Université Moulay ISMAÏL
Faculté des Sciences de Meknès
Département de Mathématiques

Cours d'Analyse Numérique II

Filière SMA-S5

Pr. Samir KHALLOUQ

Année universitaire 2021-2022

Table des matières

Analyse Numérique C'est Quoi ?

Pour aborder le calcul numérique (à l'aide d'un outil informatique) des solutions d'un problème "réel", on passe par les étapes suivantes :

1. **Description qualitative des phénomènes physiques :** Cette étape est effectuée par des spécialistes des phénomènes que l'on veut quantifier (ingénieurs, chimistes, biologistes etc.....)

2. **Modélisation :** Il s'agit, à partir de la description qualitative précédente, d'écrire un modèle mathématique. Dans la plupart des cas, on ne saura pas calculer une solution analytique, explicite, du modèle ; on devra faire appel à des techniques de résolution approchée.

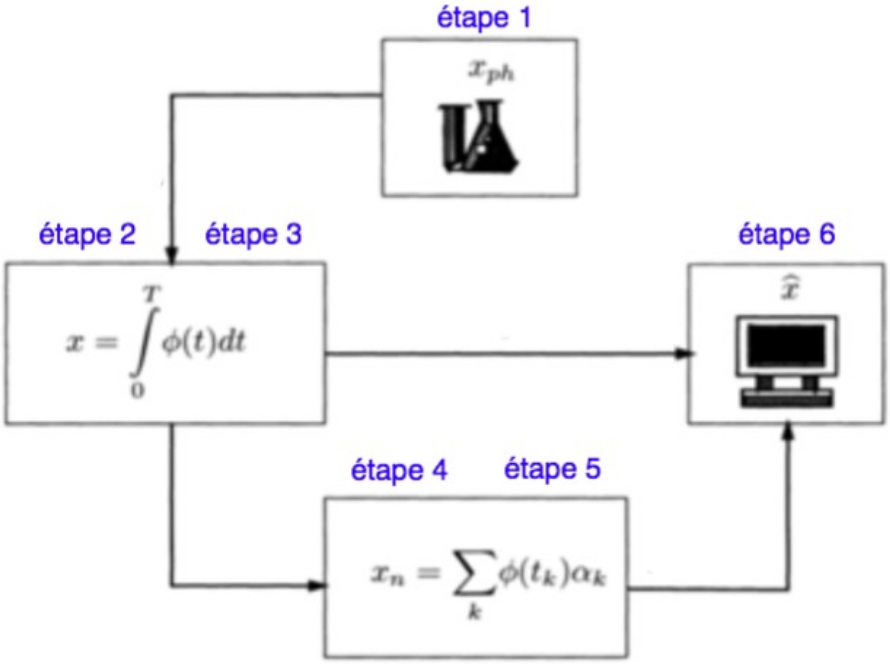
3. **Analyse mathématique :** Même si l'on ne sait pas trouver une solution explicite du modèle, il est important d'en étudier les propriétés mathématiques, dans la mesure du possible. Il est bon de se poser les questions suivantes :
   - Le problème est-il bien posé ? c'est-à-dire y-a-t'il existence et unicité de la solution ?
   - Les propriétés physiques auxquelles on s'attend sont elles satisfaites par les solutions du modèle mathématique ?

4. **Discrétisation et résolution numérique :** Un problème posé sur un domaine continu (espace - temps) n'est pas résoluble tel quel par un ordinateur, qui ne peut traiter qu'un nombre fini d'inconnues. Pour se ramener à un problème en dimension finie, on discrétise l'espace et/ou le temps.

5. **Analyse numérique :** Une fois le problème discret obtenu, il est raisonnable de se demander si la solution de ce problème est proche, et en quel sens, du problème continu.

6. **Mise en oeuvre, programmation et analyse des résultats**

La figure suivante donne les différentes étapes pour trouver une solution d'un problème réel :



Chapitre 1

Résolution numérique des équations différentielles ordinaires

1.1 Introduction

Nous abordons dans ce chapitre la résolution numérique du problème de Cauchy pour les équations différentielles ordinaires (E.D.O.). Après un bref rappel des notions de base sur les E.D.O., nous introduisons les techniques les plus couramment utilisées pour l'approximation numérique des équations scalaires. Nous présentons ensuite les concepts de consistance, convergence, zérostabilité et stabilité absolue.

Définition 1.1 *(E.D.O) Une équation différentielle ordinaire (E.D.O) d'ordre n est une relation entre la variable t, une fonction $y(t)$ et ses dérivées, $y', y'', \cdots, y^{(n)}$ au point t, définie par :*

$$f\left(t, y, y', y'', \cdots, y^{(n)}\right) = 0, \text{ pour } t \in I \subset \mathbb{R}$$

Définition 1.2 *(E.D.A) Une équation différentielle autonome (ou homogène) d'ordre n toute équation de la forme :*

$$y^{(n)} = f\left(y, y', y'', \cdots, y^{(n-1)}\right), \text{ pour } t \in I \subset \mathbb{R}$$

Autrement dit, f ne dépend pas explicitement de t.

Exemple 1.1 (Modèle Malthusien) *Soit $N(t)$ l'effectif d'une population à un instant t, l'évolution de la population $N(t)$ est décrite par une (E.D.O) de type : $\frac{d}{dt}N(t) = naissances - décés + immigration$. S'il n'y a pas de migration et que les naissances et les décés sont proportionnelle à $N(t)$, alors on a :*

$$\frac{d}{dt}N(t) = \alpha N(t) - \beta N(t) = rN(t), t \in \mathbb{R}^+$$

avec α le taux naissances, β le taux de mortalité et r le taux de croissance.

1.2 Problème de Cauchy

Le problème de Cauchy (ou problème aux valeurs initiales) consiste à trouver la solution d'une E.D.O, satisfaisant des conditions initiales. Dans le cas scalaire, le problème de Cauchy associé à une EDO du premier ordre s'écrit : Trouver une fonction $y \in C^1(I)$ telle que

$$\begin{cases} y'(t) = f(t, y(t)), t \in I, \\ y(t_0) = y_0, \end{cases} \tag{1.1}$$

où I est un intervalle de $\mathbb{R}$ contenant t_0 et $f(t, y)$ est une fonction à valeur réelle définie sur $I \times \mathbb{R}$ continue par rapport aux deux variables.
En intégrant (1.1) entre t_0 et t, on obtient :

$$y(t) = y_0 + \int_{t0}^{t} f(\tau, y(\tau))d\tau. \tag{1.2}$$

Si f est continue, on vérifie facilement que le problème de Cauchy (1.1) est équivalent à l'équation intégrale (1.2).

1.3 Existence, unicité et stabilité

Définition 1.3 *Une fonction f à deux variables t et y définie dans $I \times \mathbb{R}$ est dite localement lipschitzienne par rapport à sa seconde variable y s'il existe une boule ouverte $J \subseteq I$ centrée en t_0 de rayon r_J, une boule ouverte Σ centrée en y_0 de rayon r_Σ et une constante $L > 0$ telles que :*

$$|f(t, y_1) - f(t, y_2)| \leq L |y_1 - y_2|, \forall t \in J, \forall y_1, y_2 \in \Sigma. \tag{1.3}$$

Théorème 1.1 *(**Existence locale**) Si f est localement lipschitzienne par rapport à y, alors le problème de Cauchy* (1.1) *admet une unique solution dans une boule ouverte de centre t_0 et de rayon r_0 $(0 < r_0 < \min(r_J, r_\Sigma/M, 1/L))$, où $M = \max_{(t,y)\in J\times\Sigma} |f(t, y)|$. Cette solution est appelée solution locale.*

Remarque 1.1 *La condition* (1.3) *est automatiquement vérifiée si la dérivée de f par rapport à y est continue : en effet, dans ce cas, il suffit de prendre pour L le maximum de $|\partial f(t, y)/\partial y|$ sur $\overline{J \times \Sigma}$.*

Théorème 1.2 *(**Existence globale**) Si f est uniformément lipschitzienne par rapport à y, c-à-d il existe une constante $L > 0$ telle que pour tous $y_1, y_2 \in \Sigma = \mathbb{R}$ et $t \in J = I$ on a :*

$$|f(t, y_1) - f(t, y_2)| \leq L |y_1 - y_2|,$$

alors le problème de Cauchy (1.1) *admet une solution globale unique.*

Pour l'analyse de la stabilité du problème de Cauchy, on considère le problème perturbé suivant :

$$\begin{cases} z'(t) = f(t, z(t)) + \delta(t), t \in I, \\ z(t_0) = y_0 + \delta_0, \end{cases} \tag{1.4}$$

où $\delta_0 \in \mathbb{R}$ et δ est une fonction continue sur I. Caractérisons à présent la sensibilité de la solution z par rapport à ces perturbations.

Définition 1.4 *Soit I un intervalle borné, le problème de Cauchy* (1.1) *est stable au sens de Liapunov (ou simplement stable) sur I si, pour toute perturtation $(\delta_0, \delta(t))$ satisfaisant*

$$|\delta_0| < \varepsilon, \qquad |\delta(t)| < \varepsilon, \forall t \in I,$$

avec $\varepsilon > 0$ assez petit pour garantir l'existence de la solution du problème perturbé (1.4)*, alors il existe une constante $C > 0$ indépendante de ε telle que*

$$|y(t) - z(t)| < C\varepsilon, \forall t \in I, \tag{1.5}$$

avec y est une solution du problème de Cauchy.
Lorsque I n'est pas borné supérieurement, on dit que (1.1) *est asymptotiquement stable si, en plus de* (1.5) *on a*

$$\lim_{t \to +\infty} |y(t) - z(t)| = 0, \ \textit{si} \lim_{t \to +\infty} |\delta(t)| \to 0.$$

Proposition 1.1 *Si f est uniformément lipschitzienne par rapport à y, alors le problème de Cauchy* (1.1) *est stable.*

Preuve : Soit $y(t)$ solution du problème :

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

et soit $z(t)$ solution du problème perturbé :

$$\begin{cases} z'(t) = f(t, z(t)) + \delta(t), t \in I, \\ z(t_0) = y_0 + \delta_0. \end{cases}$$

Posons $w(t) = z(t) - y(t)$. On a

$$\begin{aligned} w'(t) &= z'(t) - y'(t), \\ &= f(t, z(t)) - f(t, y(t)) + \delta(t). \end{aligned}$$

On en déduit que

$$w(t) = \delta_0 + \int_{t0}^{t} (f(s, z(s)) - f(s, y(s)))ds + \int_{t0}^{t} \delta(s)ds.$$

Donc si f est uniformément lipschitzienne par rapport à y alors :

$$|w(t)| \leq (1 + |t - t_0|)\,\varepsilon + L \int_{t0}^{t} |w(s)|ds.$$

Pour conclure, on a besoin du lemme suivant :

Lemme 1.1 *(**Lemme de Gronwall**) Soit f une fonction réelle positive intégrable sur l'intervalle $]t_0, t_0+T[$ et soient g et φ deux fonctions réelles continues sur $[t_0, t_0+T]$, avec g croissante. Si φ satisfait linégalité :*

$$\varphi(t) \leq g(t) + \int_{t0}^{t} f(s)\varphi(s)ds, \qquad \forall t \in [t_0, t_0+T],$$

alors

$$\varphi(t) \leq g(t) \exp\left(\int_{t0}^{t} f(s)ds\right), \qquad \forall t \in [t_0, t_0+T].$$

En appliquant ce lemme, on obtient

$$|w(t)| \leq (1+|t-t_0|)\,\varepsilon \exp\left(L\,|t-t_0|\right) \leq C\varepsilon, \qquad \forall t \in I,$$

avec $C = (1+K_I)\exp(LK_I)$ où $K_I = \max_{t\in I}|t-t_0|$. D'où la stabilité.

On ne sait intégrer qu'un très petit nombre d'EDO non linéaires. De plus, même quand c'est possible, il n'est pas toujours facile d'exprimer explicitement la solution.

Exemple 1.2 *L'équation (très simple) $y' = (y-t)/(y+t)$, admet une solution qu'on ne peut définir que de manière implicite par la relation $(1/2)\log(t^2+y^2) + \tan^{-1}(y/t) = C$, où C est une constante dépendant de la condition initiale.*

Pour cette raison, nous sommes conduits à considérer des méthodes numériques. Celles-ci peuvent en effet être appliquées à n'importe quelle EDO, sous la seule condition qu'elle admette une unique solution.

1.4 Méthodes numériques à un pas

Abordons à présent l'approximation numérique du problème de Cauchy (1.1). On fixe $0 < T < +\infty$ et on note $I =]t_0, t_0+T[$ l'intervalle d'intégration. Pour $h_n > 0$, soit t_n, avec $n = 0, 1, 2, \ldots, N$, une suite de noeuds de I induisant une discrétisation de I en sous-intervalles $I_n = [t_n, t_{n+1}]$. La longueur h_n de ces sous-intervalles est appelée pas de discrétisation. Le nombre N est le plus grand entier tel que $t_N \leq t_0 + T$. Soit y_j l'approximation au noeud t_j de la solution exacte $y(t_j)$. De même, f_j désigne la valeur $f(t_j, y_j)$. On pose naturellement $y_0 = y(0)$.
Dans la plupart des méthodes, on prend h_n constante et on la note h, dans ce cas $h = \frac{T}{N}$ et $t_n = t_0 + nh \;\; (n = 0, \ldots, N)$.

Définition 1.5 *Une méthode de résolution d'équation différentielles est dite à un pas si elle est de la forme :*

$$y_{n+1} = y_n + h\Phi\left(t_n, y_n, f\left(t_n, y_n\right); h\right) \;\; \textit{pour } n = 0, \cdots, N-1 \tag{1.6}$$

*où Φ est une fonction quelconque appelée **fonction d'incrément**. La méthode est à un pas si, pour obtenir une approximation de la solution en t_{n+1}, on doit utiliser seulement la solution au temps t_n. Autrement, on dit que la méthode est multi-pas (ou à pas multiples).*

Définition 1.6 (Méthodes explicites et méthodes implicites) *Une méthode est dite explicite si la valeur y_{n+1} peut être calculée directement à l'aide des valeurs précédentes $y_k, k \leq n$ (ou d'une partie d'entre elles). Une méthode est dite implicite si y_{n+1} n'est définie que par une relation implicite. faisant intervenir la fonction f.*

1.4.1 Méthodes d'Euler

La méthode d'Euler est la méthode la plus simple de résolution numérique d'équations différentielles ordinaires et son emploi est facile. Toutefois, elle est relativement peu utilisée en raison de sa faible précision.
Le but est d'obtenir une approximation de la solution en $t = t_1 = t_0 + h$. Nous n'avons pas l'équation de la courbe $y(t)$, mais nous en connaissons la pente $y'(t)$ en $t = t_0$. En effet $y'(t_0) = f(t_0, y(t_0)) = f(t_0, y_0)$. On peut donc suivre la droite passant par (t_0, y_0) et de pente $f(t_0, y_0)$, l'équation de cette droite est

$$d_0(t) = y_0 + f(t_0, y_0)(t - t_0).$$

En $t = t_1$, on a :

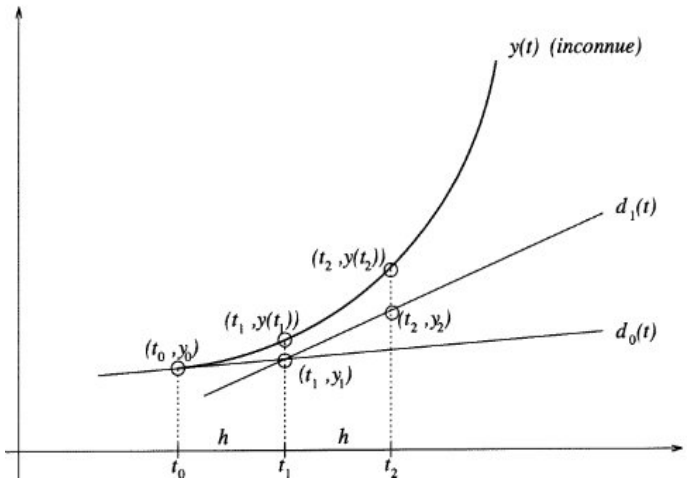


FIGURE 1.1 – Méthode d'Euler : illustration géométrique

$$d_0(t_1) = y_0 + f(t_0, y_0)(t_1 - t_0) = y_0 + hf(t_0, y_0) = y_1.$$

En d'autre termes, $d_0(t_1)$ est proche de la solution analytique $y(t_1)$, c'est-à-dire :

$$y(t_1) \simeq y_1 = d_0(t_1) = y_0 + hf(t_0, y_0)$$

Il est important de noter que, le plus souvent, $y_1 \neq y(t_1)$, donc lorsque on souhaite faire une deuxième itération et obtenir une approximation de $y(t_2)$, on dispose pas de la valeur

exacte de $y(t_1)$ mais seulement y_1. On doit alors utiliser l'expression :

$$y'(t_1) = f(t_1, y(t_1)) \simeq f(t_1, y_1)$$

et construire la droite :

$$d_1(t) = y_1 + f(t_1, y_1)(t - t_1)$$

On constate que les erreurs se propagent d'une itération à l'autre. De façon générale l'erreur $|y(t_n) - y_n|$ augmente légèrement avec n. On arrive à l'algorithme suivant :
Algorithm 1 (**Euler**)
Données : $h, (t_0, y_0)$, N et f ;
1 : $y \leftarrow 0, y(0) \leftarrow y_0$;
2 : **Pour** $n = 0$ à $N - 1$ **faire**
3 : $y(n+1) \leftarrow y(n) + hf(t_0, y(n))$
4 : $t_0 \leftarrow t_0 + h$;
5 : **Fin Pour**
6 : **Résultat** y ;

Exemples des méthodes d'Euler

- Méthode d'Euler progressive

$$\begin{cases} y_{n+1} = y_n + hf(t_n, y_n) \\ y_0 \text{ donné,} \end{cases}$$

pour $n = 0, 1, \ldots, N - 1$.
- Méthode d'Euler rétrograde

$$\begin{cases} y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}), \\ y_0 \text{ donné,} \end{cases}$$

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

Remarque 1.2

- *La méthode d'Euler progressive est explicite, tandis que celle d'Euler rétrograde est implicite.*
- *Le calcul de y_{n+1}, dans la méthode d'Euler rétrograde, nécessite la résolution d'une équation non linéaire (si f dépend non linéairement de la seconde variable).*

Exemple 1.3 *On considère l'équation différentielle suivante*

$$y'(t) = y(t) - t + 1,$$

avec la condition initiale $y(0) = 1$. La solution analytique est $y(t) = e^t + t$.

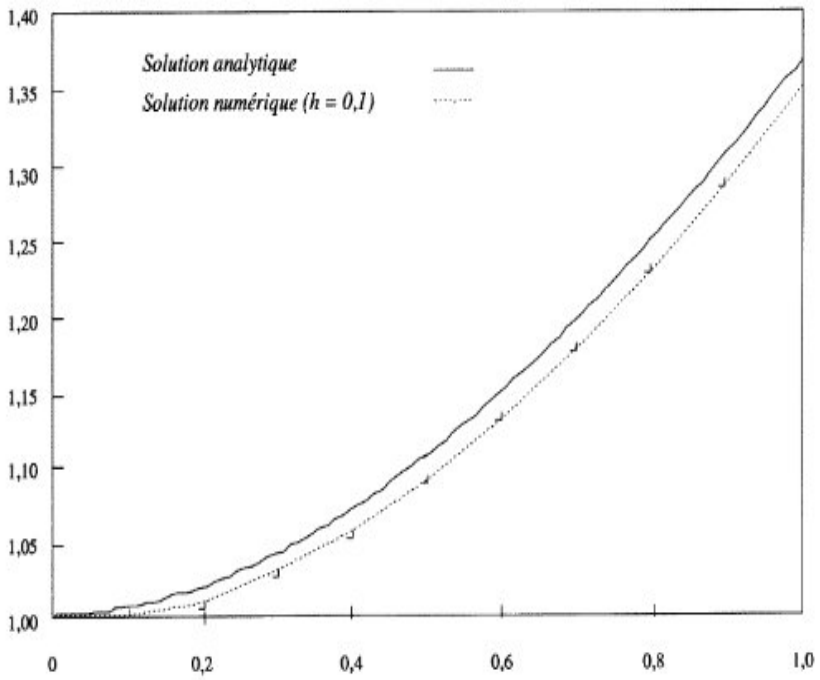


FIGURE 1.2 – Méthode d'Euler : $f(t, y) = y - t + 1$ et $y_0 = 1$

La figure représente une comparaison entre la solution exacte et celle obtenue par la méthode d'Euler explicite pour un pas $h = 0.1$.

1.4.2 Méthodes de Taylor

Le développement de Tayolor autorise une généralisation de la méthode d'Euler permettant d'obtenir une erreur de troncature locale d'ordre plus élvé. On cherche, au temps $t = t_n$ une approximation de la solution en $t = t_{n+1}$. On suppose que la solution $y(t)$ du problème de Cauchy (1.1) est suffisament régulière. En effectuant un développement de Taylor au voisinage de t_n, on obtient :

$$y\left(t_{n+1}\right) = y\left(t_n\right) + hy'\left(t_n\right) + \frac{h^2}{2}y''\left(t_n\right) + O\left(h^3\right),$$

En se servant de l'équation (1.1) on trouve :

$$y\left(t_{n+1}\right) = y\left(t_n\right) + hf\left(t_n, y\left(t_n\right)\right) + \frac{h^2}{2}f'\left(t_n, y\left(t_n\right)\right) + O\left(h^3\right).$$

La règle de dérivation assure que

$$f'\left(t_n, y\left(t_n\right)\right) = \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial t} + \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial y}y'\left(t_n\right),$$

c'est-à-dire

$$f'\left(t_n, y\left(t_n\right)\right) = \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial t} + \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial y} f\left(t_n, y\left(t_n\right)\right).$$

On obtient donc

$$y\left(t_{n+1}\right) = y\left(t_n\right) + hf\left(t_n, y\left(t_n\right)\right) + \frac{h^2}{2}\left(\frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial t} + \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial y} f\left(t_n, y\left(t_n\right)\right)\right) + O\left(h^3\right),$$

En négligeant les termes d'ordre supérieur ou égale à 3 , on en arrive à poser :

$$y\left(t_{n+1}\right) \simeq y\left(t_n\right) + hf\left(t_n, y\left(t_n\right)\right) + \frac{h^2}{2}\left(\frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial t} + \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial y} f\left(t_n, y\left(t_n\right)\right)\right), \tag{1.7}$$

qui sera la base des méthodes de Taylor. L'équation (1.7) peut s'écrire sous la forme (1.6). Dans ce cas, on a :

$$\Phi\left(t_n, y_n, f\left(t_n, y_n\right); h\right) = f\left(t_n, y_n\right) + \frac{h}{2}\left(\frac{\partial f\left(t_n, y_n\right)}{\partial t} + \frac{\partial f\left(t_n, y_n\right)}{\partial y} f\left(t_n, y_n\right)\right).$$

Algorithm 2 (**Taylor d'ordre 2**)

Données : $h, (t_0, y_0)$, N et f ;
1 : $y \leftarrow 0, y(0) \leftarrow y_0$;
2 :**Pour** $n = 0$ à $N - 1$ **faire**
3 :

$$y(n+1) \leftarrow y(n) + hf\left(t_0, y(n)\right) + \frac{h^2}{2}\left(\frac{\partial f\left(t_0, y(n)\right)}{\partial t} + \frac{\partial f\left(t_0, y(n)\right)}{\partial y} f\left(t_0, y(n)\right)\right)$$

4 : $t_0 \leftarrow t_0 + h$;
5 : **Fin Pour**
6 :**Résultat** y ;

Remarque 1.3 *Dans cette algorithme, on a remplacé la solution analytique $y\left(t_n\right)$ par son approximation y_n dans la relation* (1.7)*. On conclut que les erreurs se propagent d'une itération à une autre.*

Il est possible d'obtenir des méthodes de Taylor encore plus précises en poursuivant le développement de Taylor jusqu'à des termes plus élevés. On doit donc évaluer des dérivées de la fonction $f(t, y(t))$ d'ordre de plus en plus élevé, ce qui nécessite le calcul de :

$$\frac{\partial^2 f}{\partial t^2}, \frac{\partial^2 f}{\partial y^2}, \frac{\partial^2 f}{\partial t \partial y}.$$

Pour cette raison, les méthodes obtenues sont difficile à utiliser. Il existe cependant un moyen de contourner cette difficulté en développant les méthodes de Runge-Kutta.

1.4.3 Méthodes de Runge-Kutta

Il serait avantageux de disposer de méthodes d'ordre du plus en plus élevé tout en évitant les inconvénients des méthodes de Taylor qui nécessitent l'évaluation des dérivés partielles de la fonction f. Une voie alternative est les méthodes de Runge-Kutta.

Méthodes de Runge-Kutta d'ordre 2

Le développement de la méthode de Taylor passe par la relation :

$$\begin{aligned} y\left(t_{n+1}\right) = y\left(t_n\right) &+ hf\left(t_n, y\left(t_n\right)\right) \\ &+ \frac{h^2}{2}\left(\frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial t} + \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial y} f\left(t_n, y\left(t_n\right)\right)\right) + O\left(h^3\right) \end{aligned} \tag{1.8}$$

Le but est de remplacer cette relation par une expression équivalente possédant le même ordre de précision $O\left(h^3\right)$. On propose la forme :

$$y\left(t_{n+1}\right) = y\left(t_n\right) + a_1 hf\left(t_n, y\left(t_n\right)\right) + a_2 hf\left(t_n + a_3 h, y\left(t_n\right) + a_4 h\right), \tag{1.9}$$

où on doit déterminer les paramètres a_1, a_2, a_3 et a_4 de telle sorte que les deux expressions (1.8) et (1.9) aient une erreur en $O\left(h^3\right)$. Cette dernière expression ne contient aucune dérivée partielle. Pour y arriver, on utilise le dêveloppement de Taylor en deux variables autour du point $\left(t_n, y\left(t_n\right)\right)$.

On a ainsi :

$$\begin{aligned} f\left(t_n + a_3 h, y\left(t_n\right) + a_4 h\right) = f\left(t_n, y\left(t_n\right)\right) &+ a_3 h \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial t} \\ &+ a_4 h \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial y} + O\left(h^2\right), \end{aligned} \tag{1.10}$$

la relation (1.9) devient alors :

$$\begin{aligned} y\left(t_{n+1}\right) = y\left(t_n\right) &+ \left(a_1 + a_2\right) hf\left(t_n, y\left(t_n\right)\right) \\ &+ a_2 a_3 h^2 \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial t} + a_2 a_4 h^2 \frac{\partial f\left(t_n, y\left(t_n\right)\right)}{\partial y} + O\left(h^3\right). \end{aligned} \tag{1.11}$$

On voit que les expressions (1.8) et (1.11) sont du même ordre. Pour déterminer les coefficients a_i, il suffit de comparer ces deux expressions terme à terme, on en arrive au système suivant :

$$\begin{cases} 1 = a_1 + a_2, \\ \frac{1}{2} = a_2 a_3, \\ \dfrac{f\left(t_n, y\left(t_n\right)\right)}{2} = a_2 a_4. \end{cases} \tag{1.12}$$

Le système (1.12) est sous-déterminé, il y a moins d'équation que d'inconnues et qu'il n'a donc pas de solution unique. Ce qui favorise la mise au point de plusieurs variantes de la

méthode de Runge-Kutta. Une solution du système (1.12) est donnée par $a_1 = a_2 = \frac{1}{2}$, $a_3 = 1$ et $a_4 = f(t_n, y(t_n))$. Il suffit ensuite de remplacer ces valeurs dans l'équations (1.9). Pour ce faire, on doit négliger le terme en $O(h^3)$ et remplacer la valeur exacte $y(t_n)$ par approximation y_n. On obtient l'algorithme d'Euler modifié suivant :

Algorithm 3 (Algorithme d'Euler modifié)
Données : $h, (t_0, y_0)$, N et f ;
$1 : y \leftarrow 0, y(0) \leftarrow y_0$
$2 :$ **Pour** $n = 0$ à $N - 1$ **Faire**
$3 : \quad \bar{y} \leftarrow y(n) + hf(t_0, y(n))$
$4 : y(n+1) \leftarrow y(n) + \frac{h}{2}(f(t_0, y(n)) + f(t_0 + h, \bar{y}))$
$5 : t_0 \leftarrow t_0 + h;$
$6 :$ **Fin Pour**
$7 :$ **Résultat** y ;

Remarque 1.4 *La variable temporaire $\bar{y}$ correspond tout simplement à une itération de la méthode d'Euler et qui est une prédiction de la solution en t_{n+1} corrigée à la deuxième étape de l'algorithme. On parle d'une methode de prédiction-correction.*

Un deuxième choix possible des a_i est $a_1 = 0, a_2 = 1, a_3 = \frac{1}{2}$ et $a_4 = \frac{f(t_{n.y}(t_n))}{2}$. En remplaçant ces valeurs des coefficients a_i dans l'équation (1.9) on obtient l'algorithme du point milieu suivant :

Algorithm 4 (Algorithme du point milieu)
Données : $h, (t_0, y_0)$, N et f ;
$1 : y \leftarrow 0, y(0) \leftarrow y_0;$
$2 :$ **Pour** $n = 0$ à $N - 1$ **Faire**
$3 : \quad \bar{y} \leftarrow \frac{h}{2} f(t_0, y(n))$
$4 : \quad y(n+1) \leftarrow y(n) + hf\left(t_0 + \frac{h}{2}, y(n) + \tilde{y}\right)$
$5 : t_0 \leftarrow t_0 + h$;
$6 :$ **Fin Pour**
7. **Résultat** y ;

Remarque 1.5 *Les méthodes d'Euler modifiée et du point milieu sont d'ordre 2. D'autres choix sont possibles pour les coefficients a_i.*

Méthode de Runge-Kutta d'ordre 4
En reprenant le d'eveloppement de Taylor, mais cette fois jusqu'à l'ordre 5, un raisonnement similaire à celui qui a mené aux méthodes de Runge-Kutta d'ordre 2 aboutit à un système de 8 équations non linéaire de 10 inconnues. Le résultat est la méthode de Runge-Kutta d'ordre 4 qui est très fréquemment utilisée en raison de sa grande précision.

Algorithm 5 (Algorithme de Runge-Kutta d'ordre 4)
Require : $h, (t_0, y_0)$, N et f ;
1 : $y \leftarrow 0, y(0) \leftarrow y_0$;
2. **Pour** $n = 0$ **à** $N - 1$ **Faire**
3 : $k_1 \leftarrow hf\left(t_0, y(n)\right)$
4 : $k_2 \leftarrow hf\left(t_0 + \frac{h}{2}, y(n) + \frac{k_1}{2}\right)$
5 : $k_3 \leftarrow hf\left(t_0 + \frac{h}{2}, y(n) + \frac{k_2}{2}\right)$
6 : $k_4 \leftarrow hf\left(t_0 + h, y(n) + k_3\right)$
7 : $y(n+1) \leftarrow y(n) + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right)$
8 : $t_0 \leftarrow t_0 + h$;
9. **Fin Pour**
10 : **Résultat** y ;

Une autre approche pour décrire les méthodes de Runge-Kutta consiste à utiliser des formules d'intégration numérique pour approcher l'intégrale de l'équation (1.2). Pour obtenir une meilleure approximation de la solution du problème de Cauchy (1.1), il faut se demander comment approcher l'intégrale de (1.2) de manière précise. L'idée est donc de calculer par récurrence les points (t_n, y_n) en utilisant des points intermédiaires (t_{ni}, y_{ni}) avec $t_{ni} = t_n + c_i h \, (c_i \in [0,1], i = 1, 2, \ldots, q)$. À chacun de ces points intermédiaires, on associé les poids corresponds

$$w_{ni} = f\left(t_{ni}, y_{ni}\right)$$

Si y est solution de l'équation $y'(t) = f(t, y(t))$ pour $t \in I$, alors en intégrant cette équation entre t_n et t_{ni} on aboutit à

$$y\left(t_{ni}\right) = y\left(t_n\right) + \int_{t_n}^{t_{ni}} f(t, y(t))dt.$$

Posons $t = t_n + sh$, alors

$$y\left(t_{ni}\right) = y\left(t_n\right) + h\int_0^{c_i} f\left(t_n + sh, y\left(t_n + sh\right)\right) ds. \tag{1.13}$$

De même, on a

$$y\left(t_{n+1}\right) = y\left(t_n\right) + h\int_0^{1} f\left(t_n + sh, y\left(t_n + sh\right)\right) ds \tag{1.14}$$

Posons $\varphi(s) = f\left(t_n + sh, y\left(t_n + sh\right)\right)$ et considérons les formules d'intégration suivantes :

$$\int_0^{c_i} \varphi(s)ds \simeq \sum_{1 \le j < i} a_{ij}\varphi\left(c_j\right) \text{ et } \int_0^{1} \varphi(s)ds \simeq \sum_{1 \le j \le q} b_j\varphi\left(c_j\right) \tag{1.15}$$

On supposera que la condition suivante est vérifie :

$$c_i = \sum_{1\leq j<i} a_{ij} \text{ et } \sum_{1\leq j\leq q} b_j = 1.$$

Les formules (1.15) sont exactes pour $\varphi \equiv 1$. En appliquant ces formules d'intégration à (1.13) et (1.14), on obtient

$$y(t_{ni}) \simeq y(t_n) + h \sum_{1\leq j<i} a_{ij} f(t_{nj}, y(t_{nj})),$$

puis

$$y(t_{n+1}) \simeq y(t_n) + h \sum_{1\leq j\leq q} b_j f(t_{nj}, y(t_{nj})),$$

La méthode de Runge-Kutta correspondante est donc

$$\left.\begin{array}{l} t_{ni} = t_n + c_i h, \\ y_{ni} = y_n + h\sum_{1\leq j<i} a_{ij} w_{nj} \\ w_{ni} = f(t_{ni}, y_{ni}) \end{array}\right\} \text{ pour } i = 1, \cdots, q,$$

et

$$y_{n+1} = y_n + h \sum_{j=1}^{q} b_j w_{nj}.$$

On présente les paramétres de la méthode de Runge-Kutta dans un tableau

$$\begin{array}{c|ccccc} c_1 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \vdots & \vdots & \vdots & & & \\ c_q & a_{q1} & a_{q2} & \cdots & a_{qq-1} & \\ \hline & b_1 & b_2 & \cdots & b_{q-1} & b_q \end{array}$$

Méthode d'Euler modifiée : c'est une méthode de Runge-Kutta d'ordre 2, elle correspond aux paramétres suivants : $q = 2$

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & 1/2 & 1/2 \end{array}$$

L'algorithme d'Euler modifié pour le problème de Cauchy (1.1) s'écrit : pour $n = 0, 1, \cdots, N-1$

$$\begin{aligned} w_{n1} &= f(t_n, y_n), \\ y_{n2} &= y_n + h w_{n1}, \\ w_{n2} &= f(t_{n+1}, y_{n2}), \\ y_{n+1} &= y_n + \frac{h}{2}(w_{n1} + w_{n2}). \end{aligned}$$

Méthode du point milieu : c'est une méthode de Runge-Kutta d'ordre 2, elle correspond aux paramétres suivants : $q = 2$

$$\begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array}$$

L'algorithme du point milieu pour le problème de Cauchy (1.1) s'écrit pour $n = 0, \cdots, N-1$

$$\begin{aligned} w_{n1} &= f(t_n, y_n) \\ y_{n2} &= y_n + \frac{h}{2} w_{n1} \\ w_{n2} &= f\left(t_n + \frac{h}{2}, y_{n2}\right), \\ y_{n+1} &= y_n + h w_{n2}. \end{aligned}$$

Méthode de Heun d'odre 3 **:** c'est une méthode de Runge-Kutta d'ordre 3, elle correspond aux paramétres suivants : $q = 3$

$$\begin{array}{c|ccc} 0 & & & \\ 1/3 & 1/3 & & \\ 2/3 & 0 & 2/3 & \\ \hline & 1/4 & 0 & 3/4 \end{array}$$

L'algorithme de Heun d'odre 3 pour le problème de Cauchy (1.1) s'écrit : pour $n = 0, 1, \cdots, N-1$

$$\begin{aligned} w_{n1} &= f(t_n, y_n), \\ w_{n2} &= f\left(t_n + \frac{h}{3}, y_n + \frac{h}{3} w_{n1}\right), \\ w_{n3} &= f\left(t_n + \frac{2h}{3}, y_n + \frac{2h}{3} w_{n2}\right), \\ y_{n+1} &= y_n + \frac{h}{4}(w_{n1} + 3 w_{n3}). \end{aligned}$$

Méthodes de Runge-Kutta d'ordre 4 : Elle correspond aux paramétres suivants :

$q = 4$

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

L'algorithme de Runge-Kutta d'ordre 4 pour le problème de Cauchy (1.1) s'écrit : pour $n = 0, 1, \cdots, N-1$

$$
\begin{aligned}
w_{n1} &= f(t_n, y_n) \\
w_{n2} &= f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} w_{n1}\right), \\
w_{n3} &= f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} w_{n2}\right), \\
w_{n4} &= f(t_{n+1}, y_n + h w_{n3}) \\
y_{n+1} &= y_n + \frac{h}{6}\left(w_{n1} + 2w_{n2} + 2w_{n3} + w_{n4}\right).
\end{aligned}
$$

Généralisation de la méthode de Runge-Kutta "RK"

Les méthodes de Runge-Kutta sont des cas particuliers des méthodes à un pas. Elles s'écrivent sous la forme

$$y_{n+1} = y_n + h\Phi\left(t_n, y_n, f(t_n, y_n); h\right), \quad n \geq 0$$

où Φ est la fonction d'incrément définie par

$$
\begin{aligned}
\Phi\left(t_n, y_n, f(t_n, y_n); h\right) &= \sum_{i=1}^{d} b_i K_i, \\
K_i &= f\left(t_n + c_i h, y_n + h \sum_{j=1}^{d} a_{ij} K_j\right), \quad i = 1, 2, \ldots, d
\end{aligned}
$$

où d désigne le nombre d'étapes de la méthode. Les coefficients $\{a_{ij}\}, \{c_i\}$ et $\{b_i\}$ caractérisent complètement une méthode RK. On peut les présenter dans un tableau de **Butcher** :

c_1	a_{11}	a_{12}	$\ldots$	a_{1d}
c_2	a_{21}	a_{22}		a_{2d}
$\vdots$	$\vdots$		$\ddots$	$\vdots$
c_d	a_{d1}	a_{d2}	$\ldots$	a_{dd}
	b_1	b_2	$\ldots$	b_d

où

$\mathbf{C}$	$\mathbf{A}$
	$\mathbf{b}^T$

avec $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{d\times d}, \mathbf{b} = [b_1, \ldots, b_d]^T \in \mathbb{R}^d$ et $\mathbf{C} = [c_1, \ldots, c_d]^T \in \mathbb{R}^d$. Nous supposerons dans la suite que la condition suivante est vérifiée

$$c_i = \sum_{j=1}^{d} a_{ij}, \quad i = 1, \ldots, d.$$

Si les coefficients a_{ij} de $\mathbf{A}$ sont nuls pour $j \geq i, i = 1, 2, \ldots, d$, alors chaque K_i peut être explicitement calculé en fonction des $i-1$ coefficients $K_1, \ldots, K_{i-1}$ déjà connus. Dans ce cas la méthode RK est explicite. Autrement, elle est implicite et il faut résoudre un système non linéaire de dimension d pour calculer les K_i.
L'augmentation des calculs pour les schémas implicites rend leur utilisation coûteuse ; un compromis acceptable est obtenu avec les méthodes RK semiimplicites. Dans ce cas $a_{ij} = 0$ pour $j > i$, de sorte que chaque K_i est solution de l'équation non linéaire

$$K_i = f\left(t_n + c_i h, y_n + h a_{ii} K_i + h \sum_{j=1}^{i-1} a_{ij} K_j\right).$$

Un schéma semi-implicite implique donc la résolution de d équations non linéaires indépendantes.

1.4.4 Méthode de Crank-Nicolson

La méthode de Crank-Nicolson appelé aussi méthode du trapèze est de la forme

$$y_{n+1} = y_n + \frac{h}{2}\left(f\left(t_n, y_n\right) + f\left(t_{n+1}, y_{n+1}\right)\right). \tag{1.16}$$

Cette méthode provient de l'approximation de l'intégrale de (1.2) par la formule de quadrature du trapèze.

1.5 Analyse des méthodes à un pas

1.5.1 Consistance

Définition 1.7 *(**Erreur de consistance**) On appelle erreur de consistance la quantité obtenue en remplaçant l'inconnue par la solution exacte dans le schéma numérique. Dans le cas du schéma* (1.6) *(schéma explicite), l'erreur de consistance (ou erreur de troncature locale (ETL)) au point t_{n+1} est donc définie par :*

$$R_{n+1}(h) = \frac{y\left(t_{n+1}\right) - y\left(t_n\right)}{h} - \Phi\left(t_n, y\left(t_n\right), f\left(t_n, y\left(t_n\right)\right); h\right), n = 0, \cdots N-1. \tag{1.17}$$

L'erreur de troncature globale est alors définie par

$$R(h) = \max_{0\leq n\leq N-1} \left|R_{n+1}(h)\right|.$$

On dit qu'un schéma de discrétisation est consistant si

$$\lim_{h\to 0} R = \lim_{h\to 0} \max_{1\leq n\leq N} |R_n| = 0$$

ou N est le nombre de points de diserétisation.

L'erreur de consistance R_{n+1} est donc l'erreur qu'on commet en remplaçant l'opérateur y' par le quotient différentiel :

$$\frac{y\left(t_{n+1}\right) - y\left(t_n\right)}{h}.$$

Cette erreur peut être évaluée si y est suffisamment régulière, en effectuant des développements de Taylor.
Remarquer que $R(h)$ dépend de la solution y du problème de Cauchy (1.1).
La méthode d'Euler progressive est un cas particulier de (1.6), pour lequel

$$\Phi\left(t_n, y_n, f_n; h\right) = f_n$$

et la méthode de Heun d'odre 2 correspond à

$$\Phi\left(t_n, y_n, f_n; h\right) = \frac{1}{2}\left[f_n + f\left(t_n + h, y_n + hf_n\right)\right]$$

Un schéma explicite à un pas est entièrement caractérisé par sa fonction d'incrément Φ. Cette fonction, dans tous les cas considérés jusqu'à présent, est telle que

$$\lim_{h\to 0} \Phi\left(t_n, y(t_n), f\left(t_n, y(t_n)\right); h\right) = f\left(t_n, y(t_n)\right) \quad \forall t_n \geq t_0 \tag{1.18}$$

La propriété (1.18), jointe à la relation évidente $y(t_{n+1}) - y(t_n) = hy'\left(t_n\right) + \mathcal{O}\left(h^2\right)$, $\forall n \geq 0$, nous permet de déduire de (1.17) que $\lim_{h\to 0} R_n(h) = 0, 0 \leq n \leq N-1$, ce qui implique

$$\lim_{h\to 0} R(h) = 0$$

Cette dernière relation exprime la consistance de la méthode numérique avec le problème de Cauchy.

Propriété 1.1 *La méthode RK est coansistante si et seulemnt si $\sum_{i=1}^{d} b_i = 1$.*

Définition 1.8 ***(Ordre du schéma)*** *On dit qu'un schéma de discrétisation à N points est d'ordre p s'il existe $C > 0$, ne dépendant que de la solution exacte, tel que l'erreur de consistance satisfait la relation :*

$$\max_{1\leq n\leq N} |R_n| \leq Ch^p.$$

Par un développement de Taylor autour du point t_n, on montre que l'erreur de consistance pour la méthode d'Euler est un $\mathcal{O}(h)$.

1.5.2 Zéro-stabilité

Définition 1.9 *Soient y_n et z_n définies par*

$$y_{n+1} = y_n + h\Phi\left(t_n, y_n, f\left(t_n, y_n\right); h\right)$$
$$z_{n+1} = z_n + h\Phi\left(t_n, z_n, f\left(t_n, z_n\right); h\right) + h\delta_{n+1}$$

pour $n = 0, \ldots, N-1$ avec $z_0 = y_0 + \delta_0$. On dit que la méthode (1.6) *pour l'approximation du problème de Cauchy* (1.1) *est zero-stable, s'il existe $h_0 > 0, C > 0$ tels que $\forall h \in]\,0, h_0]$, $\forall \varepsilon > 0$ assez petit, si $|\delta_n| < \varepsilon$ $(0 \leq n \leq N)$, alors*

$$|y_n - z_n| \leq C\varepsilon. \tag{1.19}$$

Les constantes C et h_0 peuvent dependre des données du problème t_0, T, y_0 et f.

La zero-stabilité requiert donc que, sur un intervalle borné, la condition (1.19) soit vérifiée pour toute valeur $h \leq h_0$. Cette propriété concerne en particulier le comportement de la méthode numérique dans le cas limite $h \to 0$, ce qui justifie le nom de zéro-stabilité. La propriété (1.19) assure que la méthode numérique est peu sensible aux petites perturbations des données.

Théorème 1.3 ***(Zéro-stabilité)*** *Considérons la méthode explicite à un pas* (1.6) *pour la résolution numérique du probleme de Cauchy* (1.1). *On suppose que la fonction d'incrément Φ est lipschitzienne par rapport à sa seconde variable, avec une constante de Lipschitz L_Φ independante de h et des noeuds $t_j \in [t_0, t_0 + T]$, autrement dit : $\exists h_0 > 0, \exists L_\Phi > 0$ tels que $\forall h \in]\,0, h_0]$*

$$\left|\Phi\left(t_n, y_n, f\left(t_n, y_n\right); h\right) - \Phi\left(t_n, z_n, f\left(t_n, z_n\right); h\right)\right| \leq L_\phi \left|y_n - z_n\right|, 0 \leq n \leq N.$$

Alors, la méthode (1.6) *est zíro-stable.*

Preuve : Soient y_n et z_n solutions, réspectivement, des problèmes

$$\begin{cases} y_{n+1} = y_n + h\Phi\left(t_n, y_n, f\left(t_n, y_n\right), h\right), \\ y_0 \in \mathbb{R}, \end{cases} \tag{1.20}$$

et

$$\begin{cases} z_{n+1} = z_n + h\Phi\left(t_n, z_n, f\left(t_n, z_n\right), h\right) + h\delta_{n+1}, \\ z_0 = y_0 + \delta_0, \end{cases} \tag{1.21}$$

avec $|\delta_{n+1}| < \varepsilon$ et $|\delta_0| < \varepsilon$.
Posons $w_j = z_j - y_j$ pour $j = 0, 1, \cdots, N$.
En retranchant l'équation (1.20) dans l'équation (1.21), on obtient pour $j = 0, 1, \cdots, N-1$

$$w_{j+1} = w_j + h\left(\Phi\left(t_j, z_j, f\left(t_j, z_j\right), h\right) - \Phi\left(t_j, y_j, f\left(t_j, y_j\right), h\right)\right) + h\delta_{j+1}.$$

En sommant sur j on obtient, pour $n \geq 1$

$$w_n = w_0 + h\sum_{j=0}^{n-1}\delta_{j+1} + h\sum_{j=0}^{n-1}\left(\Phi\left(t_j, z_j, f\left(t_j, z_j\right), h\right) - \Phi\left(t_j, y_j, f\left(t_j, y_j\right), h\right)\right).$$

On déduit que

$$|w_n| \leq |w_0| + h\sum_{j=0}^{n-1}|\delta_{j+1}| + hL_\Phi\sum_{j=0}^{n-1}|w_j|, \text{ pour } n \geq 1.$$

Pour conclure, nous avons besoin du lemme suivant

Lemme 1.2 *(**Lemme de Gronwall discret**) Soit k_n une suite réels positfs et φ_n une suite telle que*

$$\begin{cases} \varphi_0 \leq g_0, \\ \varphi_n \leq g_0 + \sum_{s=0}^{n-1} p_s + \sum_{s=0}^{n-1} k_s\varphi_s, \ n \geq 1. \end{cases}$$

Si $g_0 \geq 0$ et $p_n \geq 0$ pour tout $n \geq 0$, alors

$$\varphi_{\mathrm{n}} \leq \left(g_0 + \sum_{s=0}^{n-1} p_s\right)\exp\left(\sum_{s=0}^{n-1} k_s\right), \ n \geq 1.$$

En appliquant le lemme de Gronwall discret avec $\varphi_n = |w_n|, g_0 = |w_0| < \varepsilon, p_s = h|\delta_n|$ et $k_s = hL_\Phi$, on obtient

$$\begin{aligned} |w_n| &\leq (1+nh)\varepsilon\exp\left(nhL_\Phi\right), n \geq 1, \\ &\leq (1+T)\varepsilon\exp\left(L_\Phi T\right). \end{aligned}$$

Par conséquent

$$|z_n - y_n| \leq C\varepsilon, \text{ avec } C = (1+T)\exp\left(L_\Phi T\right).$$

D'où la zéro-stabilite du schéma (1.6).

1.5.3 Analyse de la convergence

Définition 1.10 *Une méthode est dite convergente si*

$$\forall n = 0, \cdots, N, \quad |y\left(t_n\right) - y_n| \leq C(h),$$

où $C(h)$ est un infinement petit en h. Si de plus, il existe une constante $\tilde{C} > 0$ indépendante de h telle que $C(h) \leq \bar{C}h^p$, on dit que la méthode est convergente d'ordre p.

On peut montrer le théorème suivant :

Théorème 1.4 *(**Convergence**) Sous les mêmes hypotheses qu'au théoreme 1.3 on a*

$$|y\left(t_n\right) - y_n| \leq \left(|y\left(t_0\right) - y_0| + nhR(h)\right)\exp\left(nhL_\phi\right), 1 \leq n \leq N$$

où $R(h) = \max_{0\leq n\leq N-1}|R_{n+1}|$.

Preuve : (Exercice) Ind. Considérer les suites $y_{\mathrm{n}}, y\left(t_{\mathrm{n}}\right)$ solutions des problèmes

$$\begin{cases} y_{n+1} = y_n + h\Phi\left(t_n, y_n, f\left(t_n, y_n\right); h\right), \\ y_0 \in \mathbb{R}, \end{cases}$$

et

$$\begin{cases} y\left(t_{n+1}\right) = y\left(t_n\right) + h\Phi\left(t_n, y\left(t_n\right), f\left(t_n, y\left(t_n\right); h\right) + hR_{n+1}, \\ y\left(t_0\right) = y_0 + \delta_0. \end{cases}$$

Ensuite utiliser le lemme de Gronwall discret. Une conséquence de ce thèorème est que, si l'hypothèse de consistance est vérifiée, alors

$$\begin{aligned} \left|y\left(t_n\right) - y_n\right| &\leq T\exp\left(TL_\phi\right)R(h), 1 \leq n \leq N.(nh \leq T), \\ &\leq Ch^p. \end{aligned}$$

Le shéma est donc convergent d'ordre de convergence égale à p.

Corrollaire 1.1 *Un schéma numérique consistant et stable est convergent. La réciproque de ce résultat : un schéma convergent est stable.*

Le résultat suivant établit une relation entre l'ordre et le nombre d'étapes des méthodes RK explicites.

Propriété 1.2 *L'ordre d'une méthode RK explicite à d étapes ne peut pas être plus grand que d. De plus, il n'existe pas de méthode à d étapes d'ordre d si $d \geq 5$.*

En particulier, pour des ordres allant de 1 à 8 , on a fait figurer dans le tableau ci-après le nombre minimum d'étapes $d_{\min}$ requis pour avoir une méthode d'ordre donnée.

ordre	1	2	3	4	**5**	6	7	8
$d_{\min}$	1	2	3	4	**6**	7	9	11

1.5.4 Stabilité absolue

La propriété de stabilité absolue est, d'une certaine manière, la contrepartie de la zéro-stabilité du point de vue des rôles respectifs de h et I. De façon heuristique, on dit qu'une méthode numérique est absolument stable si, pour un h fixé, y_n reste borné quand $t_n \to +\infty$. Une méthode absolument stable offre donc une garantie sur le comportement asymptotique de y_n, alors qu'une méthode zéro-stable assure que, pour un intervalle d'intégration fixé, y_n demeure borné quand $h \to 0$. Considérons le problème de Cauchy linéaire (que nous appellerons dorénavant problème test)

$$\begin{cases} y'(t) = \lambda y(t), \quad t > 0, \\ y(0) = 1, \end{cases} \tag{1.22}$$

avec $\lambda \in \mathbb{C}$, dont la solution est $y(t) = e^{\lambda t}$. Remarquer que si $\mathrm{Re}(\lambda) < 0$ alors $\lim_{t\to+\infty} |y(t)| = 0$.

Définition 1.11 *Une méthode numérique pour l'approximation de* (1.22) *est absolument stable si*

$$|y_n| \longrightarrow 0 \quad \textit{quand} \quad t_n \longrightarrow +\infty. \tag{1.23}$$

La région de stabilité absolue de la méthode numérique est le sous-ensemble du plan complexe

$$\mathcal{A} = \{z = h\lambda \in \mathbb{C} \;\; t.q. \;\; (1.23) \textit{ est vérifiée } \}. \tag{1.24}$$

Ainsi, $\mathcal{A}$ est l'ensemble des valeurs prises par le produit $h\lambda$ pour lesquelles la méthode numérique donne des solutions qui tendent vers zéro quand t_n tend vers l'infini. Cette définition a bien un sens car y_n est une fonction de $h\lambda$.

Remarque 1.6 *Considérons le problème de Cauchy général* (1.1) *et supposons qu'il existe deux constantes strictement positives $\mu_{\min}$ et $\mu_{\max}$ telles que*

$$-\mu_{\max} < \frac{\partial f}{\partial y}(t, y(t)) < -\mu_{\min} \quad \forall t \in I.$$

Alors, $-\mu_{\max}$ est un bon candidat pour jouer le rôle de λ dans l'analyse de stabilité ci-dessus.

Vérifions si les méthodes à un pas introduites précédemment sont absolument stables.
- Méthode d'Euler progressive : Le schéma d'Euler progressive appliqué au problème (1.22) donne $y_{n+1} = y_n + h\lambda y_n$ pour $n \geq 0$, avec $y_0 = 1$. En procédant par récurrence sur n, on a

$$y_n = (1 + h\lambda)^n, \quad n \geq 0.$$

Ainsi, la condition (1.23) est satisfaite si et seulement si $|1 + h\lambda| < 1$, c'està-dire si $h\lambda$ se situe à l'intérieur du disque unité centré en $(-1, 0)$ (voir Figure 1.4), ce qui revient à

$$h\lambda \in \mathbb{C}^- \quad \text{et} \quad 0 < h < -\frac{2\,\mathrm{Re}(\lambda)}{|\lambda|^2}, \tag{1.25}$$

où

$$\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Re}(z) < 0\}.$$

Exemple 1.4 *Pour le problème de Cauchy $y'(x) = -5y(x)$ avec $x > 0$ et $y(0) = 1$, la condition* (1.25) *implique $0 < h < 2/5$. La Figure 1.3 montre, à gauche, le comportement des solutions calculées avec deux valeurs de h qui ne remplissent pas cette condition, et, à droite, les solutions obtenues avec deux valeurs de h qui la satisfont. Remarquer que dans ce second cas les oscillations, quand elles sont présentent, s'atténuent quand t croît.*

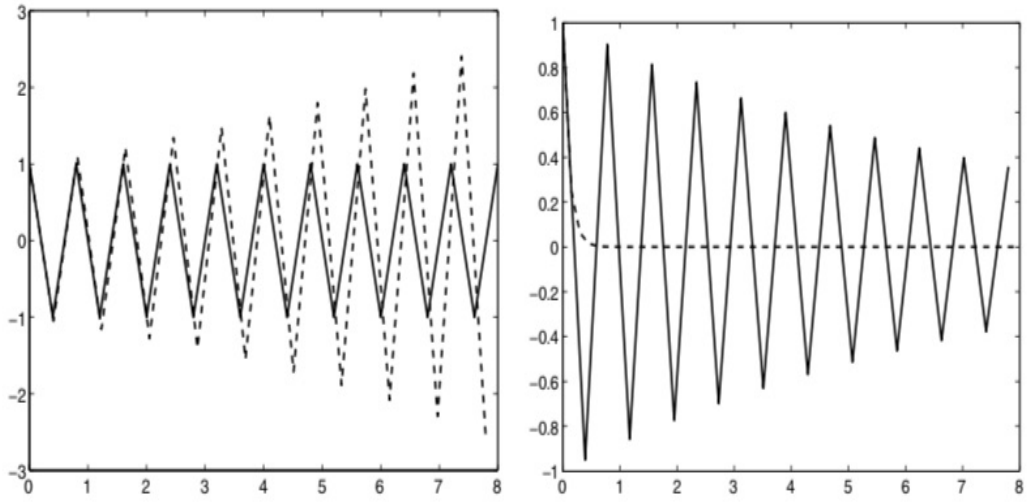


FIGURE 1.3 – A gauche : solutions calculées pour $h = 0.41 > 2/5$ (trait discontinu) et $h = 2/5$ (trait plein). Remarquer que, dans le cas limite $h = 2/5$, les oscillations n'évoluent pas quand t augmente. A droite : solutions correspondant à $h = 0.39$ (trait plein) et $h = 0.15$ (trait discontinu)

- Méthode d'Euler rétrograde : En procédant comme précédemment, on obtient cette fois

$$y_n = \frac{1}{(1-h\lambda)^n}, \quad n \geq 0.$$

On a la propriété de stabilité absolue (1.23) pour toute valeur $h\lambda$ qui n'appartient pas au disque unité centré en $(1, 0)$ (voir Figure 1.4).

- Méthode du trapèze (ou de Crank-Nicolson) : On obtient

$$y_n = \left[\left(1 + \frac{1}{2}\lambda h\right) / \left(1 - \frac{1}{2}\lambda h\right)\right]^n, \quad n \geq 0,$$

la propriété (1.23) est donc satisfaite pour tout $h\lambda \in \mathbb{C}^-$.

- Méthode de Heun : en appliquant la méthode de Heun d'odre 2 au problème (1.22) et en procédant par récurrence sur n, on trouve

$$y_n = \left[1 + h\lambda + \frac{(h\lambda)^2}{2}\right]^n, \quad n \geq 0.$$

Comme le montre la Figure 1.4, la région de stabilité absolue de la méthode de Heun est plus grande que celle de la méthode d'Euler progressive. Néanmoins, leurs restrictions à l'axe réel coïncident.

Définition 1.12 *On dit qu'une méthode est A-stable si $\mathcal{A} \cap \mathbb{C}^- = \mathbb{C}^-$, c'est-à-dire si pour* $\mathrm{Re}(\lambda) < 0$, *la condition* (1.23) *est satisfaite pour toute valeur de h.*

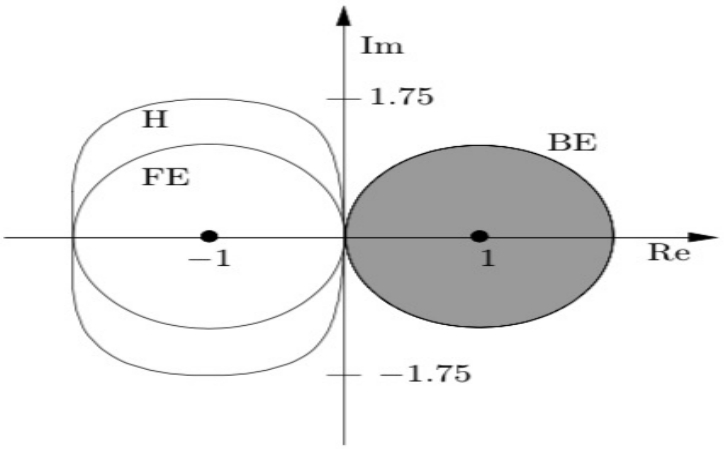


FIGURE 1.4 – Régions de stabilité absolue pour les méthodes d'Euler progressive (notée **FE** pour forward Euler), d'Euler rétrograde (notée **BE** pour backward Euler et de Heun (notée **H**). La région de stabilité absolue de la méthode d'Euler rétrograde se situe à l'extérieur du disque unité centré en $(1, 0)$ (zone grisée).

Remarque 1.7 *Les méthodes d'Euler rétrograde et de Crank-Nicolson sont A-stables, tandis que les méthodes d'Euler progressive et de Heun sont conditionnellement stables.*

Remarque 1.8 *Remarquer que (quand* $\mathrm{Re}(\lambda) \leq 0$ *dans* (1.22)*) les méthodes implicites à un pas examinées jusqu'à présent sont inconditionnellement absolument stables, alors que les schémas explicites sont conditionnellement absolument stables. Ceci n'est cependant pas une règle générale : il existe des schémas implicites instables ou seulement conditionnellement stables. En revanche, il n'y a pas de schéma explicite inconditionnellement absolument stable.*

1.6 Méthodes multi-pas "Méthodes d'Adams"

On déduit ces méthodes de la forme intégrale (1.2) en évaluant de manière approchée l'intégrale de f entre t_n et t_{n+1}. On suppose les noeuds de discrétisation équirépartis, c'est-à-dire $t_j = t_0 + jh$, avec $h > 0$ et $j \geq 1$. On intègre alors, au lieu de f, son polynôme d'interpolation aux $\tilde{p} + \vartheta$ noeuds distincts, où $\vartheta = 1$ quand les méthodes sont explicites (dans ce cas $\tilde{p} \geq 0$) et $\vartheta = 2$ quand les méthodes sont implicites (dans ce cas $\tilde{p} \geq -1$). Les schémas obtenus ont la forme suivante

$$y_{n+1} = y_n + h \sum_{j=-1}^{\tilde{p}} b_j f_{n-j}. \tag{1.26}$$

Les noeuds d'interpolation peuvent être ou bien :
1. $t_n, t_{n-1}, \ldots, t_{n-\tilde{p}}$ (dans ce cas $b_{-1} = 0$ et la méthode est explicite) ;
ou bien

2. $t_{n+1}, t_n, \ldots, t_{n-\tilde{p}}$ (dans ce cas $b_{-1} \neq 0$ et le schéma est implicite).
Ces schémas **implicites** sont appelés méthodes d'**Adams-Moulton**, et les **explicites** sont appelés méthodes d'**Adams-Bashforth**.

1.6.1 Méthodes d'Adams-Bashforth

En prenant $\tilde{p} = 0$, on retrouve la méthode d'Euler progressive, puisque le polynôme d'interpolation de degré zéro au noeud t_n est simplement $\Pi_0 f = f_n$. Pour $\tilde{p} = 1$, le polynôme d'interpolation linéaire aux noeuds t_{n-1} et t_n est

$$\Pi_1 f(t) = f_n + (t - t_n) \frac{f_{n-1} - f_n}{t_{n-1} - t_n}.$$

Comme $\Pi_1 f\,(t_n) = f_n$ et $\Pi_1 f\,(t_{n+1}) = 2f_n - f_{n-1}$, on obtient

$$\int_{t_n}^{t_{n+1}} \Pi_1 f(t) = \frac{h}{2} \left[\Pi_1 f\,(t_n) + \Pi_1 f\,(t_{n+1})\right] = \frac{h}{2} \left[3f_n - f_{n-1}\right].$$

Le schéma d'Adams-Bashforth à deux pas est donc

$$y_{n+1} = y_n + \frac{h}{2} \left[3f_n - f_{n-1}\right]. \tag{1.27}$$

Si $\tilde{p} = 2$, on trouve de façon analogue le schéma d'Adams-Bashforth à trois pas

$$y_{n+1} = y_n + \frac{h}{12} \left[23f_n - 16f_{n-1} + 5f_{n-2}\right],$$

et pour $\tilde{p} = 3$, on a le schéma d'Adams-Bashforth à quatre pas

$$y_{n+1} = y_n + \frac{h}{24} \left(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}\right).$$

Remarque 1.9 *Les schémas d'Adams-Bashforth utilisent $\tilde{p}+1$ noeuds et sont des méthodes à $\tilde{p}+1$ pas (avec $\tilde{p} \geq 0$). De plus, les schémas d'Adams-Bashforth à q pas sont d'ordre q.*

1.6.2 Méthodes d'Adams-Moulton

Si $\tilde{p} = -1$, on retrouve le schéma d'Euler rétrograde. Si $\tilde{p} = 0$, on construit le polynôme d'interpolation de degré un de f aux noeuds t_n et t_{n+1}, et on retrouve le schéma de Crank-Nicolson. Pour la méthode à deux pas ($\tilde{p} = 1$), on construit le polynôme d'interpolation de degré 2 de f aux noeuds t_{n-1}, t_n, t_{n+1}, et on obtient le schéma suivant

$$y_{n+1} = y_n + \frac{h}{12} \left[5f_{n+1} + 8f_n - f_{n-1}\right].$$

Les schémas correspondant à $\tilde{p} = 2$ et 3 sont respectivement donnés par

$$y_{n+1} = y_n + \frac{h}{24}\left(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}\right),$$

et

$$y_{n+1} = y_n + \frac{h}{720}\left(251f_{n+1} + 646f_n - 264f_{n-1} + 106f_{n-2} - 19f_{n-3}\right).$$

Les schémas d'Adams-Moulton utilisent $\tilde{p}+2$ noeuds et sont à $\tilde{p}+1$ pas si $\tilde{p} \geq 0$, la seule exception étant le schéma d'Euler rétrograde ($\tilde{p} = -1$) qui est à un pas et utilise un noeud. Les schémas d'Adams-Moulton à q pas sont d'ordre $q+1$, excepté à nouveau le schéma d'Euler rétrograde qui est une méthode à un pas d'ordre un.

1.7 Systèmes d'équations différentielles ordinaires

Considérons le système d'équations différentielles ordinaires du premier ordre

$$\mathbf{y}' = \mathbf{F}(t, \mathbf{y}), \tag{1.28}$$

où $\mathbf{F} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ est une fonction vectorielle donnée et $\mathbf{y} \in \mathbb{R}^n$ est le vecteur solution qui dépend de n constantes arbitraires fixées par les n conditions initiales

$$\mathbf{y}\left(t_0\right) = \mathbf{y}_0. \tag{1.29}$$

Rappelons la propriété suivante :

Propriété 1.3 *Soit $\mathbf{F} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ une fonction continue sur $D = [t_0, t_0 + T] \times \mathbb{R}^n$, où t_0 et T sont finis. S'il existe une constante positive L telle que*

$$\|\mathbf{F}(t, \mathbf{y}) - \mathbf{F}(t, \overline{\mathbf{y}})\| \leq L\|\mathbf{y} - \overline{\mathbf{y}}\|, \tag{1.30}$$

pour tout $(t, \mathbf{y})$ et $(t, \overline{\mathbf{y}}) \in D$, alors, pour tout $\mathbf{y}_0 \in \mathbb{R}^n$ il existe une unique fonction $\mathbf{y}$, continue et différentiable par rapport à t, solution du problème de Cauchy (1.28)-(1.29). *La condition* (1.30) *exprime le fait que $\mathbf{F}$ est lipschitzienne par rapport à sa seconde variable.*

Il est rarement possible d'écrire de manière analytique la solution du système (1.28), mais on peut le faire dans des cas très particuliers, par exemple quand le système est de la forme

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t),$$

avec $\mathbf{A} \in \mathbb{R}^{n \times n}$. On suppose que $\mathbf{A}$ possède n valeurs propres distinctes λ_j, $j = 1, \ldots, n$. La solution $\mathbf{y}$ est alors donnée par

$$\mathbf{y}(t) = \sum_{j=1}^{n} C_j e^{\lambda_j t}\mathbf{v}_j,$$

où $C_1, \ldots, C_n$ sont des constantes et $\{\mathbf{v}_j\}$ est une base formée des vecteurs propres de $\mathbf{A}$ associés aux valeurs propres λ_j pour $j = 1, \ldots, n$. La solution est déterminée en se donnant n conditions initiales.

Remarque 1.10 *Du point de vue numérique, les méthodes introduites dans le cas scalaire peuvent être étendues aux systèmes.*

Remarque 1.11 *Un système différentiel faisant intervenir des différentielles d'ordre supérieur peut toujours s'écrire sous la forme* (1.28) .

Exemple 1.5 *Prenons par exemple l'équation du second ordre décrivant le comportement de l'amortisseur d'une voiture :*

$$\begin{cases} my'' + cy' + ky = 0, \\ y(0) = \bar{x}_0, \\ y'(0) = 0, \end{cases}$$

où m est la masse de la voiture, c le coefficient d'amortissement et k la force de rappel. L'inconnue y est le déplacement de l'amortisseur par rapport à sa position d'équilibre. Pour se ramener à un système d'ordre 1, on pose $x_1 = y$, $x_2 = y'$, et le système amortisseur s'écrit alors, avec comme inconnue $x = (x_1, x_2)^t$:

$$\begin{cases} \mathbf{x}'(\mathbf{t}) = \mathbf{F}(\mathbf{t}, \mathbf{x}(\mathbf{t})), \\ \mathbf{x}(\mathbf{0}) = (\bar{x}_0, 0)^t, \end{cases} \qquad \textit{avec} \qquad \mathbf{F}(\mathbf{t}, \mathbf{x}(\mathbf{t})) = \begin{pmatrix} x_2 \\ -\dfrac{1}{m}\left(cx_2 + kx_1\right) \end{pmatrix}.$$

Chapitre 2

Méthode des différences finies

2.1 Introduction

On peut difficilement étudier les équations aux dérivées partielles (E.D.P.) dans une totale généralité comme on peut le faire pour les équations différentielles ordinaires (E.D.O.). Heureusement les E.D.P. les plus intéressantes proviennent de la modélisation d'un nombre restreint de phénomènes :

- Le transport : convection d'un polluant dans l'atmosphère ou dans un milieu poreux.
- La diffusion : diffusion de la chaleur dans un solide.
- Les vibrations : son dans l'air.

Chacun de ces phénomènes correspond à une catégorie d'E.D.P. Ce chapitre fait l'objet des E.D.P. linéaires d'ordre inférieur ou égale à 2.

2.2 Définitions et exemples des EDPs

Définition 2.1 *On appelle E.D.P. une équation comprenant au moins une dérivée partielle d'une fonction de plusieurs variables.*

Définition 2.2 *On appelle E.D.P. linéaire, d'inconnue* $u : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$*, une équation de la forme*

$$F(u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial y^2}, \frac{\partial^2 u}{\partial x \partial y}, \cdots, x, y) = 0,$$

où la fonction F *est linéaire par rapport à ses variables, sauf les deux dérnières* x *et* y*. L'ordre d'une E.D.P. est l'ordre de la dérivée de plus haut degré de l'E.D.P.*

Exemple 2.1

- $\frac{\partial^2 u}{\partial x^2} + 3\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial u}{\partial x} + u = 0$ *est une E.D.P. linéaire d'ordre* 2 *à coefficients constants.*
- $sin(xy)\frac{\partial^2 u}{\partial x^2} + 3x^2\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial u}{\partial x} = 0$ *est une E.D.P. linéaire d'ordre* 2 *à coefficients variables.*
- $\frac{\partial^2 u}{\partial x^2} + 3x^2\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} + (\frac{\partial u}{\partial x})^2 = 0$ *est une E.D.P. non linéaire d'ordre* 2.

Définition 2.3 *On appelle E.D.P. linéaire d'ordre* 1 *dans un domaine* $\Omega \subset \mathbb{R}^N$ *et d'inconnue* $u : \Omega \to \mathbb{R}$ *une E.D.P. de la forme :*

$$\beta(x) \cdot \nabla u(x) + \alpha(x)u(x) = f(x)$$

où $\nabla u(x)$ *est le vecteur gradient de* u, $(\nabla u(x))_{i=1,\dots,N} = \frac{\partial u}{\partial x_i}$, $\beta(x) = (\beta_1(x), \beta_2(x), \cdots, \beta_N(x))$ *et* $f(x)$ *un terme source donné.*

Exemple 2.2 *L'équation de continuité de la mécanique des fluides :*

$$\frac{\partial \rho}{\partial t} + div(\rho(t,x)\beta(t,x)) = 0,$$

pour une vitesse $\beta : [0,T] \times \mathbb{R}^3 \to \mathbb{R}^3$ *connue, l'inconnue est la densité* $\rho : [0,T] \times \mathbb{R}^3 \to \mathbb{R}$. *En effet l'équation se réecrit :*

$$\frac{\partial \rho}{\partial t} + \beta \cdot \nabla \rho + \rho div(\beta) = 0,$$

Définition 2.4 *On appellle E.D.P. linéaire d'ordre inférieur ou égale à* 2 *dans un domaine* $\Omega \subset \mathbb{R}^N$ *et d'inconnue* $u : \Omega \to \mathbb{R}$, *une équation de type*

$$\sum_{i,j=1}^{N} a_{ij}(x)\frac{\partial^2 u(x)}{\partial x_i \partial x_j} + \sum_{i=1}^{N} \beta_i(x)\frac{\partial u}{\partial x_i} + \alpha(x)u(x) = f(x). \tag{2.1}$$

On supposera que $a_{ij}(x) = a_{ji}(x)$.

$f(x)$ est souvent appelé terme source de l'équation.
On introduit les notations suivantes :

- $A(x) = (a_{ij}(x))_{1 \le i,j \le N}$ la matrice $N \times N$ symétrique des coefficients devant les termes d'ordre 2.
- $\beta(x) = (\beta_i(x))_{1 \le i \le N}$ le vecteur de taille N des coefficients devant les termes d'ordre 1.

- $\nabla u(x)$ le vecteur gradient de u, $\nabla u(x) = (\dfrac{\partial u}{\partial x_i})_{1\leq i\leq N}$.

- $Hu(x)$ la matrice Hessienne de u, $Hu(x) = (\dfrac{\partial^2 u(x)}{\partial x_i \partial x_j})_{1\leq i,j\leq N}$.

Ainsi que la notation

$$A : B = \sum_{i,j=1}^{N} a_{ij} b_{ij}$$

appelée produit scalaire de **Frobenius**, pour A et B deux matrices de composantes respectivement a_{ij} et b_{ij}.
Avec ces notations, l'E.D.P. d'ordre 2 se réecrit :

$$A(x) : Hu(x) + \beta(x) \cdot \nabla u(x) + \alpha(x) u(x) = f(x) \tag{2.2}$$

Exemple 2.3 *Soit l'E.D.P. suivante :*

$$\frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial x \partial y}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y) + \frac{\partial u}{\partial x}(x,y) = sin(\pi xy), \ \textit{pour } x, y \in \mathbb{R}.$$

Cette équation s'écrit sous la forme (2.2) avec

$$A = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}, \ \beta^t = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ \alpha(x,y) = 0, \ f(x,y) = sin(\pi xy).$$

2.3 Classification des E.D.P. linéaires du second ordre

On peut classer les E.D.P. linéaires du second ordre en trois grandes familles : "**elliptique**'', "**parabolique**" et "**hyperbolique**".

2.3.1 E.D.P. elliptiques

Définition 2.5 *Une E.D.P. linéaire du second ordre est dite elliptique en $x \in \Omega$ si la matrice $A(x)$ n'admet que des valeurs propres non nulles et qui sont toutes de même signe.*

Si l'E.D.P. est elliptique dans tout le domaine Ω, elle modélise en général un problème d'équilibre ou un problème stationnaire.

Exemple 2.4 *L'équation*

$$-\Delta u(x) = f(x),$$

(où $\Delta u(x) = \sum_{i=1}^{N} \frac{\partial^2 u(x)}{\partial x_i^2}$ *est le laplacien de* $u(x)$*) est elliptique, en effet toutes les valeurs propres de la matrice* $A(x)$ *sont égales à* -1*.*
Plus généralement, les équations de la forme :

$$-div(k(x)\nabla u(x)) = f(x),$$

avec $k(x) > 0$ *sont elliptiques.*

2.3.2 E.D.P. paraboliques

Définition 2.6 *Une E.D.P. linéaire du second ordre est dite parabolique en* $x \in \Omega$ *si la matrice* $A(x)$ *admet* $N-1$ *valeurs propres non nulles de même signe et une valeur propre nulle. De plus, soit* $v(x)$ *un vecteur propre associé à la valeur propre nulle, on doit avoir* $v(x) \cdot \beta(x) \neq 0$*.*

Si l'E.D.P. est parabolique sur tout le domaine Ω, elle modélise en général un phènomène de diffusion.

Exemple 2.5 *L'équation de la chaleur*

$$\frac{\partial u}{\partial t}(t,x) - k\Delta u(t,x) = f(t,x) \text{ pour } t \geq 0,\ x \in \mathbb{R}^N.$$

est parabolique, en effet la matrice $A(x)$ *admet une valeur propre nulle, toutes les autres valant* $-k$*. De plus* $\beta = (1, 0, \cdots, 0)^T$*, la condition* $v(t,x) \cdot \beta \neq 0$ *est vérifiée.*
Plus généralement, les équations de la forme :

$$\frac{\partial u}{\partial t}(t,x) - div(k(x)\nabla u(t,x)) = f(t,x) \text{ pour } t \geq 0,\ x \in \mathbb{R}^N,$$

avec $k(x) > 0$ *sont paraboliques.*

2.3.3 E.D.P. hyperboliques

Définition 2.7 *Une E.D.P. linéaire du second ordre est dite hyperbolique en* $x \in \Omega$ *si la matrice* $A(x)$ *n'admet que des valeurs propres non nulles et qui sont toutes de même signe sauf une de signe opposé.*

Si l'E.D.P. est hyperbolique sur tout le domaine Ω, elle modélise en général un problème de propagation d'onde.

Exemple 2.6 *L'équation d'onde*

$$\frac{\partial^2 u}{\partial t^2}(t,x) - c^2\Delta u(t,x) = f(t,x) \text{ pour } t \geq 0,\ x \in \mathbb{R}^N.$$

(où $\Delta u(t,x) = \sum_{i=1}^{N} \frac{\partial^2 u(x)}{\partial x_i^2}$ *est le laplacien de* $u(t,x)$ *par rapport à* x*) est hyperbolique, en effet la matrice* $A(x)$ *à comme valeurs propres* 1 *de multiplicité* 1 *et* $-c^2$ *de multiplicité* N*.*
Plus généralement, les équations de la forme :

$$\frac{\partial^2 u}{\partial t^2}(t,x) - div(c(x)\nabla u(t,x)) = f(t,x) \text{ pour } t \geq 0,\ x \in \mathbb{R}^N,$$

avec $c(x) > 0$ *sont hyperboliques.*

Remarque 2.1 *La terminologie "elliptique", "parabolique" et "hyperbolique" vient du fait que lorsque la matrice* $A(x)$ *est constante, les courbes* $x^T Ax = cte$ *sont respectivement des ellipsoïdes, des paraboloïdes et des hyperboloïdes.*

2.3.4 E.D.P. linéaires du second ordre dans $\mathbb{R}^2$

Dans cette section, on présente quelques éléments de classification des EDPs linéaires du second ordre portant sur des fonctions de deux variables réelles $u(x,y)$. Une telle équation s'écrit

$$A\partial^2_{xx}u + 2B\partial^2_{xy}u + C\partial^2_{y}u + D\partial_x u + E\partial_y u + Fu = G, \tag{2.3}$$

où, pour simplifier A, B, C, D, E, F et G sont supposés constants. On lui associe son polynôme caractéristique en α et β :

$$P(\alpha,\beta) = A\alpha^2 + 2B\alpha\beta + C\beta^2 + D\alpha + E\beta + F.$$

Les propriétés de $P(\alpha,\beta)$ déterminent la nature de l'EDP. Selon le discrimant $B^2 - AC$, le polynôme et l'EDP sont dits :

- Elliptique si $B^2 - AC < 0$,
- Parabolique si $B^2 - AC = 0$,
- Hyperbolique si $B^2 - AC > 0$.

2.3.5 Conditions aux limites

Sur un domaine borné $\Omega \subset \mathbb{R}^N$ et de frontière suffisament régulière, en imposant des conditions aux limites sur le bord du domaine noté $\partial\Omega$, on peut aboutir à un problème bien posé : la solution existe et est unique.
Les condition aux limites peuvent être de type :
- **Dirichlet** si on impose, sur une partie de $\partial\Omega$,

$$u = g, \quad \text{sur } \Gamma_D \subset \partial\Omega.$$

- **Neumann** si on impose sur une partie de $\partial\Omega$,

$$\frac{\partial u}{\partial n} = g, \quad \text{sur } \Gamma_N \subset \partial\Omega.$$

où $\dfrac{\partial u}{\partial n} = \langle \operatorname{grad} u, \boldsymbol{n} \rangle$ avec $\boldsymbol{n}$ normale exterieure unitaire à Ω.
- **Robin** si on impose sur une partie de $\partial\Omega$

$$\frac{\partial u}{\partial n} + \alpha u = g, \quad \text{sur } \Gamma_R \subset \partial\Omega.$$

2.4 Méthode des différences finies

2.4.1 Principe de la méthode

- Cas de la dimension 1

On considère le problème unidimensionnel

$$-u''(x) = f(x), \quad \forall x \in]0,1[, \tag{2.4}$$

$$u(0) = u(1) = 0, \tag{2.5}$$

où $f \in C([0,1])$. Les conditions aux limites (2.5) considérées ici sont dites de type **Dirichlet homogène** (le terme homogène désigne les conditions nulles). Cette équation modélise par exemple la diffusion de la chaleur dans un barreau conducteur chauffé (terme source f) dont les deux extrémités sont plongées dans de la glace. Soit $(x_k)_{k=0,\ldots,N+1}$ une subdivision de $[0,1]$, avec :

$$x_0 = 0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = 1.$$

Pour $i = 0, \ldots, N$, on note $h_{i+1/2} = x_{i+1} - x_i$ et on définit le "pas" du maillage par :

$$h = \max_{i=0,\ldots,N} h_{i+1/2}. \tag{2.6}$$

Pour simplifier le cours, on se limitera à un pas constant :

$$h_{i+1/2} = h \quad \forall i \in [|0, N|].$$

On écrit l'équation aux dérivées partielles (2.4) aux points x_i

$$-u''(x_i) = f(x_i), \quad \forall i = 1, \ldots, N.$$

Effectuons un développement de Taylor en x_i :

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\zeta_i),$$
$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\eta_i),$$

avec $\zeta_i \in]x_i, x_{i+1}[$, $\eta_i \in]x_{i-1}, x_i[$,. En additionnant, on obtient :

$$u(x_{i+1}) + u(x_{i-1}) = 2u(x_i) + h^2 u''(x_i) + O\left(h^4\right).$$

Il semble donc raisonnable d'approcher la dérivée seconde $-u''(x_i)$ par le "quotient différentiel"

$$\frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1})}{h^2}.$$

Sous des hypothèses de régularité sur u, on peut montrer (voir la sousection suivante) que cette approximation est d'ordre 2 au sens

$$R_i = u''(x_i) + \frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1})}{h^2} = O\left(h^2\right).$$

On appelle erreur de consistance au point x_i la quantité R_i.

- Cas de la dimension 2

On considère le problème bidimensionnel

$$-\Delta u(x) = f(x), \quad \forall x \in \Omega, \tag{2.7}$$

$$u(x) = 0 \quad \forall x \in \partial\Omega, \tag{2.8}$$

où Ω est un ouvert de $\mathbb{R}^2$ et $\partial\Omega$ désigne la frontière de Ω.
Supposons (pour simplifier) que le domaine Ω soit un carré (le cas rectangulaire se traite tout aussi facilement). On se donne un pas de maillage constant h et des points $x_{i,j} = (ih, jh), i = 0, \ldots, N+1,\ j = 0, \ldots, N+1$. En effectuant les développements limités de Taylor dans les deux directions, on approche $-\partial_i^2 u(x_{i,j})$ (resp. $-\partial_j^2 u(x_{i,j})\big)$ par

$$\frac{2u(x_{i,j}) - u(x_{i+1,j}) - u(x_{i-1,j})}{h^2} \left(\text{ resp. par } \frac{2u(x_{i,j}) - u(x_{i,j+1}) - u(x_{i,j-1})}{h^2} \right).$$

- Questions d'analyse numérique

Voici les questions auxquelles nous tenterons de répondre dans la suite :

1. Le problème qu'on a obtenu en dimension finie, (avec des inconnues localisées aux noeuds du maillage) admet-il une (unique) solution ?

2. La solution du problème discret converge-t-elle vers la solution du problème continu lorsque le pas du maillage h tend vers 0 ?

2.4.2 Etude de la méthode différences finies pour un problème stationnaire unidimensionnel

On prend comme modèle de problème stationnaire l'équation elliptique suivante :

$$\begin{cases} -u''(x) + c(x)u(x) = f(x), \quad 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \tag{2.9}$$

où $c \in C([0,1], \mathbb{R}_+)$, et $f \in C([0,1], \mathbb{R})$, qui peut modéliser par exemple un phénomène de diffusion réaction d'une espèce chimique. On se donne un pas du maillage constant $h = \frac{1}{N+1}$, et une subdivision de $]\, 0, 1\,[$, notée $(x_k)_{k=0,\ldots,N+1}$, avec : $x_0 = 0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = 1$. Soit u_i l'inconnue discrète associée au noeud $i \quad (i = 1, \ldots, N)$. On pose $u_0 = u_{N+1} = 0$. On obtient les équations discrètes en approchant $u''(x_i)$ par quotient différentiel par développement de Taylor, comme on l'a vu précédemment.

$$\begin{cases} \dfrac{1}{h^2}(2u_i - u_{i-1} - u_{i+1}) + c_i u_i = f_i, \quad i = 1, \ldots, N, \\ u_0 = u_{N+1} = 0, \end{cases} \tag{2.10}$$

avec $c_i = c(x_i)$ et $f_i = f(x_i)$. On peut écrire ces équations sous forme matricielle :

$$A_h U_h = b_h, \text{ avec } U_h = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix} \text{ et } b_h = \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix} \tag{2.11}$$

$$\text{et } A_h = \frac{1}{h^2} \begin{pmatrix} 2 + c_1 h^2 & -1 & 0 & \ldots & 0 \\ -1 & 2 + c_2 h^2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 + c_{N-1} h^2 & -1 \\ 0 & \ldots & 0 & -1 & 2 + c_N h^2 \end{pmatrix}. \tag{2.12}$$

Dans la suite nous allons répondre aux questions suivantes :

1. Le système (2.11) admet-il une unique solution ?

2. A-t-on la convergence de U_h vers u et en quel sens ?

Proposition 2.1 (existence et unicité de la solution) *Soit $c = (c_1, \ldots, c_N)^t \in \mathbb{R}^N$ tel que $c_i \geq 0$ pour $i = 1, \ldots, N$; alors la matrice A_h définie par (2.12) est symétrique définie positive, et donc inversible.*

Démonstration:
La matrice A_h est évidemment symétrique. Montrons qu'elle est définie positive. Soit $v =$

$(v_1 \dots v_N)^t$, on pose $v_0 = v_{N+1} = 0$. Calculons le produit scalaire $A_h v \cdot v = v^t A_h v$. On a :

$$A_h v \cdot v = \frac{1}{h^2} (v_1 \dots v_N) \begin{pmatrix} 2 + c_1 h^2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & & -1 \\ 0 & & -1 & 2 + c_N h^2 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ \vdots \\ v_N \end{pmatrix},$$

c'est-à -dire :

$$A_h v \cdot v = \frac{1}{h^2} \sum_{i=1}^{N} v_i \left(-v_{i-1} + \left(2 + c_i h^2\right) v_i - v_{i+1} \right).$$

On a donc, par changement d'indice :

$$A_h v \cdot v = \frac{1}{h^2} \left[\sum_{i=1}^{N} (-v_{i-1} v_i) + \sum_{i=1}^{N} \left(2 + c_i h^2\right) v_i^2 - \sum_{j=2}^{N+1} v_{j-1} v_j \right].$$

Et comme on a posé $v_0 = 0$ et $v_{N+1} = 0$, on peut écrire :

$$A_h v \cdot v = \frac{1}{h^2} \sum_{i=1}^{N} \left(2 + c_i h^2\right) v_i^2 + \frac{1}{h^2} \sum_{i=1}^{N} (-2 v_i v_{i-1}),$$

soit encore :

$$A_h v \cdot v = \sum_{i=1}^{N} c_i v_i^2 + \frac{1}{h^2} \sum_{i=1}^{N} \left(-2 v_i v_{i-1} + v_i^2 + v_{i-1}^2\right) + \frac{1}{h^2} v_N^2.$$

On a donc finalement :

$$A_h v \cdot v = \sum_{i=1}^{N} c_i v_i^2 + \frac{1}{h^2} \sum_{i=1}^{N} (v_i - v_{i-1})^2 + \frac{1}{h^2} v_N^2 \geq 0, \forall v = (v_1, \dots, v_N) \in \mathbb{R}^N.$$

Si on suppose $A_h v \cdot v = 0$, on a alors

$$v_N^2 = 0, \sum_{i=1}^{N} c_i v_i^2 = 0 \text{ et } (v_i - v_{i-1})^2 = 0, \quad \forall i = 1, \dots, N.$$

On a donc $v_1 = v_2 = \dots = v_N = v_0 = v_{N+1} = 0$. Remarquons que ces égalités sont vérifiées même si les c_i sont nuls. Ceci démontre que la matrice A_h est bien définie.
On a A_h est symétrique définie positive, donc inversible, ce qui entraine l'existence et l'unicité de la solution de (2.11). ■

Remarque 2.2 *On peut aussi démontrer l'existence et l'unicité de la solution de* (2.11) *directement, en montrant que* $\text{Ker}\,(A_h) = 0$. *On rappelle qu'en dimension finie, toute application linéaire injective ou surjective est bijective.*

Nous allons répondre maintenant à la question de la convergence.

Définition 2.8 (Matrices monotones) *Soit* $A \in \mathcal{M}_N(\mathbb{R}), A = (a_{ij})_{i=1,\dots,N}, j = 1, \dots, N.$

- *On dit que* A *est positive (ou* $A \geq 0$ *) si* $a_{ij} \geq 0, \quad \forall i, j = 1, \dots, N.$
- *On dit que* A *est monotone si* A *est inversible et* $A^{-1} \geq 0.$

L'avantage des schémas à matrices monotones est de satisfaire la propriété de conservation de la positivité, qui peut être cruciale dans les applications physiques :

Définition 2.9 (Conservation de la positivité)
Soit $A \in \mathcal{M}_N(\mathbb{R}), A = (a_{ij})_{i=1,\dots,N, j=1,\dots,N}$ *; on dit que* A *conserve la positivité si* $Av \geq 0$ *entraine* $v \geq 0.$

Proposition 2.2 (Monotonie et conservation de la positivité) *Soit* $A \in \mathcal{M}_N(\mathbb{R})$. *Alors* A *conserve la positivité si et seulement si* A *est monotone.*

Démonstration:
Supposons d'abord que A conserve la positivité, et montrons que A inversible et que A^{-1} a des coefficients ≥ 0. Si x est tel que $Ax = 0$, alors $Ax \geq 0$ et donc, par hypothèse, $x \geq 0$. Mais on a aussi $Ax \leq 0$, soit $A(-x) \geq 0$ et donc par hypothèse, $x \leq 0$. On en déduit $x = 0$, ce qui prouve que $\text{Ker}\,(A_h) = 0$ (remarque 2.2) donc A est inversible.
La conservation de la positivité donne alors que $y \geq 0 \Rightarrow A^{-1}y \geq 0$. En prenant $y = e_1$ on obtient que la première colonne de A^{-1} est positive, puis en prenant $y = e_i$ on obtient que la i-ème colonne de A^{-1} est positive, pour $i = 2, \dots, N$. Donc A^{-1} a tous ses coefficients positifs.
Réciproquement, supposons maintenant que A est inversible et que A^{-1} a des coefficients positifs. Soit $x \in \mathbb{R}^N$ tel que $Ax = y \geq 0$, alors $x = A^{-1}y \geq 0$. Donc A conserve la positivité. ■

Remarque 2.3 (Principe du maximum) *On appelle principe du maximum continu le fait que si* $f \geq 0$ *alors le minimum de la fonction* u *solution du problème* (2.9) *est atteint sur les bords.*

Lemme 2.1 *Soit* $c = (c_1, \dots, c_N)^t \in \mathbb{R}^N$, *et* $A_h \in \mathcal{M}_N(\mathbb{R})$ *définie par* (2.12). *Si* $c_i \geq 0$ *pour tout* $i = 1, \dots, N$, *alors* A_h *est monotone.*

Démonstration:
On va montrer que si $v \in \mathbb{R}^N, A_h v \geq 0$ alors $v \geq 0$. On peut alors utiliser la proposition 2.2 pour conclure. Soit $v = (v_1, \dots, v_N)^t \in \mathbb{R}^N$. Posons $v_0 = v_{N+1} = 0$. Supposons que $A_h v \geq 0$. On a donc

$$-\frac{1}{h^2}v_{i-1} + \left(\frac{2}{h^2} + c_i\right)v_i - \frac{1}{h^2}v_{i+1} \geq 0, \quad i = 1, \dots, N. \tag{2.13}$$

Soit

$$p = \min \left\{ i \in \{1, \ldots, N\}; v_p = \min_{j=1,\ldots,N} v_j \right\}.$$

On a alors $p \geq 1$ et

$$\frac{1}{h^2}(v_p - v_{p-1}) + c_p v_p + \frac{1}{h^2}(v_p - v_{p+1}) \geq 0.$$

On en déduit que

$$c_p v_p \geq \frac{1}{h^2}(v_{p-1} - v_p) + \frac{1}{h^2}(v_{p+1} - v_p) \geq 0.$$

Si $c_p > 0$, on a donc $v_p \geq 0$, et donc $v_i \geq 0, \quad \forall i = 1, \ldots, N$. Si $c_p = 0$, on doit alors avoir $v_{p-1} = v_p = v_{p+1}$ ce qui est impossible car p est le plus petit indice j tel que $v_j = \min_{i=1,\ldots,N} v_i$. Donc dans ce cas le minimum ne peut pas être atteint pour $j = p \geq 1$. On a ainsi finalement montré que $\min_{i\in\{1,\ldots,N\}} v_i \geq 0$, on a donc $v \geq 0$. ■

Définition 2.10 (Erreur de consistance) *On appelle erreur de consistance la quantité obtenue en remplaçant l'inconnue par la solution exacte dans le schéma numérique. Dans le cas du schéma* (2.10)*, l'erreur de consistance au point x_i est donc définie par :*

$$R_i = \frac{1}{h^2}(2u(x_i) - u(x_{i-1}) - u(x_{i+1})) + c(x_i)\,u(x_i) - f(x_i). \tag{2.14}$$

L'erreur de consistance R_i est donc l'erreur qu'on commet en remplaçant l'opérateur $-u''$ par le quotient différentiel

$$\frac{1}{h^2}(2u(x_i) - u(x_{i-1}) - u(x_{i+1})).$$

Cette erreur peut être évaluée si u est suffisamment régulière, en effectuant des développements de Taylor.

Définition 2.11 (Ordre du schéma) *On dit qu'un schéma de discrétisation à N points de discrétisation est d'ordre p s'il existe $C \in \mathbb{R}$, ne dépendant que de la solution exacte, tel que l'erreur de consistance satisfasse :*

$$\max_{i=1,\ldots,N} |R_i| < Ch^p,$$

où h est le le pas du maillage défini par (2.6) *(c.àd. le maximum des écarts $x_{i+1} - x_i$). On dit qu'un schéma de discrétisation est consistant si*

$$\max_{i=1,\ldots N} |R_i| \to 0 \text{ lorsque } h \to 0,$$

où N est le nombre de points de discrétisation.

Lemme 2.2 *Si la solution de* (2.9) *vérifie* $u \in C^4([0,1])$*, alors le schéma* (2.10) *est consistant d'ordre* 2*, et on a plus precisément :*

$$|R_i| \leq \frac{h^2}{12} \sup_{[0,1]} \left|u^{(4)}\right|, \forall i = 1, \ldots, N. \tag{2.15}$$

Démonstration:
Par développement de Taylor, on a :

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i),$$
$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\eta_i).$$

En additionnant ces deux égalités, on obtient que :

$$\frac{1}{h^2}\left(u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)\right) = u''(x_i) + \frac{h^2}{24}\left(u^{(4)}(\xi_i) + u^{(4)}(\eta_i)\right),$$

ce qui entraine que :

$$|R_i| \leq \frac{h^2}{12} \sup_{[0,1]} \left|u^{(4)}\right|. \tag{2.16}$$

■

Remarque 2.4 (Sur l'erreur de consistance)

1. *Si on note* $\bar{U}_h : (u(x_i))_{i=1\ldots N}$ *le vecteur dont les composantes sont les valeurs exactes de la solution de* (2.9)*, et* $U_h = (u_1 \ldots u_N)^t$ *la solution de* (2.10)*, on a :*

$$R = A_h\left(U_h - \bar{U}_h\right). \tag{2.17}$$

2. *On peut remarquer que si* $u^{(4)} = 0$*, les développements de Taylor effectués ci-dessus se résument à :*

$$-u''(x_i) = \frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1})}{h^2},$$

et on a donc $R_i = 0$*, pour tout* $i = 1, \ldots, N$*, et donc* $u_i = u(x_i)$*, pour tout* $i = 1 \ldots N$*. Dans ce cas (rare !), le schéma de discrétisation donne la valeur exacte de la solution en* x_i*, pour tout* $i = 1, \ldots, N$*. Cette remarque est bien utile lors de la phase de validation de méthodes numériques pour la résolution de l'équation* (2.9)*. En effet, si on choisit* f *telle que la solution soit un polynôme de degré inférieur ou égal à* 3*, alors on doit avoir une erreur entre la solution exacte et approchée inférieure à l'erreur machine.*

Nous introduisons maintenant une notion de stabilité qui sera utilisée avec la notion de consistance pour démontrer la convergence du schéma.

Proposition 2.3 *On dit que le schéma* (2.10) *est stable, au sens où la matrice de discrétisation* A_h *satisfait :*

$$\left\|A_h^{-1}\right\|_\infty \leq \frac{1}{8}. \tag{2.18}$$

On peut réécrire cette inégalité comme une estimation sur les solutions du système (2.11) *:*

$$\|U_h\| \leq \frac{1}{8}\|f\|_\infty. \tag{2.19}$$

Démonstration:
On rappelle que par définition, si $M \in \mathcal{M}_N(\mathbb{R})$,

$$\|M\|_\infty = \sup_{\substack{v\in\mathbb{R}^N \\ v\neq 0}} \frac{\|M.v\|_\infty}{\|v\|_\infty}, \text{ avec } \|v\|_\infty = \sup_{i=1,\ldots,N} |v_i|\,.$$

Pour montrer que $\left\|A_h^{-1}\right\|_\infty \leq \frac{1}{8}$, on décompose la matrice A_h sous la forme $A_h = A_{0h} + \mathrm{diag}\,(c_i)$ où A_{0h} est la matrice de discrétisation de l'opérateur $-u''$ avec conditions aux limites de Dirichlet homogènes, et

$$A_{0h} = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \ldots & & \ldots & 0 \\ -1 & 2 & -1 & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & -1 & 2 & -1 \\ 0 & \ldots & & \ldots & 0 & -1 & 2 \end{bmatrix}, \tag{2.20}$$

et $\mathrm{diag}\,(c_i)$ désigne la matrice diagonale de coefficients diagonaux c_i. Les matrices A_{0h} et A_h sont inversibles, et on a :

$$A_{0h}^{-1} - A_h^{-1} = A_{0h}^{-1}A_hA_h^{-1} - A_{0h}^{-1}A_{0h}A_h^{-1} = A_{0h}^{-1}\left(A_h - A_{0h}\right)A_h^{-1}.$$

Comme $\mathrm{diag}\,(c_i) \geq 0$, on a $A_h \geq A_{0h}$, et comme A_{0h} et A_h sont monotones, on en déduit que :

$$0 \leq A_h^{-1} \leq A_{0h}^{-1}, (\text{ composante par composante }).$$

On peut maintenant remarquer que si $B \in \mathcal{M}_N(\mathbb{R})$, et si $B \geq 0$ (c.à.d. $B_{ij} \geq 0$ pour tout i et j), on a

$$\|B\|_\infty = \sup_{\substack{v\in\mathbb{R}^N \\ \|v\|_\infty=1}} \sup_{i=1,\ldots,N} |(Bv)_i| = \sup_{\substack{v\in\mathbb{R}^N \\ \|v\|_\infty=1}} \sup_{i=1,\ldots,N} \left|\sum_{j=1}^N B_{ij}v_j\right| = \sup_{i=1,\ldots,N} \sum_{j=1}^N B_{ij}.$$

On a donc $\sup_{i=1,\dots,N} \sum_{j=1}^{N} \left(A_h^{-1}\right)_{ij} \leq \sup_{i=1,\dots,N} \sum_{j=1}^{N} \left(A_{0h}^{-1}\right)_{ij}$ car $A_h^{-1} \leq A_{0h}^{-1}$; d'où on déduit que $\left\|A_h^{-1}\right\|_\infty \leq \left\|A_{0h}^{-1}\right\|_\infty$. Il ne reste plus qu'à estimer $\left\|A_{0h}^{-1}\right\|_\infty$. Comme $A_{0h}^{-1} \geq 0$, on a

$$\left\|A_{0h}^{-1}\right\|_\infty = \left\|A_{0h}^{-1} e\right\|_\infty \text{ avec } e = (1, \dots, 1)^t.$$

Soit $d = A_{0h}^{-1} e \in \mathbb{R}^N$. On veut calculer $\|d\|_\infty$, où d vérifie $A_{0h} d = e$. Or le système linéaire $A_{0h} d = e$ n'est autre que la discrétisation par différences finies du problème

$$\begin{cases} -u'' = 1, \\ u(0) = u(1) = 0, \end{cases} \tag{2.21}$$

dont la solution exacte est :

$$u_{ex}(x) = \frac{x(1-x)}{2},$$

qui vérifie $u_{ex}^{(4)}(x) = 0$. On en conclut, par la remarque 2.4, que

$$u_{ex}(x_i) = d_i, \quad \forall i = 1 \dots N.$$

Donc $\|d\|_\infty = \sup_{i=1,N} \dfrac{ih(ih-1)}{2}$ où $h = \dfrac{1}{N+1}$ est le pas de discrétisation. Ceci entraîne que

$$\|d\|_\infty \leq \sup_{[0,1]} \left| \frac{x(x-1)}{2} \right| = \frac{1}{8}, \text{ et donc que } \left\|A_h^{-1}\right\|_\infty \leq \frac{1}{8}.$$

■

Définition 2.12 (Erreur de discrétisation) *On appelle erreur de discrétisation en x_i, la différence entre la solution exacte en x_i et la i-ème composante de la solution donnée par le schéma numérique*

$$e_i = u(x_i) - u_i, \quad \forall i = 1, \dots, N. \tag{2.22}$$

Théorème 2.1 (Convergence) *Soit u la solution exacte de*

$$\begin{cases} -u'' + cu = f, \\ u(0) = u(1) = 0. \end{cases}$$

On suppose $u \in C^4([0,1])$. Soit u_h la solution de (2.10)*. Alors l'erreur de discrétisation définie par* (2.22) *satisfait*

$$\max_{i=1,\dots,N} |e_i| \leq \frac{1}{96} \left\|u^{(4)}\right\|_\infty h^2.$$

Le schéma est donc convergent d'ordre 2.

Démonstration:
Soit $U_h = (U_1, \dots, U_n)^t$ et $\bar{U}_h = (u(x_1), \dots, u(x_N))^t$, on cherche à majorer $\|\bar{U}_h - U_h\|_\infty$. On a $A_h \left(U_h - \bar{U}_h\right) = R$ où R est l'erreur de consistance (voir remarque 2.4). On a donc

$$\left\|\bar{U}_h - U_h\right\|_\infty \leq \left\|A_h^{-1}\right\|_\infty \|R\|_\infty \leq \frac{1}{8} \times \frac{h^2}{12} \left\|u^{(4)}\right\|_\infty = \frac{h^2}{96} \left\|u^{(4)}\right\|_\infty.$$

■

Remarque 2.5 (Sur la convergence) *On peut remarquer que la preuve de la convergence s'appuie sur la stabilité (elle-même déduite de la conservation de la positivité) et sur la consistance. Dans certains livres d'analyse numérique, vous trouverez la "formule" : stabilité + consistance $\Longrightarrow$ convergence. Il faut toutefois prendre garde au fait que ces notions de stabilité et convergence peuvent être variables d'un type de méthode à un autre (comme la méthode des volumes finis, par exemple).*

Remarque 2.6 (Contrôle des erreurs d'arrondi) *On cherche à calculer la solution approchée de $-u'' = f$. Le second membre f est donc une donnée du problème. Supposons que des erreurs soient commises sur cette donnée (par exemple des erreurs d'arrondi, ou des erreurs de mesure). On obtient alors un nouveau système, qui s'écrit $A_h\tilde{U}_h = b_h+\varepsilon_h$, où ε_h représente la discrétisation des erreurs commises sur le second membre. Si on résout $A_h\tilde{U}_h = b_h + \varepsilon_h$ au lieu de $A_hU_h = b_h$, l'erreur commise sur la solution du système s'écrit*

$$E_h = \tilde{U}_h - U_h = A_h^{-1}\varepsilon_h.$$

On en déduit que

$$\|E_h\|_\infty \leq \frac{1}{8}\|\varepsilon_h\|_\infty .$$

On a donc une borne d'erreur sur l'erreur qu'on obtient sur la solution du système par rapport à l'erreur commise sur le second membre.

2.4.3 Etude de la méthode différences finies pour un problème non-stationnaire unidimensionnel

On prend comme modèle de problème non-stationnaire l'équation parabolique suivante : Au temps $t = 0$, on se donne une condition initiale u_0, et on considère des conditions aux limites de type Dirichlet homogène. Le problème unidimensionnel s'écrit :

$$\begin{cases} u_t - u_{xx} = 0, & \forall x \in]0,1[, \forall t \in]0,T[, \\ u(x,0) = u_0(x), & \forall x \in]0,1[, \\ u(0,t) = u(1,t) = 0, & \forall t \in]0,T[, \end{cases} \tag{2.23}$$

où $u(x,t)$ représente la température au point x et au temps t. On admettra le théorème d'existence et unicité suivant :

Théorème 2.2 (Résultat d'existence et unicité) *Si $u_0 \in C([0,1[,\mathbb{R})$ alors il existe une unique fonction $u \in C^2(]0,1[\times]0,T[,\mathbb{R}) \cap C([0,1]\times[0,T],\mathbb{R})$ qui vérifie* (2.23).

On a même $u \in C^\infty(]0,1[\times]0,T[,\mathbb{R})$. Ceci est appelé, effet "régularisant" de l'équation de la chaleur.

Proposition 2.4 (Principe du maximum) *Sous les hypothèses du théorème 2.2, soit u la solution du problème* (2.23) *;*

1. *si* $u_0(x) \geq 0$ *pour tout* $x \in [0,1]$, *alors* $u(x,t) \geq 0$, *pour tout* $t \geq 0$ *pour tout* $x \in]\,0,1[$.
2. $\|u\|_{L^\infty([0,1[\times]0,T[)} \leq \|u_0\|_{L^\infty(]0,1[)}$.

Ces dernières propriétés sont importantes dans le modèle physique, et il est souhaitable que les solutions approchées les vérifient également.
Pour calculer une solution approchée, on se donne une discrétisation en temps et en espace, qu'on notera $\mathcal{D}$. La discrétisation consiste donc à se donner un ensemble de points $t_n, n = 0, \ldots, M$ de l'intervalle $[0,T]$, et un ensemble de points $x_i, i = 0, \ldots, N+1$ de l'intervalle $[0,1]$. Pour simplifier, on considère un pas constant en temps et en espace. Soit : $h = \frac{1}{N+1} = \Delta x$ le pas de discrétisation en espace, et $k = \Delta t = \frac{T}{M}$, le pas de discrétisation en temps. On pose alors $t_n = nk$ pour $n = 0, \ldots, M$ et $x_i = ih$ pour $i = 0, \ldots, N+1$. On cherche à calculer une solution approchée $u_\mathcal{D}$ du problème (2.23) ; plus précisement, on cherche à déterminer $u_\mathcal{D}(x_i, t_n)$ pour $i = 1, \ldots, N$, et $n = 1, \ldots, M$. Les inconnues discrètes sont notées $u_i^{(n)}, i = 1, \ldots, N$ et $n = 1, \ldots, M$.

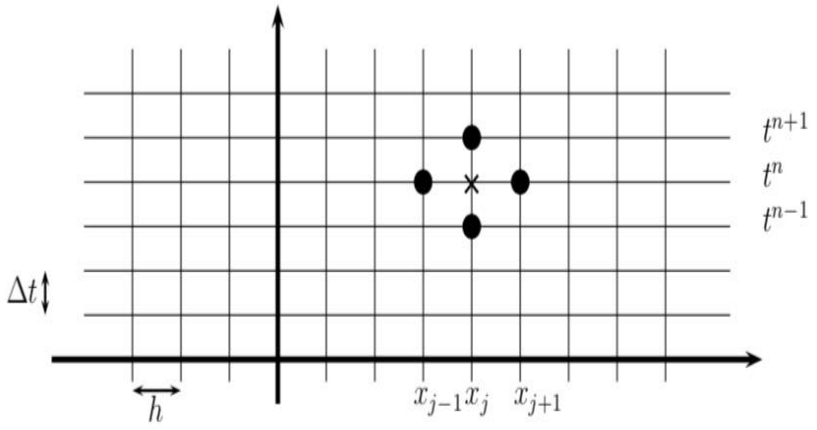


FIGURE 2.1 – Grille espace temps

2.4.3.1 Discrétisation par Euler explicite en temps

L'approximation en temps par la méthode d'Euler explicite consiste à écrire la première équation de (2.23) en chaque point x_i et temps t_n, à approcher $u_t(x_i, t_n)$ par le quotient différentiel :

$$\frac{u(x_i, t_{n+1}) - u(x_i, t_n)}{k},$$

et $-u_{xx}(x_i, t_n)$ par

$$\frac{1}{h^2}\left(2u(x_i, t_n) - u(x_{i-1}, t_n) - u(x_{i+1}, t_n)\right).$$

On obtient le schéma suivant :

$$\begin{cases} \dfrac{u_i^{(n+1)} - u_i^{(n)}}{k} + \dfrac{1}{h^2}\left(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}\right) = 0, \quad i = 1, \dots N, \quad n = 1, \dots, M-1, \\ u_i^0 = u_0\left(x_i\right), \quad i = 1, \dots, N, \\ u_0^{(n)} = u_{N+1}^{(n)} = 0, \quad \forall n = 1, \dots, M, \end{cases} \tag{2.24}$$

le schéma est dit explicite, car la formule ci-dessus donne $u_i^{(n+1)}$ de manière explicite en fonction des $\left(u_i^{(n)}\right)_{i=1,\dots,N}$. En effet on a :

$$u_i^{(n+1)} = u_i^{(n)} - \lambda\left(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}\right),$$

avec $\lambda = \dfrac{k}{h^2}$.

- Consistance du schéma

Soit $\bar{u}_i^{(n)} = u\left(x_i, t_n\right)$ la valeur exacte de la solution en x_i et t_n : L'erreur de consistance R_i en (x_i, t_n) peut s'écrire comme la somme des erreurs de consistance en temps et en espace : $R_i^{(n)} = \widetilde{R}_i^{(n)} + \widehat{R}_i^{(n)}$ avec :

$$\begin{cases} \widetilde{R}_i^{(n)} = \dfrac{\bar{u}_i^{(n+1)} - \bar{u}_i^{(n)}}{k} - u_t\left(x_i, t_n\right) \text{ et} \\ \widehat{R}_i^{(n)} = \dfrac{1}{h^2}\left(2\bar{u}_i^{(n)} - \bar{u}_{i-1}^{(n)} - \bar{u}_{i+1}^{(n)}\right) + u_{xx}\left(x_i, t_n\right). \end{cases}$$

Donc

$$R_i^{(n)} = \frac{\bar{u}_i^{(n+1)} - \bar{u}_i^{(n)}}{k} + \frac{1}{h^2}\left(2\bar{u}_i^{(n)} - \bar{u}_{i-1}^{(n)} - \bar{u}_{i+1}^{(n)}\right). \tag{2.25}$$

Proposition 2.5 *Le schéma* (2.24) *est consistant d'ordre* 1 *en temps et d'ordre* 2 *en espace, c'est à dire qu'il existe* $C \in \mathbb{R}_+$ *ne dépendant que de* u *tel que :*

$$\left|R_i^{(n)}\right| \leq C\left(k + h^2\right). \tag{2.26}$$

Démonstration:

On a vu lors de l'étude des problèmes elliptiques que l'erreur de consistance en espace $\widehat{R}_i^{(n)}$ est d'ordre 2. Dans le chapitre 1 on a vu que $\widetilde{R}_i^{(n)}$ est d'ordre 1 en temps. ■

- Stabilité

On a vu à la proposition 2.4 que la solution exacte vérifie :

$$\|u\|_{L^\infty(]0,1[\times]0,T[)} \leq \|u_0\|_{L^\infty(]0,1[)}\,.$$

Si on choisit correctement les pas de temps et d'espcace, nous allons voir qu'on peut avoir l'équivalent discret sur la solution approchée.

Définition 2.13 *On dit qu'un schéma est L^∞-stable si la solution approchée est bornée dans L^∞ indépendamment du pas du maillage.*

Proposition 2.6 *Si la condition de stabilité*

$$\lambda = \frac{k}{h^2} \leq \frac{1}{2}, \tag{2.27}$$

est vérifiée, alors le schéma (2.24) *est L^∞-stable au sens où :*

$$\sup_{\substack{i=1,\dots,N\\ n=1,\dots,M}} \left|u_i^{(n)}\right| \leq \|u_0\|_\infty\,.$$

***Démonstration**:*

On peut écrire le schéma sous la forme :

$$u_i^{(n+1)} = u_i^{(n)} - \lambda\left(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}\right),$$

soit encore :

$$u_i^{(n+1)} = (1-2\lambda)u_i^{(n)} + \lambda u_{i-1}^{(n)} + \lambda u_{i+1}^{(n)}.$$

Si $0 \leq \lambda \leq \frac{1}{2}$, on a $\lambda \geq 0$ et $1-2\lambda \geq 0$, et la quantité $u_i^{(n+1)}$ est donc combinaison convexe de $u_i^{(n)}, u_{i-1}^{(n)}$ et $u_{i+1}^{(n)}$. Soit $M^{(n)} = \max_{i=1,\dots,N} u_i^{(n)}$, on a alors :

$$u_i^{(n+1)} \leq (1-2\lambda)M^{(n)} + \lambda M^{(n)} + \lambda M^{(n)}, \quad \forall i = 1,\dots,N$$

et donc $u_i^{(n+1)} \leq M^{(n)}$. On en déduit en passant au maximum que :

$$M^{(n+1)} \leq M^{(n)}.$$

On montre de la même manière que

$$\min_{i=1,\dots,N} u_i^{(n+1)} \geq \min_{i=1,\dots,N} u_i^{(n)}.$$

On en déduit $\max_{i=1,\dots,N}\left(u_i^{(n+1)}\right) \leq \max u_i^0$ et $\min_{i=1,\dots,N}\left(u_i^{(n+1)}\right) \geq \min u_i^0$ d'où le résultat. ■

- Convergence

Définition 2.14 *Soit u la solution du problème (2.23) et $\left(u_i^{(n)}\right)_{i=1,\dots N, n=1,\dots,M}$ la solution de (2.24). On appelle erreur de discrétisation au point (x_i, t_n) la quantité $e_i^{(n)} = u(x_i, t_n) - u_i^{(n)}$ et on note $e^{(n)} = (e_i^{(n)})_{i=1,\dots,N}$.*

Théorème 2.3 *Sous les hypothèses du théorème 2.2, et sous la condition de stabilité 2.27, il existe, $C \in \mathbb{R}_+$ ne dépendant que de u tel que*

$$\left\|e^{(n+1)}\right\|_\infty \leq \left\|e^{(0)}\right\|_\infty + TC\left(k+h^2\right) \text{ pour tout } i = 1,\dots,N \text{ et } n = 0,\dots,M-1.$$

Ainsi, si $\left\|e^{(0)}\right\|_\infty = 0$, alors $\max_{i=1,\dots N}\left\|e_i^{(n)}\right\|$ tend vers 0 lorsque k et h tendent vers 0, pour tout $n = 1, \dots, M$. Le schéma (2.24) est donc convergent.

Démonstration:
On note $\bar{u}_i^{(n)} = u(x_i, t_n)$. On a donc, par définition de l'erreur de consistance (voir (2.25)),

$$\frac{\bar{u}_i^{(n+1)} - \bar{u}_i^{(n)}}{k} + \frac{1}{h^2}\left(2\bar{u}_i^{(n)} - \bar{u}_{i-1}^{(n)} - \bar{u}_{i+1}^{(n)}\right) = R_i^{(n)}. \tag{2.28}$$

D'autre part, le schéma numérique s'écrit :

$$\frac{u_i^{(n+1)} - u_i^{(n)}}{k} + \frac{1}{h^2}\left(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}\right) = 0. \tag{2.29}$$

(2.28)−(2.29) $\Longrightarrow$

$$\frac{e_i^{(n+1)} - e_i^{(n)}}{k} + \frac{1}{h^2}\left(2e_i^{(n)} - e_{i+1}^{(n)} - e_{i-1}^{(n)}\right) = R_i^{(n)},$$

soit encore :

$$e_i^{(n+1)} = (1-2\lambda)e_i^{(n)} + \lambda e_{i-1}^{(n)} + \lambda e_{i+1}^{(n)} + kR_i^{(n)}.$$

Or $(1-2\lambda)e_i^{(n)} + \lambda e_{i-1}^{(n)} + \lambda e_{i+1}^{(n)} \leq \left\|e^{(n)}\right\|_\infty$, car $\lambda \leq \frac{1}{2}$, et donc comme le schéma est consistant, l'inégalité (2.26) entraine que :

$$\left|e_i^{(n+1)}\right| \leq \left\|e^{(n)}\right\|_\infty + kC\left(k+h^2\right).$$

On a donc par récurrence :

$$\left\|e^{(n+1)}\right\|_\infty \leq \left\|e^{(0)}\right\|_\infty + MkC\left(k+h^2\right),$$

ce qui démontre le théorème. ■

2.4.3.2 Discrétisation par Euler implicite en temps

L'approximation en temps par la méthode d'Euler implicite consiste, à écrire la première équation de (2.23) en chaque point x_i et temps t_{n+1}, à approcher $u_t(x_i,t_n)$ par le quotient différentiel :

$$\frac{u(x_i,t_{n+1}) - u(x_i,t_n)}{k},$$

et $-u_{xx}(x_i,t_{n+1})$ par

$$\frac{1}{h^2}\left(2u(x_i,t_{n+1}) - u(x_{i-1},t_{n+1}) - u(x_{i+1},t_{n+1})\right).$$

On obtient le schéma suivant pour $i = 1,\dots N$ et $n = 1,\dots,M-1$:

$$\begin{cases} \dfrac{u_i^{(n+1)} - u_i^{(n)}}{k} + \dfrac{1}{h^2}\left(2u_i^{(n+1)} - u_{i-1}^{(n+1)} - u_{i+1}^{(n+1)}\right) = 0, \\ u_i^0 = u_0(x_i), \\ u_0^{(n)} = u_{N+1}^{(n)} = 0. \end{cases} \tag{2.30}$$

On peut écrire le schéma sous la forme :

$$(1+2\lambda)u_i^{(n+1)} - \lambda u_{i-1}^{(n+1)} - \lambda u_{i+1}^{(n+1)} = u_i^{(n)}. \tag{2.31}$$

On va montrer de plus qu'il est $L^\infty-$stable.

Proposition 2.7 (Stabilité L^∞ pour Euler implicite)
Si $\left(u_i^{(n)}\right)_{i=1,\dots,N}$ *est solution du schéma* (2.31)*, alors :*

$$\max_{i=1,\dots,N} u_i^{(n+1)} \leq \max_{i=1,\dots,N} u_i^{(n)} \leq \max_{i=1,\dots,N} u_i^{(0)}, \tag{2.32}$$

de même :

$$\min_{i=1,\dots,N} u_i^{(n+1)} \geq \min_{i=1,\dots,N} u_i^{(n)} \geq \min_{i=1,\dots,N} u_i^{(0)}. \tag{2.33}$$

Le schéma (2.31) *est donc* $L^\infty-$*stable.*

Démonstration:
Prouvons l'estimation (2.32), la preuve de (2.33) est similaire. Soit i_0 tel que $u_{i_0}^{(n+1)} = \max_{i=1,\dots,N} u_i^{(n+1)}$. Par définition du schéma d'Euler implicite (2.31), On a :

$$u_{i_0}^{(n)} = (1+2\lambda)u_{i_0}^{(n+1)} - \lambda u_{i_0-1}^{(n+1)} - \lambda u_{i_0+1}^{(n+1)}.$$

On en déduit : $u_{i_0}^{(n+1)} \leq \max_{i=1,\dots,N} u_i^{(n)}$, ce qui prouve que

$$\max_{i=1,\dots,N} u_i^{(n+1)} \leq \max_{i=1,\dots,N} u_i^{(n)}.$$

Donc le schéma (2.31)est L^∞ stable. ■

Théorème 2.4 *Soit* $e^{(n)} = (e_j^n)_{j=1,\dots,N}$ *l'erreur de discrétisation, définie par*

$$e_j^{(n)} = u\left(x_j, t_n\right) - u_j^{(n)} \text{ pour } j = 1, \dots, N.$$

Alors $\left\|e^{(n+1)}\right\|_\infty \leq \left\|e^{(0)}\right\|_\infty + TC\left(k+h^2\right)$. *Si* $\left\|e^{(0)}\right\|_\infty = 0$, *le schéma est donc convergent (d'ordre* 1 *en temps et* 2 *en espace).*

Démonstration:
En utilisant la définition de l'erreur de consistance, on obtient :

$$(1+2\lambda)e_i^{(n+1)} - \lambda e_{i-1}^{(n+1)} - \lambda e_{i+1}^{(n+1)} = e_i^{(n)} + kR_i^{(n+1)},$$

et donc :

$$\left\|e^{(n+1)}\right\|_\infty \leq \left\|e^n\right\|_\infty + kC\left(k+h^2\right).$$

On en déduit, par récurrence sur n, que :

$$\left\|e^{(n+1)}\right\|_\infty \leq \left\|e^0\right\|_\infty + TC\left(k+h^2\right),$$

d'où la convergence du schéma. ■

2.4.3.3 Discrétisation par θ-schéma

En faisant une combinaison convexe de (2.24) et (2.30), on obtient le **θ-schéma** ($0 \leq \theta \leq 1$). Pour $i = 1, \ldots N$ et $n = 1, \ldots, M-1$, on a :

$$\begin{cases} \dfrac{u_i^{(n+1)} - u_i^{(n)}}{k} = \dfrac{\theta}{h^2}\left(-2u_i^{(n+1)} + u_{i-1}^{(n+1)} + u_{i+1}^{(n+1)}\right) \\ \qquad\qquad + \dfrac{1-\theta}{h^2}\left(-2u_i^{(n)} + u_{i-1}^{(n)} + u_{i+1}^{(n)}\right), \\ u_i^0 = u_0\left(x_i\right), \\ u_0^{(n)} = u_{N+1}^{(n)} = 0. \end{cases} \tag{2.34}$$

- Si $\theta = 0$ on retrouve le schéma d'**Euler explicite**.
- Si $\theta \neq 0$, le θ-schéma est implicite.
- Si $\theta = 1$ on retrouve le schéma d'**Euler implicite**.
- Si $\theta = \dfrac{1}{2}$ on retrouve le schéma de **Crank-Nicholson**.

.

Proposition 2.8 (Consistance du θ-schéma) *Le θ-schéma* (2.34) *pour la discrétisation du problème* (2.23) *est d'ordre* 2 *en espace. Il est d'ordre* 2 *en temps si* $\theta = \dfrac{1}{2}$, *et d'ordre* 1 *sinon.*

Chapitre 3

Calcul numérique des vecteurs et valeurs propres

3.1 Introduction

On distingue deux grandes classes de problèmes en algèbre linéaire :

1. Résoudre un système d'équations linéaire $Ax = b$.

2. Résoudre un problème aux valeurs propres $Ax = \lambda x$ et c'est l'objet de ce chapitre. La détermination des éléments propres d'une matrice à de multiples applications, on rencontre ce type de problème en traitement d'image, la résolution de certaines équations aux dérivées partielles ou des équations différentielles ordinaire, en mécanique, etc.

Il existe deux grandes familles de méthodes numériques d'approximation des valeurs et vecteurs propres.

- Méthodes partielles : Approcher des valeurs propres optimales telles que la plus grande en module, la plus petite ou la plus proches d'une valeur donnée. Un exemple de ces méthodes est la méthode de la puissance et la méthode de la puissance inverse.

- Méthodes globales : Approcher tout le spectre en utilisant des transformations de la matrice. Un exemple de ces méthode et la méthode QR pour une matrice quelconque ou Jacobi pour une matrice symétrique.

3.2 Rappels mathématiques

Définition 3.1 *Une matrice carrée A d'ordre n est dite inversible (ou régulière ou non singulière) s'il existe une matrice carrée B d'ordre n telle que* $AB = BA = I$. *On dit que B est la matrice inverse de A et on la note* A^{-1}. *Une matrice qui n'est pas inversible est dite singulière.*

Si A est inversible, son inverse est aussi inversible et $\left(\mathrm{A}^{-1}\right)^{-1} = \mathrm{A}$. De plus, si A et B sont deux matrices inversibles d'ordre n, leur produit AB est aussi inversible et $(\mathrm{AB})^{-1} = \mathrm{B}^{-1}\ \mathrm{A}^{-1}$. On a la propriété suivante :

Propriété 3.1 *Une matrice carrée est inversible si et seulement si ses vecteurs colonnes sont linéairement indépendants.*

Définition 3.2 *On appelle transposée d'une matrice* $\mathrm{A} \in \mathbb{R}^{m\times n}$ *la matrice* $n \times m$, *notée* A^T, *obtenue en échangeant les lignes et les colonnes de* A.

On a clairement, $\left(\mathrm{A}^T\right)^T = \mathrm{A}, (\mathrm{A}+\mathrm{B})^T = \mathrm{A}^T + \mathrm{B}^T, (\mathrm{AB})^T = \mathrm{B}^T\ \mathrm{A}^T$ et $(\alpha\mathrm{A})^T = \alpha\mathrm{A}^T \forall \alpha \in \mathbb{R}$. Si A est inversible, on a aussi $\left(\mathrm{A}^T\right)^{-1} = \left(\mathrm{A}^{-1}\right)^T = \mathrm{A}^{-T}$.

Définition 3.3 *Soit* $\mathrm{A} \in \mathbb{C}^{m\times n}$ *; la matrice* $\mathrm{B} = \mathrm{A}^* \in \mathbb{C}^{n\times m}$ *est appelée adjointe (ou transposée conjuguée) de A si* $b_{ij} = \bar{a}_{ji}$, *où* $\bar{a}_{ji}$ *est le complexe conjugué de* a_{ji}.

On a $(\mathrm{A}+\mathrm{B})^* = \mathrm{A}^* + \mathrm{B}^*, (\mathrm{AB})^* = \mathrm{B}^*\ \mathrm{A}^*$ et $(\alpha\mathrm{A})^* = \bar{\alpha}\mathrm{A}^* \forall \alpha \in \mathbb{C}$.

Définition 3.4 *Une matrice* $\mathrm{A} \in \mathbb{R}^{n\times n}$ *est dite symétrique si* $\mathrm{A} = \mathrm{A}^T$, *et antisymétrique si* $\mathrm{A} = -\mathrm{A}^T$. *Elle est dite orthogonale si* $\mathrm{A}^T\ \mathrm{A} = \mathrm{A}\mathrm{A}^T = \mathrm{I}$, *c'est-à-dire si* $\mathrm{A}^{-1} = \mathrm{A}^T$.

Les matrices de permutation sont orthogonales et le produit de matrices orthogonales est orthogonale.

Définition 3.5 *Une matrice* $\mathrm{A} \in \mathbb{C}^{n\times n}$ *est dite hermitienne ou autoadjointe si* $\mathrm{A}^T = \overline{\mathrm{A}}$, *c'est-à-dire si* $\mathrm{A}^* = \mathrm{A}$, *et elle est dite unitaire si* $\mathrm{A}^*\ \mathrm{A} = \mathrm{A}\mathrm{A}^* = \mathrm{I}$. *Enfin, si* $\mathrm{A}\mathrm{A}^* = \mathrm{A}^*\ \mathrm{A}$, A *est dite normale.*

Par conséquent, une matrice unitaire est telle que $\mathrm{A}^{-1} = \mathrm{A}^*$. Naturellement, une matrice unitaire est également normale, mais elle n'est en général pas hermitienne.
On notera enfin que les coefficients diagonaux d'une matrice hermitienne sont nécessairement réels.

3.3 Problème aux valeurs propres

De façon générale, la résolution d'un problème aux valeurs propres consiste à trouver λ et $x = (x_1\, x_2\, \cdots\, x_n)^T$ solution de

$$Ax = \lambda x. \tag{3.1}$$

Le vecteur x est le vecteur propre associé à la valeur propre λ et l'ensemble des valeurs propres de A est appelé spectre de A, on le note $\sigma(A)$.
Nous supposons que la matrice A est soit symétrique (ou hermitienne), soit diagonalisable, ou soit toutes ses valeurs propres sont différents.
Puisque $X = 0$ est toujours solution de $(A - \lambda I)X = 0$ et que l'on cherche des solutions $X \neq 0$, on se place dans le cas où $M = A - \lambda I$ est singulière. Autrement dit on cherche les valeurs propres λ de A en résolvant l'équation :

$$det(A - \lambda I) = 0, \tag{3.2}$$

Théoriquement, La résolution se fait en 3 étapes

- Calcul du déterminant de $A - \lambda I$.
- Détermination des racines du polynôme caractéristique obtenu en écrivant : $det(A - \lambda I) = 0$.
- Pour chaque racine (chaque valeur propre), résoudre le système linéaire $AX = \lambda X$ afin de déterminer le ou les vecteurs propres associés.

Mais si la taille de la matrice est grande (1000×1000 par exemple), le calcul d'un déterminants symbolique 1000×1000, et les racines d'un polynôme de degré 1000 deviennent très rapidement monstrueux. Donc c'est impossible d'appliquer la définition dans la pratique. C'est pour cette raison qu'on fait appel à des méthodes itératives.

Quelques propriétés :

- $B = P^{-1}AP$ et A ont les mêmes valeurs propres et si X est le vecteur propre de A associé à λ, le vecteur propre de B associé à la mêmes valeur propre λ est $Z = P^{-1}X$.
- Si A est symétrique, ses valeurs propres sont réelles.
- Si A est symétrique, les vecteurs propres de A forment une base de $\mathbb{R}^n$.
- Si A est diagonalisable, on a $det(A) = \prod_i \lambda_i$ et $tr(A) = \sum_i \lambda_i$.

Localisation des valeurs propres

Théorème 3.1 *Soit A une matrice carrée d'ordre n, alors les valeurs propres de A appartiennent à la réunion des n disques $R_{i=1\ldots n}$ (appelés **disques de Gershgorin**) définis par $R_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j,j\neq i} |a_{ij}|\}$.*

Preuve : Soit λ une valeur propre de A associée au vecteur propre v, c'est à dire $Av = \lambda v$. Soit i tel que $|v_i| \geq |v_j| \; \forall j$, alors $v_i \neq 0$.

$$Av = \lambda v \iff \sum_{j=1}^{n} a_{ij}v_j = \lambda v_i \iff \sum_{j=1,j\neq i}^{n} a_{ij}v_j = (\lambda - a_{ii})v_i,$$

donc

$$|v_i|\,|\lambda - a_{ii}| \leq \sum_{j=1,j\neq i}^{n} |a_{ij}||v_j|.$$

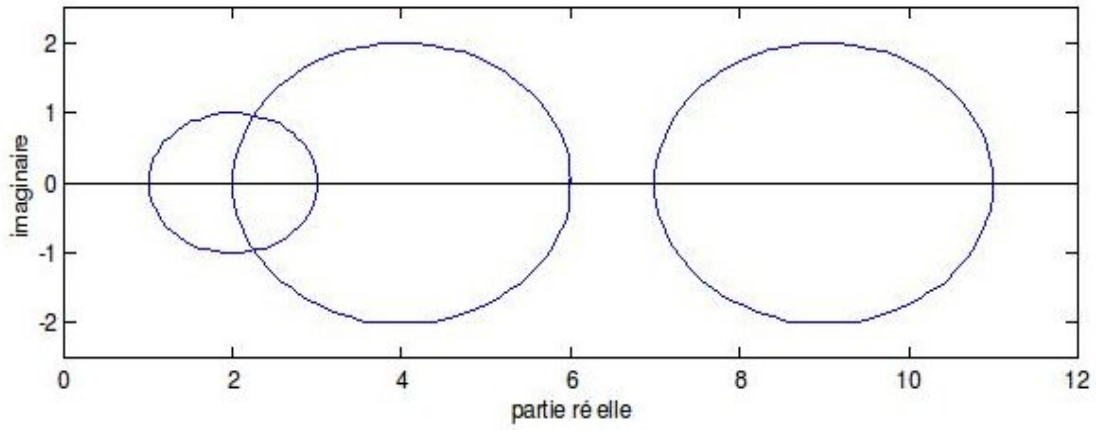
D'ou

$$|\lambda - a_{ii}| \leq \sum_{j=1,j\neq i}^{n} |a_{ij}|.$$

Exemple 3.1

$$A = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ 2 & 0 & 9 \end{pmatrix}.$$

Les valeurs propres de A sont $\lambda_1 = 2.1943972$, $\lambda_2 = 3.3867702$ et $\lambda_3 = 9.4188327$.



Les matrices A et A^T ayant le même spectre, le Théoorème 3.1 s'écrit aussi sous la forme :

Remarque 3.1 *Les valeurs propres de A appartiennent aussi à la réunion des n disques définis par $R_i = \{z \in \mathbb{C} : |z - a_{jj}| \leq \sum_{i,i\neq j} |a_{ij}|\}$.*

3.4 Méthodes partielles de recherche de valeurs propres

3.4.1 Méthode de la puissance

Soit A $\in \mathbb{C}^{n\times n}$ une matrice diagonalisable et soit X $\in \mathbb{C}^{n\times n}$ la matrice de ses vecteurs propres $\mathbf{x}_i$, pour $i = 1, \ldots, n$. Supposons les valeurs propres de A ordonnées de la façon suivante :

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \ldots \geq |\lambda_n|\,, \tag{3.3}$$

et supposons que λ_1 ait une multiplicité algébrique égale à 1. Sous ces hypothèses, λ_1 est appelée valeur propre dominante de la matrice A.

On veut construire une méthode pour approcher numériquement λ_1. Pour cela, on se donne un vecteur initial arbitraire $\mathbf{q}^{(0)} \in \mathbb{C}^n$ de norme euclidienne égale à 1, considérons pour $k = 1, 2, \ldots$ la méthode itérative suivante,

connue sous le nom de **méthode de la puissance** :

$$\begin{aligned} \mathbf{z}^{(k)} &= \mathrm{A}\mathbf{q}^{(k-1)}, \\ \mathbf{q}^{(k)} &= \mathbf{z}^{(k)} / \left\|\mathbf{z}^{(k)}\right\|_2, \\ \nu^{(k)} &= \left(\mathbf{q}^{(k)}\right)^* \mathrm{A}\mathbf{q}^{(k)}. \end{aligned} \tag{3.4}$$

Analysons les propriétés de convergence de la méthode (3.4). Par récurrence sur k, on peut vérifier que

$$\mathbf{q}^{(k)} = \frac{\mathrm{A}^k \mathbf{q}^{(0)}}{\left\|\mathrm{A}^k \mathbf{q}^{(0)}\right\|_2}, \quad k \geq 1. \tag{3.5}$$

Cette relation rend explicite le rôle joué par les puissances de A. Ayant supposé la matrice A diagonalisable, ses vecteurs propres $\mathbf{x}_i$ forment une base de $\mathbb{C}^n$ sur laquelle on peut décomposer $\mathbf{q}^{(0)}$:

$$\mathbf{q}^{(0)} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i, \quad \alpha_i \in \mathbb{C}, \quad i = 1, \ldots, n. \tag{3.6}$$

De plus, comme $\mathrm{A}\mathrm{x}_i = \lambda_i \mathbf{x}_i$, on a (si $\alpha_1 \neq 0$)

$$\mathrm{A}^k \mathbf{q}^{(0)} = \alpha_1 \lambda_1^k \left(\mathbf{x}_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1}\right)^k \mathbf{x}_i \right), k = 1, 2, \ldots \tag{3.7}$$

Quand k augmente, comme $|\lambda_i/\lambda_1| < 1$ pour $i = 2, \ldots, n$, la composante le long de $\mathbf{x}_1$ du vecteur $\mathrm{A}^k \mathbf{q}^{(0)}$ (et donc aussi celle de $\mathbf{q}^{(k)}$ d'après (3.5)) augmente, tandis que ses composantes suivant les autres directions $\mathbf{x}_j$ diminuent.
En utilisant (3.5) et (3.7), on obtient

$$\mathbf{q}^{(k)} = \frac{\alpha_1 \lambda_1^k \left(\mathbf{x}_1 + \mathbf{y}^{(k)}\right)}{\left\|\alpha_1 \lambda_1^k \left(\mathbf{x}_1 + \mathbf{y}^{(k)}\right)\right\|_2} = \mu_k \frac{\mathbf{x}_1 + \mathbf{y}^{(k)}}{\left\|\mathbf{x}_1 + \mathbf{y}^{(k)}\right\|_2},$$

où $\mu_k = \dfrac{\alpha_1 \lambda_1^k}{|\alpha_1 \lambda_1^k saoudivsar|}$ et $\mathbf{y}^{(k)}$ désigne un vecteur qui tend vers zéro quand $k \to \infty$.

Le vecteur $\mathbf{q}^{(k)}$ s'aligne donc le long de la direction du vecteur propre $\mathbf{x}_1$ quand $k \to \infty$. On a de plus l'estimation d'erreur suivante à l'étape k :

Théorème 3.2 *Soit* $\mathrm{A} \in \mathbb{C}^{n\times n}$ *une matrice diagonalisable dont les valeurs propres satisfont* (3.3). *En supposant* $\alpha_1 \neq 0$, *il existe une constante* $C > 0$ *telle que*

$$\left\| \tilde{\mathbf{q}}^{(k)} - \mathbf{x}_1 \right\|_2 \leq C \left| \frac{\lambda_2}{\lambda_1} \right|^k, \quad k \geq 1, \tag{3.8}$$

où

$$\tilde{\mathbf{q}}^{(k)} = \frac{\mathbf{q}^{(k)} \left\| \mathrm{A}^k \mathbf{q}^{(0)} \right\|_2}{\alpha_1 \lambda_1^k} = \mathbf{x}_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{x}_i, \quad k = 1, 2, \ldots \tag{3.9}$$

Démonstration:
On peut, sans perte de généralité, supposer que les colonnes de la matrice X ont une norme euclidienne égale à 1, c'est-à-dire $\|\mathbf{x}_i\|_2 = 1$ pour $i = 1, \ldots, n$. D'après (3.7), on a alors

$$\begin{aligned} \|\mathbf{x}_1 + \sum_{i=2}^{n} \left[\frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{x}_i \right] - \mathbf{x}_1 \|_2 &= \| \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{x}_i \|_2 \\ &\leq \left(\sum_{i=2}^{n} \left[\frac{\alpha_i}{\alpha_1} \right]^2 \left[\frac{\lambda_i}{\lambda_1} \right]^{2k} \right)^{1/2} \leq \left| \frac{\lambda_2}{\lambda_1} \right|^k \left(\sum_{i=2}^{n} \left[\frac{\alpha_i}{\alpha_1} \right]^2 \right)^{1/2}, \end{aligned}$$

c'est-à-dire (3.8) avec $C = \left(\sum_{i=2}^{n} \left(\alpha_i / \alpha_1 \right)^2 \right)^{1/2}$.

■

L'estimation (3.8) exprime la convergence de la suite $\tilde{\mathbf{q}}^{(k)}$ vers $\mathbf{x}_1$. La suite des quotients de **Rayleigh**

$$\left(\left(\tilde{\mathbf{q}}^{(k)} \right)^* \mathrm{A} \tilde{\mathbf{q}}^{(k)} \right) / \left\| \tilde{\mathbf{q}}^{(k)} \right\|_2^2 = \left(\mathbf{q}^{(k)} \right)^* \mathrm{A} \mathbf{q}^{(k)} = \nu^{(k)},$$

converge donc vers λ_1. Par conséquent, $\lim_{k\to\infty} \nu^{(k)} = \lambda_1$, et la convergence est d'autant plus rapide que le quotient $|\lambda_2/\lambda_1|$ est petit.

Si la matrice A est réelle et symétrique avec $\alpha_1 \neq 0$, on a le résultat suivant :

$$\left|\lambda_1 - \nu^{(k)}\right| \leq |\lambda_1 - \lambda_n| \tan^2(\theta_0) \left|\frac{\lambda_2}{\lambda_1}\right|^{2k}, \tag{3.10}$$

où $\cos(\theta_0) = \left|\mathbf{x}_1^T \mathbf{q}^{(0)}\right| \neq 0$. L'inégalité (3.10) montre que la convergence de la suite $\nu^{(k)}$ vers λ_1 est quadratique par rapport à $|\lambda_2/\lambda_1|$.
L'analyse de convergence montre que l'efficacité de la méthode de la puissance dépend fortement des valeurs propres dominantes. Plus précisément, la méthode est d'autant plus efficace que les valeurs dominantes sont bien séparées, i.e. $|\lambda_2| / |\lambda_1| \ll 1$. Analysons à présent le comportement des itérations (3.4) quand il y a deux valeurs propres dominantes de même module (c'est-à-dire quand $|\lambda_2| = |\lambda_1|$). On doit distinguer trois cas :

1. $\lambda_2 = \lambda_1$: les deux valeurs propres dominantes coincident. La méthode est encore convergente, puisque pour k assez grand, (3.7) implique

   $$\mathrm{A}^k \mathbf{q}^{(0)} \simeq \lambda_1^k (\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2)$$

   qui est un vecteur propre de A. Pour $k \to \infty$, la suite $\tilde{\mathbf{q}}^{(k)}$ (convenablement redéfinie) converge vers un vecteur appartenant à l'espace engendré par $\mathbf{x}_1$ et $\mathbf{x}_2$. La suite $\nu^{(k)}$ converge encore vers λ_1.

2. $\lambda_2 = -\lambda_1$: les deux valeurs propres dominantes sont opposées. Dans ce cas, la valeur propre de plus grand module peut être approchée en appliquant la méthode de la puissance à la matrice A^2. En effet, pour $i = 1, \ldots, n$, $\lambda_i\left(\mathrm{A}^2\right) = [\lambda_i(\mathrm{A})]^2$, donc $\lambda_1^2 = \lambda_2^2$ et on est ramené au cas précédent avec la matrice A^2.

3. $\lambda_2 = \bar{\lambda}_1$: les deux valeurs propres dominantes sont complexes conjuguées. Cette fois, des oscillations non amorties se produisent dans la suite $\mathbf{q}^{(k)}$ et la méthode de la puissance ne converge pas.

3.4.2 Méthode de la puissance inverse

Dans cette section, nous recherchons une approximation de la valeur propre d'une matrice $\mathrm{A} \in \mathbb{C}^{n \times n}$ la plus proche d'un nombre $\mu \in \mathbb{C}$ donné, avec $\mu \notin \sigma(\mathrm{A})$. On peut pour cela appliquer la méthode de la puissance (3.4)

à la matrice $(\mathrm{M}_\mu)^{-1} = (\mathrm{A} - \mu\mathrm{I})^{-1}$, ce qui conduit à **la méthode des itérations inverses ou méthode de la puissance inverse**. Le nombre μ est appelé shift en anglais.
Les valeurs propres de M_μ^{-1} sont $\xi_i = (\lambda_i - \mu)^{-1}$; supposons qu'il existe un entier m tel que :

$$|\lambda_m - \mu| < |\lambda_i - \mu|, \quad \forall i = 1, \ldots, n \quad \text{et } i \neq m. \tag{3.11}$$

Cela revient à supposer que la valeur propre λ_m qui est la plus proche de μ a une multiplicité égale à 1 . De plus, (3.11) montre que ξ_m est la valeur propre de M_μ^{-1} de plus grand module ; en particulier, si $\mu = 0, \lambda_m$ est la valeur propre de A de plus petit module.
Etant donné un vecteur initial arbitraire $\mathbf{q}^{(0)} \in \mathbb{C}^n$ de norme euclidienne égale à 1, on construit pour $k = 1, 2, \ldots$ la suite définie par :

$$\begin{aligned} &(\mathrm{A} - \mu\mathrm{I})\mathbf{z}^{(k)} = \mathbf{q}^{(k-1)}, \\ &\mathbf{q}^{(k)} = \mathbf{z}^{(k)} / \left\|\mathbf{z}^{(k)}\right\|_2, \\ &\sigma^{(k)} = \left(\mathbf{q}^{(k)}\right)^* \mathrm{A}\mathbf{q}^{(k)}. \end{aligned} \tag{3.12}$$

Remarquer que les vecteurs propres de M_μ sont les mêmes que ceux de A puisque $\mathrm{M}_\mu = \mathrm{X}(\Lambda - \mu\mathrm{I}_n)\mathrm{X}^{-1}$, où $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Pour cette raison, on calcule directement le quotient de Rayleigh dans (3.12) à partir de la matrice A (et non à partir de M_μ^{-1}). La différence principale par rapport à (3.4) est qu'on doit résoudre à chaque itération k un système linéaire de matrice $\mathrm{M}_\mu = \mathrm{A} - \mu\mathrm{I}$.
Donc la méthode de la puissance inverse est plus coûteuse que la méthode de la puissance (3.4), mais elle a l'avantage de pouvoir converger vers n'importe quelle valeur propre de A (la plus proche de μ). Les itérations inverses se prêtent donc bien au raffinement de l'approximation μ d'une valeur propre de A. Cette approximation peut être par exemple obtenue en appliquant les techniques de localisation introduites à la Section 3.3. Les itérations inverses peuvent aussi être utilisées efficacement pour calculer le vecteur propre associé à une valeur propre (approchée) donnée.
En vue de l'analyse de convergence des itérations (3.12), supposons A diagonalisable et décomposons $\mathbf{q}^{(0)}$ sous la forme (3.6). En procédant de la

même manière que pour la méthode de la puissance, on a

$$\tilde{\mathbf{q}}^{(k)} = \mathbf{x}_m + \sum_{i=1, i\neq m}^{n} \frac{\alpha_i}{\alpha_m} \left(\frac{\xi_i}{\xi_m}\right)^k \mathbf{x}_i,$$

où les $\mathbf{x}_i$ sont les vecteurs propres de M_μ^{-1} (et donc aussi ceux de A), et les α_i sont comme en (3.6). Par conséquent, en rappelant la définition des ξ_i et en utilisant (3.11), on obtient

$$\lim_{k\to\infty} \tilde{\mathbf{q}}^{(k)} = \mathbf{x}_m, \quad \lim_{k\to\infty} \sigma^{(k)} = \lambda_m.$$

La convergence sera d'autant plus rapide que μ est proche de λ_m.

3.4.3 Méthode de déflation (Déflation de Householder)

On cherche à calculer les premières (plus grandes ou plus petites) valeurs propres d'une matrice $\mathrm{A} \in \mathbb{R}^{n\times n}$. Supposon que les valeurs propres ordonnées vérifient la condition (3.3) et que les paires valeurs propres/vecteurs propres $(\lambda_1, \mathbf{x}_1)$ aient été calculées en utilisant la méthode de la puissance. La matrice A peut alors être transformée en :

$$\mathrm{A}_1 = \mathrm{HAH} = \begin{bmatrix} \lambda_1 & \mathbf{b}^T \\ 0 & \mathrm{A}_2 \end{bmatrix},$$

où $\mathbf{b} \in \mathbb{R}^{n-1}$, H est la matrice de Householder telle que $\mathrm{H}\mathbf{x}_1 = \alpha\mathbf{x}_1$ pour $\alpha \in \mathbb{R}$, et où les valeurs propres de $\mathrm{A}_2 \in \mathbb{R}^{(n-1)\times(n-1)}$ sont les mêmes que celles de A, exceptée λ_1. La matrice H est donnée par :

$$\mathrm{H} = \mathrm{I} - 2\mathbf{v}\mathbf{v}^T/\|\mathbf{v}\|_2^2 \text{ avec } \mathbf{v} = \mathbf{x}_1 \pm \|\mathbf{x}_1\|_2 \mathbf{e}_1.$$

La méthode de déflation consiste à calculer la seconde valeur propre dominante (ou "sous-dominante") de A en appliquant la méthode de la puissance à A_2, à condition que λ_2 et λ_3 aient des modules distincts. Une fois calculée λ_2, le vecteur propre correspondant $\mathbf{x}_2$ peut être calculé en appliquant la méthode de la puissance inverse à la matrice A avec $\mu = \lambda_2$ (voir Section 3.4.2). On procède de même pour les autres valeurs propres et vecteurs propres de A.

3.5 Méthodes globales de recherche de valeurs propres

3.5.1 Méthode de Jacobi

La méthode de Jacobi est une méthode itérative applicable à une matrice $\mathrm{A} = (a_{ij})_{1\leq i,j\leq N}$ symétrique. Elle consiste à faire opérer le groupe des rotations planes sur A, c'est-à-dire à multiplier A par des transformations orthogonales afin de la mettre sous forme diagonale, les éléments diagonaux étant les valeurs propres de la matrice A.
On considére la matrice $H_{pq}(\alpha) = H = (h_{ij})_{1\leq i,j\leq N}$ dont ses éléments sont égaux à ceux de la matrice identité sauf pour les quatre valeurs suivantes $h_{pp} = \cos(\alpha), h_{pq} = \sin(\alpha), h_{qq} = \cos(\alpha)$ et $h_{qp} = -\sin(\alpha)$, avec $p < q$ et $\alpha \in [-\pi, \pi]$. La matrice H est orthogonale $H^t H = I$.
Exemple : Dans IR^2 on a $H = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$.
Pour un vecteur $\boldsymbol{x} \in \mathbb{R}^2, H\boldsymbol{x}$ est la rotation plane de $\boldsymbol{x}$ autour de l'origine d'un angle α (dans le sens des aiguilles d'une montre lorsque $\alpha > 0$ et dans le sens inverse des aiguilles d'une montre lorsque $\alpha < 0$).
Soit $\mathrm{A} = (a_{ij})_{1\leq i,j\leq N}$ une matrice symétrique avec $N \geq 2$, considérons le produit $A' = AH$. Évidemment, la seule différence entre A' et A réside dans les colonnes p et q ; ces colonnes de A' sont des combinaisons linéaires des deux mêmes colonnes de A, pour $c = \cos\alpha$ et $s = \sin\alpha$ on a :

$$\left.\begin{array}{ll} a'_{ip} & = a_{ip}c - a_{iq}s \\ a'_{iq} & = a_{ip}s + a_{iq}c \end{array}\right\}, \quad i = 1, 2, \ldots, N.$$

La multiplication de A' par H^T à gauche donne un résultat similaire, mais affecte les lignes p et q, plutôt que les colonnes p et q. L'écriture $A_1 = H^T A' =$ donne

$$\left.\begin{array}{ll} a^{(1)}_{pj} & = a'_{pj}c - a'_{qj}s \\ a^{(1)}_{qj} & = a'_{pj}s + a'_{qj}c \end{array}\right\}, \quad j = 1, 2, \ldots, N.$$

La combinaison de ces équations montre que

$$\left.\begin{array}{ll} a^{(1)}_{pp} & = a_{pp}c^2 - 2a_{pq}sc + a_{qq}s^2, \\ a^{(1)}_{qq} & = a_{pp}s^2 + 2a_{pq}sc + a_{qq}c^2, \\ a^{(1)}_{pq} & = (a_{pp} - a_{qq})\, sc + a_{pq}\left(c^2 - s^2\right) = a^{(1)}_{qp}. \end{array}\right\}. \tag{3.13}$$

Les éléments restants de $A_1 = H^T AH$ dans les colonnes p et q sont donnés par les expressions

$$\left.\begin{aligned} a_{ip}^{(1)} &= a_{ip}c - a_{iq}s \\ a_{iq}^{(1)} &= a_{ip}s + a_{iq}c \end{aligned}\right\}, \quad i = 1, 2, \ldots, N, \quad i \neq p, q.$$

La matrice $A_1 = H^T AH$ est évidemment symétrique, donc les éléments non diagonaux de A_1 dans les lignes p et q sont également donnés par les mêmes expressions.
Enfin, on remarque que tous les éléments de H qui ne se trouvent ni en ligne p ni q ni en colonne p ou q sont les mêmes que les éléments correspondants de A, c'est-à-dire,

$$a_{ij}^{(1)} = a_{ij}, \qquad \text{si } i \neq p, q \quad \text{et } j \neq p, q.$$

On peut donc choisir α dans (3.13) de sorte que $a_{pq}^{(1)} = a_{qp}^{(1)} = 0$, c'est-à-dire tel que

$$\operatorname{tg}(2\alpha) = \frac{2a_{pq}}{a_{pp} - a_{qq}}, \tag{3.14}$$

donc

$$\alpha = \frac{1}{2} \tan^{-1} \frac{2a_{pq}}{a_{qq} - a_{pp}} \in [-\pi/4, \pi/4]. \tag{3.15}$$

Pour voir cela, appliquez les identités trigonométriques $c^2 - s^2 = \cos(2\alpha)$ et $sc = \frac{1}{2}\sin(2\alpha)$ à $a_{pq}^{(1)}$ dans (3.13), avec $a_{pq}^{(1)} = 0$.
Nous pouvons éviter les calculs trigonométriques dans la formule (3.15) pour α en écrivant $t = s/c$, et en voyant que t vérifie

$$(a_{pp} - a_{qq})\, t + a_{pq} \left(1 - t^2\right) = 0. \tag{3.16}$$

- Si $a_{pq} = 0$, nous pouvons nous assurer que (3.16) est valable pour $t = 0$ (ce qui correspond au choix $\alpha = 0$).
- Si $a_{pq} \neq 0$ et $a_{pp} = a_{qq}$, on met $t = 1$ (correspondant à $\alpha = \pi/4$).
- Enfin, si $a_{pq} \neq 0$ et $a_{pp} \neq a_{qq}$, on résout l'équation quadratique (3.16) ; il y aura deux racines réelles distinctes, nous définissons donc t comme celle qui est la plus petite en valeur absolue. Après avoir détermié t,

on utilise ensuite la relation $\dfrac{1}{\cos^2\alpha} = 1 + \tan^2\alpha$ pour calculer c par $c = 1/\left(1+t^2\right)^{1/2}$, puis s de $s = ct$.

On aura alors

$$\left(a_{pp}^{(1)}\right)^2 + \left(a_{qq}^{(1)}\right)^2 = a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2,$$

et comme $a_{ii}^{(1)} = a_{ii}$ pour $i \neq p$ ou q

$$\sum_{i=1}^{n} \left(a_{ii}^{(1)}\right)^2 = \sum_{i=1}^{n} a_{ii}^2 + 2a_{pq}^2,$$

en passant de A à A_1 la somme des carrés des éléments diagonaux augmente de la quantité $2a_{pq}^2$. En itérant ce processus, on obtient

$$A_{k+1} = \left(H_1 H_2 \dots H_k\right)^t A \left(H_1 H_2 \dots H_k\right).$$

La suite des matrices A_k converge vers une matrice diagonale dont les éléments diagonaux sont les valeurs propres de la matrice initiale A. La suite des matrices $P_k = H_1 H_2 \dots H_k$ converge vers la matrice dont les colonnes sont constituées de vecteurs propres. Au cours des itérations un terme peut redevenir nul, mais on a :

$$\lim_{k\to\infty} \sum_{i\neq j} \left(a_{ij}^{(k)}\right)^2 = 0.$$

On arrête l'itération quand

$$1 - \frac{\sum_{i=1}^{n} \left(a_{ii}^{(k)}\right)^2}{\sum_{i=1}^{n} \left(a_{ii}^{(k+1)}\right)^2} < \varepsilon.$$

En pratique, on a le choix à chaque pas d'itération du couple (p, q). On définit différentes stratégies.

- Dans la méthode de **Jacobi classique**, on choisit (p, q) tels que

$$\left|a_{pq}^{(k)}\right| = \sup_{i\neq j} \left|a_{ij}^{(k)}\right|.$$

- Dans la méthode de **Jacobi cyclique**, on effectue un balayage systématique en prenant pour (p, q) les couples $(1, 2), (1, 3), \dots, (1, n)$ puis $(2, 3), \dots, (2, n)$, etc., jusqu'à $(n-1, n)$.

Remarque 3.2 *La méthode de Jacobi est stable, mais sa convergence est lente, ce qui en fait une méthode très peu utilisée.*

3.5.2 Méthode de Rutishauser (LU)

La méthode de Rutishauser est fondée sur la décomposition LU où L est une matrice triangulaire inférieure dont les éléments diagonaux sont égaux à 1 et U une matrice triangulaire supérieure. L'algorithme est le suivant : On décompose A en $A = L_1U_1$ selon les principes de la décomposition LU. Connaissant les matrices U_1 et L_1, on forme le produit $B_1 = U_1L_1$ qui a les mêmes valeurs propres que A. On cherche alors la décomposition LU de $B_1 = L_2U_2$. On itère le processus

$$\begin{aligned} B_k &= U_kL_k, \\ B_k &= L_{k+1}U_{k+1}. \end{aligned}$$

En remarquant que $L_kB_k = L_kU_kL_k = B_{k-1}L_k = U_{k-1}L_{k-1}L_k$ on peut écrire $B_k = P_k^{-1}AP_k$. La suite des matrices triangulaires supérieures $B_k = P_k^{-1}AP_k$ où $P_k = L_1L_2 \dots L_k$ converge vers une matrice B dont les éléments diagonaux sont les valeurs propres de A. Les vecteurs propres de A s'expriment en fonction des vecteurs propres de B. Soit V les vecteurs propres de B. La matrice $E_k = P_kV$ converge vers la matrice des vecteurs propres de A. Si A est une matrice symétrique définie positive, la méthode converge. Au-delà d'un certain indice d'itération la convergence devient très lente, il faut un grand nombre d'itérations pour gagner en précision sur le calcul des valeurs propres. En particulier, lorsque les valeurs propres sont égales ou peu différentes, la convergence peut être très lente.

3.5.3 Méthode de Francis (QR)

La méthode de Francis est identique à la méthode de Rutishauser mais au lieu de la décomposition LU, elle utilise la décomposition QR. À chaque étape, la matrice B_k est mise sous la forme d'un produit Q_kR_k où Q_k est une matrice unitaire et R_k une matrice triangulaire supérieure. Ces matrices sont réutilisées pour former la matrice $B_{k+1} = R_kQ_k$ qui est à son tour décomposée.

L'algorithme est le suivant : On décompose la matrice A en $A = Q_1 R_1$. Connaissant les matrices R_1 et Q_1, on forme le produit $B_1 = R_1 Q_1$. Puis, on décompose B_1 en $B_1 = Q_2 R_2$. A chaque étape, on décompose B_k en

$$B_k = Q_{k+1} R_{k+1},$$

les matrices sont réutilisées pour calculer

$$B_{k+1} = R_{k+1} Q_{k+1}.$$

La matrice B_k

$$B_{k+1} = P_k^* A P_k \quad \text{avec } P_k = Q_1 Q_2 \dots Q_k,$$

est une matrice triangulaire supérieure ayant sur sa diagonale les valeurs propres de A. La matrice P_k est la matrice des vecteurs propres de A, i.e. dont les colonnes sont les vecteurs propres associés. Pour obtenir une décomposition QR, on introduit les matrices de Jacobi $H_{p,q}$ en choisissant α de façon à annuler les coefficients triangulaires inférieurs $(a_{2,1}, \dots, a_{n,1})$, puis $a_{3,2}, \dots a_{n,2}$, etc. Les matrices

$$R_1 = H_{n,n-1}^t \dots H_{n,1}^t H_{n-1,n-2}^t \dots H_{n-1,1}^t \dots H_{3,2}^t H_{3,1}^t H_{2,1}^t A,$$

et

$$Q_1 = H_{2,1} H_{3,1} H_{3,2} \dots H_{n-1,1} \dots H_{n-1,n-2} H_{n,1} \dots H_{n,n-1} A,$$

satisfont la décomposition QR.

Bibliographie