n+1} = \frac{1}{1+\Delta t\ln(1+u_n)}u_n.$$

On a une formule de récurrence rendue explicite par un calcul élémentaire, la fraction positive et plus petite que 1 assure que la suite est positive décroissante, d'où la stabilité.

Exercice 5.4. 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

$$\begin{cases} T'(t) = K(T(t)-T_{\text{ext}}), \\ T(0) = T_0. \end{cases}$$

5.4.1 Soit Δt le pas temporel. Écrire le schéma d'Euler implicite pour approcher la solution de cette équation différentielle.
5.4.2 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 ?
5.4.3 *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.
5.4.4 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.

5.4.1 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

$$T_{n+1} = T_n + K\Delta t(T_{n+1} - T_{\text{ext}}).$$

5.4.2 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.

5.4.3 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 $t_0 = t_n - n\Delta t = 2H20 - 1H20 = 01H00$.

5.4.4 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 :

$$T_0 = T(0) = De^{K\cdot 0} \implies D = -T_0 \implies T(t) = T_0 e^{Kt}$$

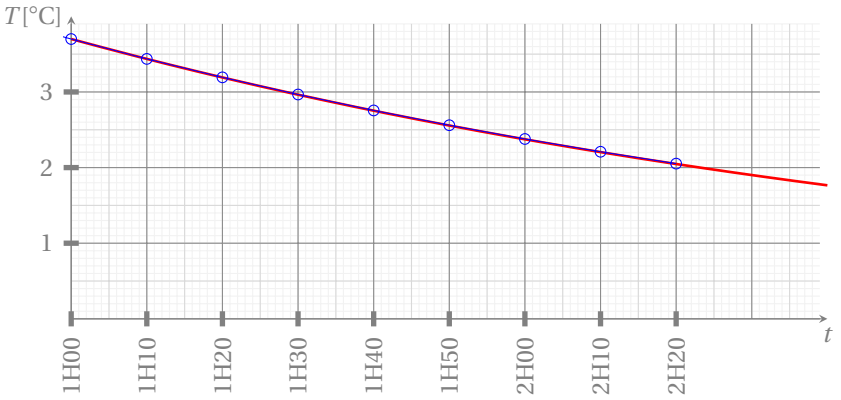
Ici $T_0 = 37°\text{C}$ donc la température du cadavre suit la loi

$$T(t) = 37e^{Kt}.$$

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 5.5. 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

$$\begin{cases} I'(t) = f(t, I(t)), \quad \forall t \in \mathbb{R}^+, \\ I(0) = I_0 > 0, \end{cases} \tag{5.8}$$

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

   $$I_{n+1} \leq A \iff (1 + kA\Delta t) I_n \leq (1 + kI_n\Delta t) A \iff I_n \leq A$$

   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 $I(t) \equiv A$.

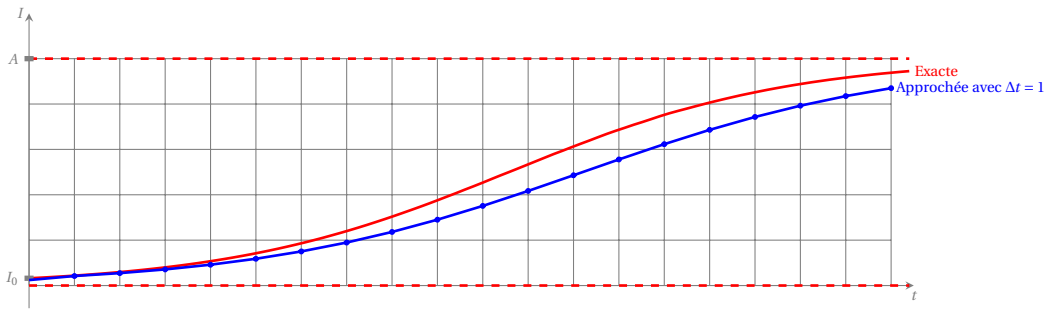
Pour chercher toutes les solutions non constantes on remarque qu'il s'agit d'une EDO à variables séparables donc on a

$$I(t) = \frac{A}{De^{-Akt} + 1}$$

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 5.6 (Café ☕)**.** Considérons une tasse de café à la température de 75 dans une salle à 25. 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'est-à-dire

$$T(5) = 50.$$

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 on a

$$\begin{aligned}
T'(t) &= K(T(t) - 25) \\
\frac{T'(t)}{(T(t) - 25)} &= K \\
\frac{\mathrm{d}\,T}{(T - 25)} &= K\mathrm{d}\,t \\
\int \frac{1}{(T - 25)}\mathrm{d}\,T &= K\int \mathrm{d}\,t \\
\ln(T - 25) &= Kt + c \\
T - 25 &= De^{Kt} \\
T(t) &= 25 + De^{Kt}
\end{aligned}$$

2. La valeur numérique de la constante d'intégration D est obtenue grâce à la CI :

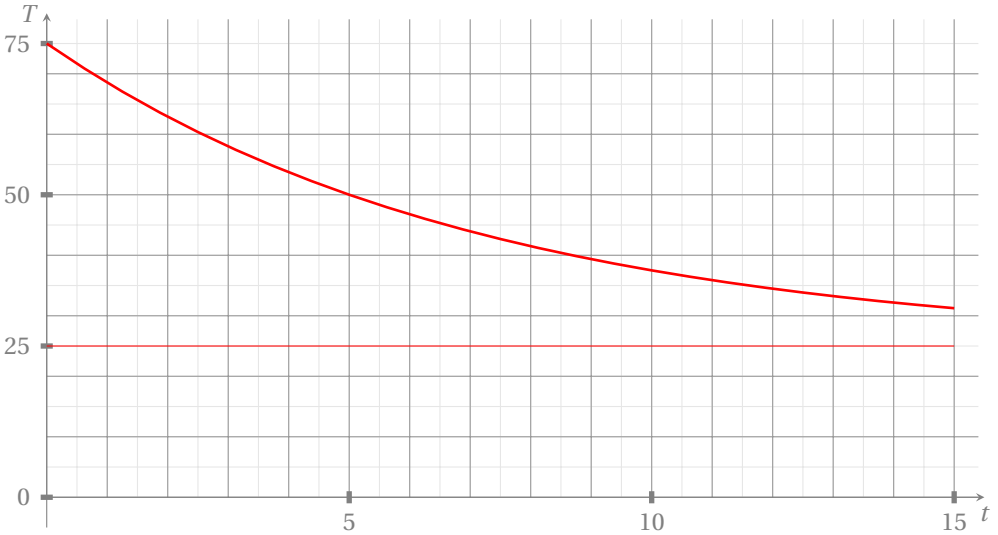
$$\begin{aligned}
75 = T(0) &= 25 + De^{K\cdot 0} \\
D &= 50 \\
T(t) &= 25 + 50e^{Kt}
\end{aligned}$$

3. Il ne reste qu'à établir la valeur numérique de la constante de refroidissement K grâce à l'«indice» :

$$\begin{aligned}
50 = T(5) &= 25 + 50e^{Kt} \\
K &= -\frac{\ln(2)}{5} \\
T(t) &= 25 + 50e^{-\frac{\ln(2)}{5}t}
\end{aligned}$$

On peut donc conclure que la température du café évolue selon la fonction

$$T(t) = 25 + 50e^{-\frac{\ln(2)}{5}t}.$$



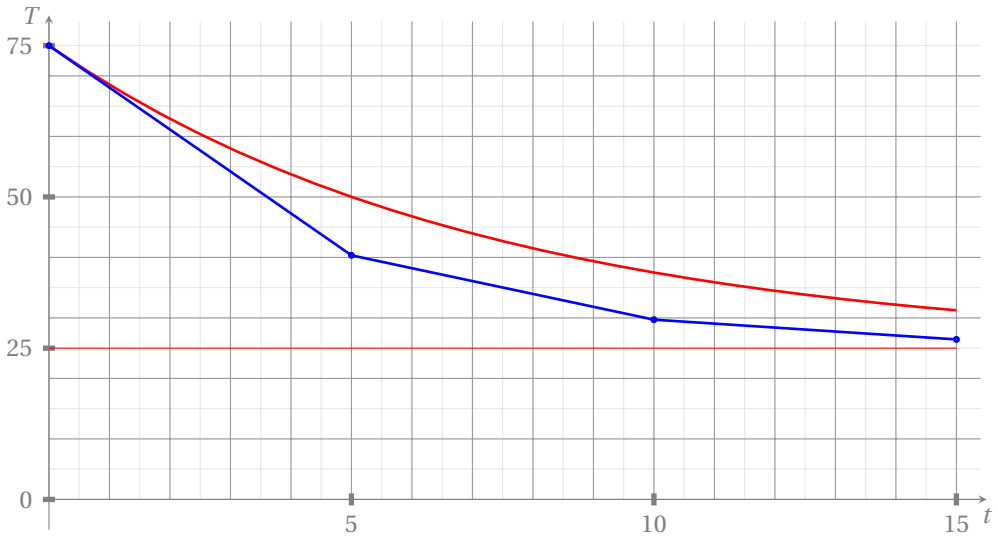
Solution approchée par la méthode d'Euler Supposons de connaître K mais de ne pas vouloir/pouvoir calculer la fonction $T(t)$. Grâce à la méthode d'Euler on peut estimer la température à différentes instantes t_i en faisant une discrétisation temporelle du futur (i.e. on construit une suite de valeurs $\{t_i = 0 + i\Delta t\}_i$) 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

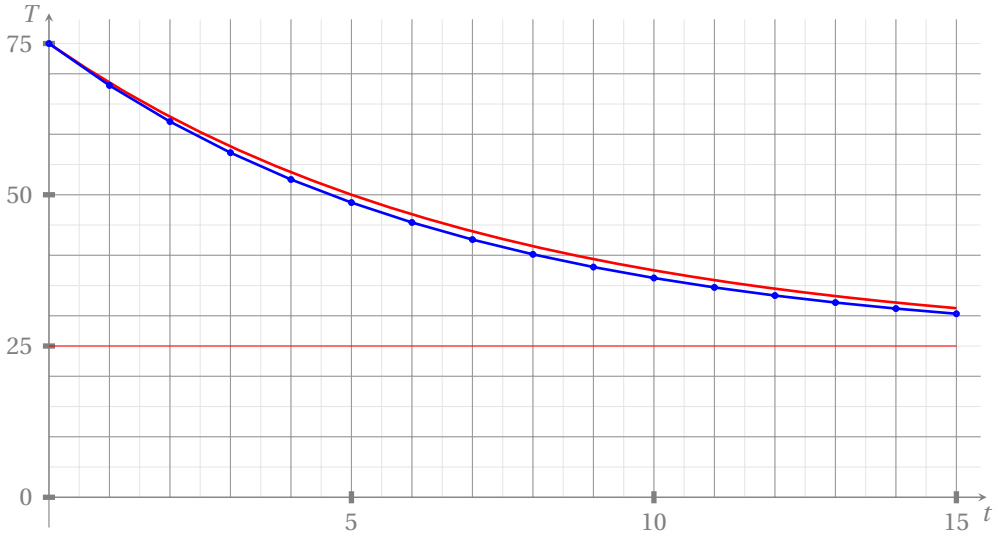
$$\begin{cases} T_{i+1} = (1 - \frac{\ln(2)}{5}\Delta t)\,T_i + 5\ln(2)\Delta t, \\ T_0 = 75. \end{cases}$$

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 5.7. 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 $t \in \mathbb{R}^+$. On peut alors écrire :

$$p'(t) = kp(t) - hp^2(t)$$

$$\begin{aligned}
\frac{p'(t)}{kp(t)-hp^2(t)} &= 1\\
\frac{\mathrm{d}\,p}{kp-hp^2} &= 1\mathrm{d}\,t\\
\int \frac{1}{p(k-hp)}\mathrm{d}\,p &= \int \mathrm{d}\,t\\
\frac{1}{k}\int \frac{1}{p}\mathrm{d}\,p - \frac{1}{k}\int \frac{-h}{k-hp}\mathrm{d}\,p &= \int \mathrm{d}\,t\\
\frac{1}{k}\ln(p) - \frac{1}{k}\ln(k-hp) &= t+c\\
\ln\left(\frac{p}{k-hp}\right) &= kt+kc\\
\frac{p}{k-hp} &= De^{kt}\\
p(t) &= \frac{k}{\frac{1}{De^{kt}}+h}.
\end{aligned}$$

2. La valeur numérique de la constante d'intégration D est obtenue grâce à la CI :

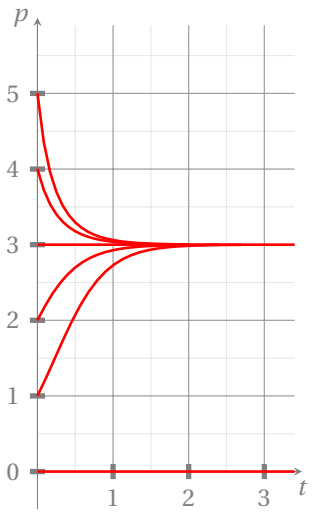
$$\begin{aligned}
p_0 = p(0) &= \frac{kD}{1+hDe^{0k}}\\
D &= \frac{p_0}{k-hp_0}.
\end{aligned}$$

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.

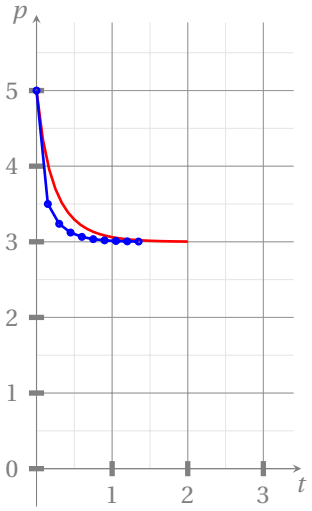
Solution approchée Supposons de ne pas vouloir/pouvoir calculer la fonction $p(t)$. Grâce à la méthode d'Euler on peut estimer le nombre d'indivus à différentes instantes t_i en faisant une discrétisation temporelle du futur (i.e. on construit une suite de valeurs $\{t_i = 0+i\Delta t\}_i$) et en construisant 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 :

$$\begin{cases} p_{i+1} = p_i + \Delta t\ p_i(k-hp_i),\\ p_0 \text{ donné,}\end{cases}$$

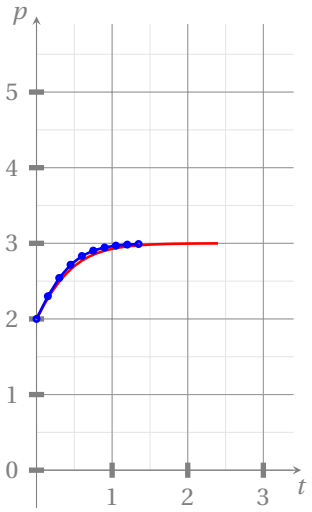
qu'on peut réécrire comme

$$\begin{cases} p_{i+1} = (1 + k\Delta t - h\Delta t\, p_i)\, p_i, \\ p_0 \text{ donné.} \end{cases}$$

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	$p(t_i) - p_i$
0.000000	5.000000	5.000000	0.000000
0.150000	4.027123	3.500000	0.527123
0.300000	3.582637	3.237500	0.345137
0.450000	3.347079	3.122164	0.224915
0.600000	3.212403	3.064952	0.147451
0.750000	3.132046	3.035091	0.096956
0.900000	3.082874	3.019115	0.063759
1.050000	3.052319	3.010459	0.041861
1.200000	3.033151	3.005736	0.027415
1.350000	3.021054	3.003150	0.017904
1.500000	3.013390	3.001731	0.011659
1.650000	3.008524	3.000952	0.007573
1.800000	3.005430	3.000523	0.004907



t_i	$p(t_i)$	p_i	$p(t_i) - p_i$
0.000000	2.000000	2.000000	0.000000
0.150000	2.274771	2.300000	−0.025229
0.300000	2.493175	2.541500	−0.048325
0.450000	2.655760	2.716292	−0.060532
0.600000	2.770980	2.831887	−0.060907
0.750000	2.849816	2.903298	−0.053483
0.900000	2.902469	2.945411	−0.042942
1.050000	2.937070	2.969529	−0.032459
1.200000	2.959567	2.983102	−0.023535
1.350000	2.974092	2.990663	−0.016571
1.500000	2.983429	2.994852	−0.011423
1.650000	2.989412	2.997164	−0.007752
1.800000	2.993240	2.998439	−0.005199