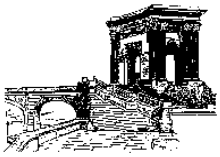
Université de Montpellier

Année Universitaire 2024-2025



Faculté des Sciences

Mathématiques

Analyse Numérique des Équations Différentielles.

Pascal Azerad

8 avril 2025

Table des matières

Les sections précédées d'une étoile (⋆) sont des compléments facultatifs.

0.1 Motivations

La science moderne tire son origine d'un livre fondateur Origine Philosophiae Naturalis Principia Mathematica, Londres, 1687, Sir Isaac Newton, Trinity college, Cambridge où il énonce le principe fondamental de la mécanique

$$m\ddot{x} = F(x)$$

Par exemple pour un ressort

$$m\ddot{x} = -kx$$

un pendule

$$L\ddot{\theta} = -g\sin\theta$$

ou bien la chute libre (non recommandée)

$$\ddot{x} = -g$$

le problème à N corps soumis à l'attraction universelle de la gravitation :

$$m_j\,\ddot{x_j} = G\sum_{k\neq j} m_j\,m_k\frac{(x_k - x_j)}{||x_k - x_j||^3.}$$

Pour $N = 2$ Képler a montré que les trajectoires sont des coniques.

Pour $N \geq 3$ c'est beaucoup plus difficile et on peut obtenir des trajectoires surprenantes, dont certaines n'ont été démontrées que très récemment, par exemple le célèbre « huit » de Alain Chenciner. On renvoie au site suivant pour des illustrations.
`http://ciel.mmi-lyon.fr/deux-astres-en-tete-a-tete/choregraphies/`

Les équations différentielles interviennent dans de nombreuses disciplines et ne sont pas seulement utilisées pour décrire les mouvements des systèmes de points matériels comme le pendule, les planètes idéalisées en des points matériels. Elles permettent de modéliser aussi que variations de courants ou de différences de potentiels dans les circuits électriques ou encore l'évolution des espèces en écologie ou encore l'évolution des épidémies, sujet d'actualité. Par exemple le fameux système SIR (sound, infected, recovered) qui s'écrit ainsi

$$\begin{cases} \dfrac{dS}{dt} = & -p\cdot I\cdot S \\ \dfrac{dI}{dt} = & p\cdot S\cdot I - \alpha I \\ \dfrac{dR}{dt} = & \alpha\cdot I \end{cases}$$

La plupart des équations issues de situations réelles n'ont pas de solution exprimables à l'aide des fonctions usuelles donc il est nécessaire de les *calculer numériquement par des approximations rigoureuses.* C 'est le but de ce cours que de présenter des méthodes numériques pour approcher la solution

$$\mathbf{y} :\ t\in\mathbb{R}\ \to \mathbf{y}(t)\in\mathbb{R}^n$$

d'une équation différentielle ordinaire :

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

où la fonction $\mathbf{f} :\ (t,\mathbf{y})\in\mathbb{R}\times\mathbb{R}^n \to \mathbf{f}(t,\mathbf{y})\in\mathbb{R}^n$ est donnée et ainsi que $\mathbf{y}(t_0)\in\mathbb{R}^n$ (problème de Cauchy).

Enfin on conseille vivement de visionner sur `youtube` la série 3BLUE1BROWN SERIES Saison 4 Episode 1 Differential equations, studying the unsolvable `https://youtu.be/p_di4Zn4wz4`

Les sections précédées d'une étoile ($\star$) sont des compléments facultatifs.

Vous trouverez à la fin de ce document une bibliographie non exhaustive de livres que vous pouvez éventuellement consulter.

Ce document, qui s'appuie sur un poly autrefois rédigé par Michel Cuer, est un version 1.0 et contient de nombreuses coquilles, merci de les signaler à l'auteur qui les corrigera au fur et à mesure.

Chapitre 1

Résolution de systèmes d'équations différentielles ordinaires (problèmes de Cauchy)

Ce chapitre est une introduction aux méthodes numériques de résolution des équations différentielles ordinaires et nous allons traiter successivement :

— quelques rappels ou compléments de théories mathématiques classiques ;

— une introduction aux méthodes numériques de résolution des problèmes de conditions initiales (problème de Cauchy) pour les équations différentielles ordinaires où on donnera des notions concernant les méthodes à un pas, les méthodes multipas et les problèmes raides.

On renvoie à la bibliographie en fin de document pour des compléments, en particulier à

E. Hairer, S.P. Nørsett, G. Wanner,1993, Solving ordinary differential equations I. Nonstiff problems : Springer-Verlag, Berlin.

Les documents suivants, en ligne sur internet, sont particulièrement recommandés. Le cours de G. Wanner `http://www.unige.ch/~wanner/Numi.html` ainsi que celui de E. Hairer `http://www.unige.ch/~hairer/poly/chap3.p`

1.1 Le problème de Cauchy pour les systèmes d'équations différentielles ordinaires

Étant donnés une fonction $\mathbf{f} : [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$, un réel $t_0 \in [a,b]$ où les réels a, b vérifient $a < b$ et en pratique t_0 est une des extrémités de $[a,b]$, et un vecteur $\mathbf{y}_0 \in \mathbb{R}^m$, par problème de Cauchy pour le système d'équations [1] différentielles $\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t))$, on entend la recherche d'une fonction $\mathbf{y}$ de $[a,b]$ dans $\mathbb{R}^m$ telle que :

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) \text{ pour } t \in [a,b] \ \mathbf{y}(t_0) = \mathbf{y}_0, \tag{1.1}$$

où $\mathbf{y}'(t) = \lim_{h\to 0} \frac{\mathbf{y}(t+h)-\mathbf{y}(t)}{h}$ est la dérivée en t de la fonction vectorielle $\mathbf{y}$.

Le ***théorème de Cauchy-Lipschitz*** énonce que si $\mathbf{f}$ est une fonction continue de $[a,b] \times \mathbb{R}^m$ dans $\mathbb{R}^m$ telle qu'il existe une norme sur $\mathbb{R}^m \|.\|$ et une constante $L > 0$ pour lesquelles (on dit alors que $\mathbf{f}$ est Lipschitzienne en $\mathbf{y}$ de constante de Lipschitz L) :

$$\|\mathbf{f}(t, \mathbf{y}^{(2)}) - \mathbf{f}(t, \mathbf{y}^{(1)})\| \leq L\|\mathbf{y}^{(2)} - \mathbf{y}^{(1)}\|, \text{ pour tout } (t, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}) \in [a,b] \times \mathbb{R}^m \times \mathbb{R}^m, \tag{1.2}$$

alors (1.1) a une solution et une seule définie sur tout l'intervalle $t \in [a,b] \to \mathbf{y}(t) \in \mathbb{R}^m$ continûment différentiable.

Preuve. Voici une démonstration, basée sur le théorème du point fixe de Banach-Picard, dans le cas où $t_0 = a$ (il n'est pas difficile de la modifier pour l'étendre au cas $t_0 = b$ et ensuite de traiter le cas général). Les conditions (1.1) sont équivalentes à :

1. Si $\mathbf{f}$ ne dépend pas de t on parle de système autonome.

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))ds, \tag{1.3}$$

équation fonctionnelle à laquelle on peut appliquer la méthode des approximations successives qui engendre une suite de fonctions de $[a, b]$ dans $\mathbb{R}^m$, $t \to \mathbf{y}^{(k)}(t)$ définies par :

$$\mathbf{y}^{(0)}(t) = \mathbf{y}_0,\ \mathbf{y}^{(k+1)}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}^{(k)}(s))ds,\ k \geq 0 \tag{1.4}$$

Pour obtenir le résultat il suffit donc d'établir que dans une espace fonctionnel complet convenable X, l'application[2] $\Phi : \mathbf{y} \in X \to \Phi(\mathbf{y}) : t \in [a = t_0, b] \to \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))ds$ est contractante donc a un unique point fixe. On choisit alors l'espace X des fonctions continues de $[a, b]$ dans $\mathbb{R}^m$ muni de la norme $\|\mathbf{y}\|_X = \max_{t\in[a=t_0,b]} e^{-k(t-a)}\|\mathbf{y}(t)\|$:

$$X = \{\mathbf{y} \in C([a, b]; \mathbb{R}^m); \|\mathbf{y}\|_X = \max_{t\in[a=t_0,b]} e^{-k(t-a)}\|\mathbf{y}(t)\|\};$$

cet espace vectoriel normé est complet et on a :

$$\begin{aligned}
&\|\Phi(\mathbf{y}^{(2)}) - \Phi(\mathbf{y}^{(1)})\|_X = \max_{t\in[a=t_0,b]} e^{-k(t-t_0)}\| \textstyle\int_{t_0}^{t} (\mathbf{f}(s, \mathbf{y}^{(2)}(s)) - \mathbf{f}(s, \mathbf{y}^{(1)}(s)))ds\| \\
&\leq \max_{t\in[a=t_0,b]} e^{-k(t-t_0)} \textstyle\int_{t_0}^{t} \|\mathbf{f}(s, \mathbf{y}^{(2)}(s)) - \mathbf{f}(s, \mathbf{y}^{(1)}(s))\|ds \\
&\leq \max_{t\in[a=t_0,b]} e^{-k(t-t_0)} \textstyle\int_{t_0}^{t} \underbrace{L}_{Le^{k(s-t_0)}e^{-k(s-t_0)}} \|\mathbf{y}^{(2)}(s) - \mathbf{y}^{(1)}(s)\|ds \\
&\leq \max_{t\in[a=t_0,b]} e^{-k(t-t_0)} \textstyle\int_{t_0}^{t} Le^{k(s-t_0)} \underbrace{\max_{s'\in[a=t_0,b]} e^{-k(s'-t_0)}\|\mathbf{y}^{(2)}(s') - \mathbf{y}^{(1)}(s'))\|}_{\|\mathbf{y}^{(2)}-\mathbf{y}^{(1)}\|_X} ds \\
&= L \max_{t\in[a=t_0,b]} e^{-k(t-t_0)} \tfrac{e^{k(t-t_0)-1}}{k}\|\mathbf{y}^{(2)} - \mathbf{y}^{(1)}\|_X \\
&= \tfrac{L}{k}(1 - e^{-k(b-a)})\|\mathbf{y}^{(2)} - \mathbf{y}^{(1)}\|_X < \tfrac{L}{k}(1 - e^{-k(b-a)})\|\mathbf{y}^{(2)} - \mathbf{y}^{(1)}\|_X
\end{aligned}$$

et en choisissant $k > L$ on a bien une application contractante.

Remarques ou compléments.

i) Il y a un théorème de Cauchy-Peano[3] qui donne l'existence dès que $\mathbf{f}$ est continue mais, sans conditions supplémentaires, il peut ne pas y avoir unicité. Le cas $y'(t) = 2|y|^{1/2}(1+y)$ pour $t \geq 0$, $y(0) = 0$ est un premier exemple ; $y(t) = 0$ est solution sur $[0, +\infty[$, mais quel que soit $a \geq 0$, $y_a(t) = \begin{cases} 0 \,\text{pour}\, t \in [0, a] \\ \tan^2(t-a) \,\text{pour}\, t \in [a, a + \frac{\pi}{2}[\end{cases}$ est encore une solution dont le domaine d'existence est $[0, a + \frac{\pi}{2}[$. Un cas plus élémentaire est proposé en exercice de TD.

La formulation précédente du théorème de Cauchy Lipschitz est assez restrictive ; avec le même schéma de démonstration, on peut établir un résultat d'existence et d'unicité ***locale*** en changeant « pour tout $(t, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}) \in [a, b] \times \mathbb{R}^m \times \mathbb{R}^m$ » dans l'hypothèse (1.2) en « pour tout $(t, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}) \in [a, b] \times O \times O$ » où O est un ouvert de $\mathbb{R}^m$, $\mathbf{f}$ étant une application continue de $[a, b] \times O$ dans O. Dans ce cas plus général mais quand même éclairant, le domaine d'existence n'est alors pas forcément $\mathbb{R}$ tout entier ou l'intervalle I sur lequel on pose le problème comme le montre l'exemple $y'(t) = 2ty(t)^2$ pour $t \in \mathbb{R}$, $y(0) = 1$ dont la seule solution est $y(t) = \frac{1}{1-t^2}$ et n'est définie que pour $t \in]-1, 1[$. Il convient, dans ce cadre général, d'introduire la notion de solution maximale (les exemples cités sont tous de telles solutions maximales).

ii) Les techniques utilisables pour les équations différentielles intégrables "à la main" sont importantes et on peut en trouver par exemple avec un logiciel de calcul formel ou le moteur `wolframalpha.com` (Bernoulli, Clairault, linéaire, variable séparable ...) ; voir aussi le début du livre de E. Hairer, S.P. Nørsett, G. Wanner (1993). Mais ce sont des cas particuliers très rares. Dans la réalité, la plupart du temps on ne sait pas intégrer exactement une équation différentielle.

iii) Une équation différentielle d'ordre $p > 1$ est de la forme :

2. Ce langage fonctionnel est tel que par $\mathbf{y}$ on entend la fonction $t \in [a, b] \to \mathbf{y}(t) \in \mathbb{R}^m$.

3. La démonstration de ce théorème n'est pas au programme de L3 parce qu'elle fait appel au théorème d'Ascoli qui donne un critère pour savoir si un ensemble de fonctions continues sur un compact est compact, si bien qu'on peut en extraire une sous suite qui converge.

$$\frac{d^p y}{dt^p}(t) = f(t, y(t), y'(t), ..., \frac{d^{p-1} y}{dt^{p-1}}(t)) \tag{1.5}$$

et le problème de Cauchy correspondant consiste à calculer la fonction $t \to y(t)$ vérifiant en plus des conditions initiales "$y(t_0)$, $y'(t_0)$, ..., $\frac{d^{p-1}y}{dt^{p-1}}(t_0)$ données". Il est important de comprendre, aussi bien sur le plan formel ou théorique que pratique [4] qu'une telle équation se ramène, en posant $\mathbf{z}(t) = (y(t), y'(t), ..., \frac{d^{p-1}y}{dt^{p-1}}(t))^T \in \mathbb{R}^p$ au système différentiel du premier ordre :

$$\mathbf{z}'(t) = \mathbf{f}(t, \mathbf{z}(t)) \text{ avec } \mathbf{f}(t, \mathbf{z}(t)) = \begin{pmatrix} z_2(t) \\ \vdots \\ z_p(t) \\ f(z_1(t), z_2(t), .., z_p(t)) \end{pmatrix} \tag{1.6}$$

où, bien sûr, $z_j(t) = \frac{d^{j-1}y}{dt^{j-1}}$ pour $j = 1, ..., p$ (par convention $\frac{d^0 y}{dt^0} = y$). Détaillons cela sur un exemple essentiel, l'équation fondamentale de la mécanique, qui traduit la loi de Newton « Force = masse × accélération ». Soit m la masse du corps, repéré par sa position de son centre d'inertie $y(t) = (y_1(t), y_2(t), y_3(t)) \in \mathbb{R}^3$ à l'instant t et sa vitesse $y'(t) \in \mathbb{R}^3$, la loi de Newton s'écrit

$$m\frac{d^2 y}{dt^2} = F(y, y') = \big(F_1(y, y'), F_2(y, y'), F_3(y, y')\big)$$

La force $F(y, y')$ ne dépend pas en général explicitement de t donc on a un système différentiel autonome du second ordre. On le réécrit sous forme d'un système d'ordre 1 en posant $z(t) = (y(t), y'(t)) = (y_1, y_2, y_3, y_1', y_2', y_3')$:

$$\frac{dz}{dt} = (y', \frac{1}{m}F(y, y')$$

qui est donc une équation différentielle de la forme

$$\frac{dz}{dt} = G(z)$$

où la fonction $t \in R \mapsto z(t) \in \mathbb{R}^6$ et $G(z) = (z_4, z_5, z_6, \frac{1}{m}F(z_1, z_2, z_3, z_4, z_5, z_6) \in \mathbb{R}^6$

iv) Une équation différentielle s'interprète géométriquement comme la donnée d'un champ de vecteurs. La figure 1.1 correspond à l'exemple simple de l'équation $y' = y$. On a tracé le champ de vecteur $(1, y)$: A chaque point (t, y), on associe le vecteur $(1, y)$. Les solutions de l'équation différentielle sont les courbes $t \mapsto (t, y(t))$ telles que le vecteur tangent $(1, y'(t))$ est égal au champ de vecteur $(1, y)$. Cette interprétation est très fructueuse et permet de visualiser l'allure des solutions, même si on ne sait pas intégrer l'équation différentielle.

Sur le site `demonstrations.wolfram.com` on trouve des documents cdf [5] très instructifs permettant de visualiser divers champs de vecteurs associés à des EDO (Cf SlopeFields.cdf).

Ci dessous sur la figure 1.1 une copie d'écran où l'on voit le champ de vecteur associé à l'EDO $y' = t^2 \cdot y$ ainsi qu'une courbe intégrale, solution d'un problème de Cauchy particulier.

1.2 Introduction aux méthodes numériques de résolution d'équations différentielles.

On considère donc le problème de Cauchy pour une équation différentielle ordinaire, qui consiste à calculer une fonction $t \in [a, b] \to \mathbf{y}(t) \in \mathbb{R}^m$ telle que :

4. parce que, bien qu'il existe des schémas numériques adaptés aux équations d'ordre 2 par exemple, la majorité des codes, en particulier les codes `MATLAB`, sont écrits pour des systèmes différentiels d'ordre 1.

5. Wolfram computable document format : ce sont des fichiers pdf interactifs où l'on peut changer des données grâce à des menus.

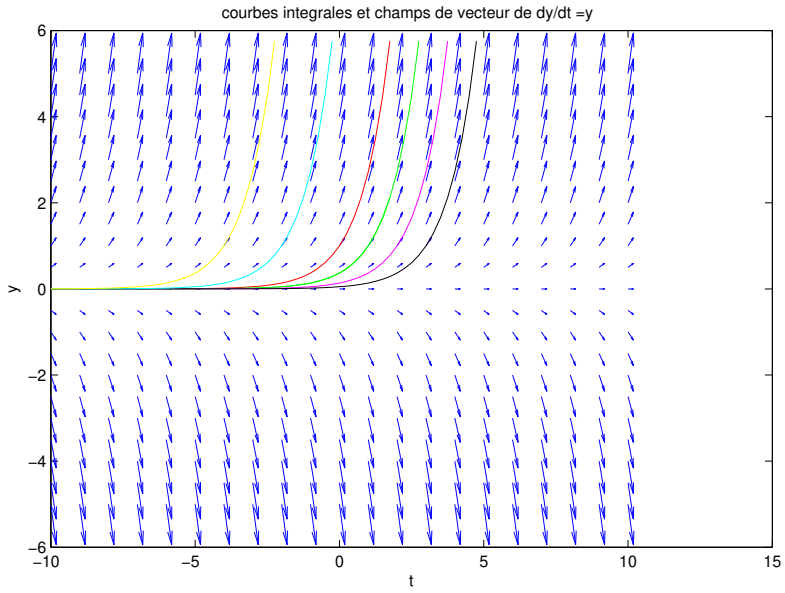


FIGURE 1.1 – Champ de vecteurs associé à l'équation différentielle $y' = y$.

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)) \text{ pour } t \in [a, b],\ \mathbf{y}(t_0) = \mathbf{y}_0, \tag{1.7}$$

où $\mathbf{y}'(t) = \frac{d\mathbf{y}}{dt}(t)$ désigne la dérivée de $\mathbf{y}$ par rapport à t au point (ou à l'instant) t et la fonction $\mathbf{f} : (t, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^m \to \mathbf{f}(t, \mathbf{y}) \in \mathbb{R}^m$ ainsi que le réel t_0 et le vecteur $\mathbf{y}_0 \in \mathbb{R}^m$ sont donnés. On suppose aussi que $\mathbf{f}$ est continue de $[a, b] \times \mathbb{R}^m$ dans $\mathbb{R}^m$ et vérifie la condition de Lipschitz "en $\mathbf{y}$" :

$$\|\mathbf{f}(t, \mathbf{y}^{(1)}) - \mathbf{f}(t, \mathbf{y}^{(2)})\| \leq L\|\mathbf{y}^{(1)} - \mathbf{y}^{(2)}\| \text{ pour tout } (t, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}) \in [a, b] \times \mathbb{R}^m \times \mathbb{R}^m, \tag{1.8}$$

pour une norme $\|.\|$ quelconque dans $\mathbb{R}^m$, par exemple $\|\mathbf{z}\| = \|\mathbf{z}\|_\infty = \max_{1 \leq j \leq m} |z_j|$ et une constante $L > 0$. Pour la simplicité on supposera en plus $t_0 = a$.

1.2.1 Les méthodes à un pas

Étant donnée une suite de réels t_0, t_1, ..., t_N telle que [6] $a = t_0 < t_1 < ... < t_N = b$, on pose $h_n = t_{n+1} - t_n$ et $h = \max_{0 \leq n \leq N-1} h_n$. La *méthode d'Euler*, archétype des méthodes à un pas [7], essentiellement *méthodes de Runge Kutta* pour la résolution des problèmes de Cauchy pour les équations différentielles ordinaires, appliqué à (1.7) consiste à calculer les quantités $\mathbf{y}_n \in \mathbb{R}^m$ qu'on espère être des approximations de $\mathbf{y}(t_n)$, définies par :

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h_n \mathbf{f}(t_n, \mathbf{y}_n) \text{ pour } 0 \leq n \leq N - 1. \tag{1.9}$$

La formule (1.9) est appelé schéma d'Euler.

Remarque. On peut la comprendre de deux façons : En partant de $y(t + h) = y(t) + \int_t^{t+h} y'(s)ds$ on écrit

$$y(t + h) = y(t) + \int_t^{t+h} f(s, y(s))ds \tag{1.10}$$

et on approche l'intégrale par la méthode du rectangle :

$$\int_t^{t+h} f(s, y(s))ds \approx hf(t, y(t)).$$

6. Ce qui suit s'applique aussi aux cas où $t_N < t_{N-1} < ... < t_1 < t_0$ en changeant les signes des h_n et en posant $h = \max_{0 \leq n \leq N-1} |h_n|$.

7. On dit aussi méthode à pas séparés.

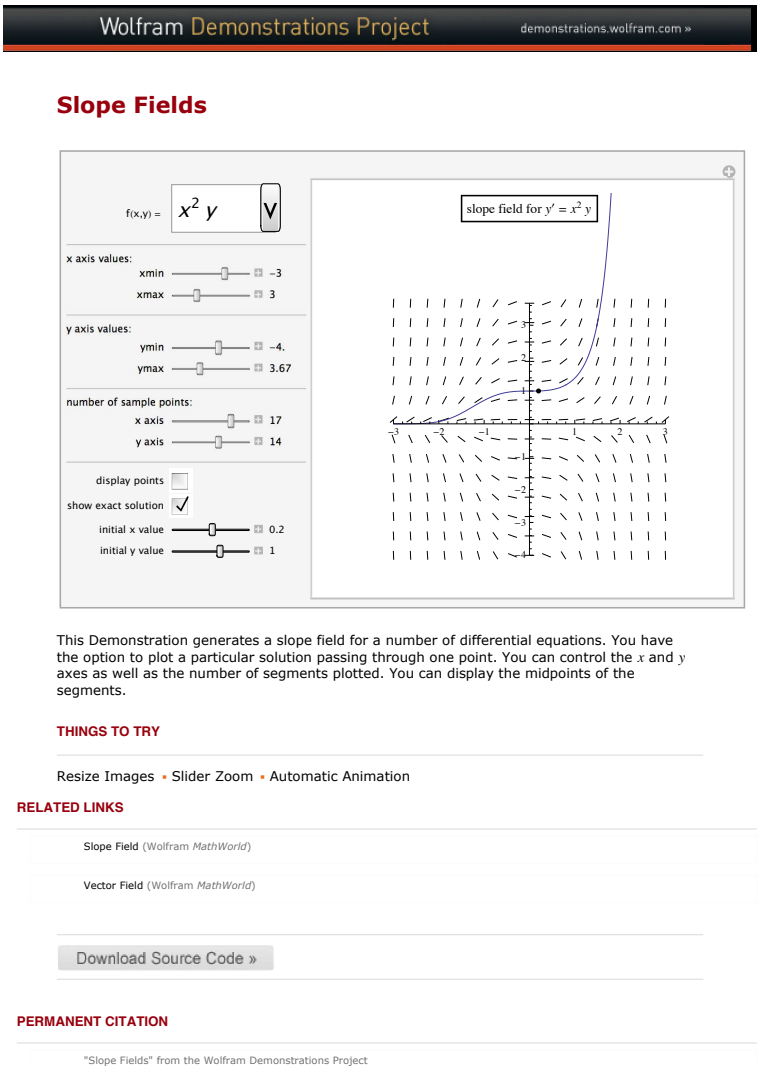


FIGURE 1.2 – champ de vecteurs associé à $y' = t^2\, y$.

On obtient ainsi

$$y(t+h) \approx y(t) + hf(t, y(t)).$$

Ou en approchant la dérivée par un taux d'accroissement.

$$y'(t) \approx \frac{y(t+h) - y(t)}{h}.$$

d'où l'on tire

$$\frac{y(t+h) - y(t)}{h} \approx f(t, y(t)).$$

□

Définition 1 *Soit $t \mapsto \mathbf{y}(t)$ la solution exacte de l'équation différentielle (1.7) On appelle* ***erreur de consistance*** *du schéma d'Euler 1.9 la quantité*

$$h \mapsto \boldsymbol{\epsilon}(h) := \mathbf{y}(t_0 + h) - \mathbf{y}(t_0) - h\,\mathbf{f}(t_0, \mathbf{y}(t_0)). \tag{1.11}$$

C'est en quelque sorte *l'erreur locale commise par le schéma* au point t_0. On espère que cette erreur $\epsilon(h)$ tend vers zéro quand le pas h tend vers zéro. En effet, en effectuant un développement de Taylor-Lagrange à l'ordre 2,

$$\boldsymbol{\epsilon}(h) = \mathbf{y}(t_0+h) - \mathbf{y}(t_0) - h\,\mathbf{f}(t_0,\mathbf{y}(t_0) = \mathbf{y}(t_0+h) - \mathbf{y}(t_0) - h\,\mathbf{y}'(t_0) = \frac{h^2}{2}\mathbf{y}''(t_0+\theta h) \leq Ch^2$$

si $\mathbf{y}''(t)$ est bornée. On a montré la proposition suivante.

Proposition 1 *Si la solution exacte $t \mapsto y(t)$ est $\mathcal{C}^2([a,b])$, l'erreur de consistance du schéma d'Euler est en $\mathcal{O}(h^2)$ quand le pas h tend vers 0. On dit que* ***l'erreur de consistance est d'ordre deux.***

1.2.1.1 Erreurs de consistance (locales) et erreurs globales

Pour établir la convergence d'une telle méthode, on introduit d'abord la suite $\{\boldsymbol{\epsilon}_n\}_{0\leq n\leq N-1}$ dans $\mathbb{R}^m$, qu'on appele ***suite des erreurs de consistance*** dans le schéma d'Euler pour la solution exacte[8] $\mathbf{y}(t)$, définie par :

$$\boldsymbol{\epsilon}_n = \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - h_n\mathbf{f}(t_n,\mathbf{y}(t_n)). \tag{1.12}$$

Alors la ***suite des erreurs globales*** $\mathbf{e}_n = \mathbf{y}(t_n) - \mathbf{y}_n$ ("solution exacte - solution approchée") satisfait :

$$\left\{\begin{array}{c} \mathbf{e}_0 = 0 \\ \mathbf{e}_{n+1}{=}\mathbf{e}_n + h_n(\mathbf{f}(t_n,\mathbf{y}(t_n)) - \mathbf{f}(t_n,\mathbf{y}_n)) + \boldsymbol{\epsilon}_n \,\text{pour}\, 0\leq n\leq N-1 \end{array}\right. . \tag{1.13}$$

En effet :

$$\mathbf{e}_{n+1} = \mathbf{y}(t_{n+1}) - \underbrace{\mathbf{y}_{n+1}}_{\mathbf{y}_n+h_n\mathbf{f}(t_n,\mathbf{y}_n)} = \underbrace{\mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - h_n\mathbf{f}(t_n,\mathbf{y}(t_n))}_{\boldsymbol{\epsilon}_n} + \underbrace{\mathbf{y}(t_n) + h_n\mathbf{f}(t_n,\mathbf{y}(t_n)) - \mathbf{y}_n - h_n\mathbf{f}(t_n,\mathbf{y_n})}_{\mathbf{e}_n+h_n(\mathbf{f}(t_n,\mathbf{y}(t_n))-\mathbf{f}(t_n,\mathbf{y}_n))}$$

Il en résulte que :

$$\left\{\begin{array}{c} \mathbf{e}_0 = 0 \\ \|\mathbf{e}_{n+1}\| \leq (1+h_nL)\|\mathbf{e}_n\| + \|\boldsymbol{\epsilon}_n\| \,\text{pour}\, 0\leq n\leq N-1 \end{array}\right. . \tag{1.14}$$

En effet la formule (1.13), l'inégalité triangulaire et la propriété (1.8) montrent que :

$$\|\mathbf{e}_{n+1}\| \leq \|\mathbf{e}_n\| + h_nL\|\underbrace{\mathbf{y}(t_n) - \mathbf{y}_n}_{\mathbf{e}_n}\| + \|\boldsymbol{\epsilon}_n\| = (1+h_nL)\|\mathbf{e}_n\| + \|\boldsymbol{\epsilon}_n\|.$$

Remarque. Si le schéma d'Euler (1.9) est remplacé par une formule plus précise (voir des exemples plus loin) de la forme :

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h_n\boldsymbol{\Phi}(t_n,\mathbf{y}_n,h_n) \quad \text{pour}\, 0\leq n\leq N-1, \tag{1.15}$$

où $\boldsymbol{\Phi}$ vérifie, pour une constante M :

$$\|\boldsymbol{\Phi}(t,\mathbf{y}^{(1)},h) - \boldsymbol{\Phi}(t,\mathbf{y}^{(2)},h)\| \leq M\|\mathbf{y}^{(1)} - \mathbf{y}^{(2)}\| \,\text{pour tout}\, (t,\mathbf{y}^{(1)},\mathbf{y}^{(2)}) \in [a,b]\times\mathbb{R}^m\times\mathbb{R}^m \,\text{et}\, h\geq 0 \,\text{assez petit} \tag{1.16}$$

alors introduisant la suite des erreurs de consistance :

$$\boldsymbol{\epsilon}_n = \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - h_n\boldsymbol{\Phi}(t_n,\mathbf{y}(t_n),h_n) \tag{1.17}$$

on voit que la suite des erreurs globales $\mathbf{e}_n = \mathbf{y}(t_n) - \mathbf{y}_n$ (même définition évidemment) satisfait encore (1.14) à condition de remplacer L par M. Un schéma sous la forme (1.15) vérifiant (1.17) est appelé un *schéma à un pas.*

8. Cette suite est en quelque sorte la suite des erreurs locales dues à la discrétisation.

1.2.1.2 Convergence de la méthode d'Euler et des méthodes à un pas consistantes.

On va établir que l'erreur globale $\max_{a\leq t_n\leq b}|y(t_n)-y_n|$ tend vers 0 quand le pas h tend vers 0. On dit alors que la méthode d'Euler converge.

Pour obtenir le résultat de convergence cherché, on va d'abord montrer (et le lien avec (1.14) est évident), avec les mêmes définitions de $h_n \geq 0$ et $L \geq 0$ le lemme suivant.

Lemme 1 *si $\{\theta_n\}_{0\leq n\leq N}$ et $\{\alpha_n\}_{0\leq n\leq N-1}$ sont deux suites de réels positifs telles que :*

$$\theta_{n+1} \leq (1+h_nL)\theta_n + \alpha_n \text{ pour } 0 \leq n \leq N-1 \tag{1.18}$$

alors :

$$\theta_n \leq e^{L(t_n-t_0)}\theta_0 + \sum_{i=0}^{n-1} e^{L(t_n-t_{i+1})}\alpha_i,\ 1 \leq n \leq N. \tag{1.19}$$

Preuve. On procède par récurrence.
Pour $n=0$ (1.18) donne

$$\theta_1 \leq (1+h_0L)\theta_0 + \alpha_0 \leq e^{Lh_0}\theta_0 + e^{L(t_1-t_1)}\alpha_0$$

où l'on a utilisé l'inégalité élementaire

$$(1+x) \leq e^x.$$

Donc (1.19) est vraie si $n=1$.
Supposons donc (1.19) vraie jusqu'au rang $n-1$: $\theta_{n-1} \leq e^{L(t_{n-1}-t_0)}\theta_0 + \sum_{i=0}^{n-2} e^{L(t_{n-1}-t_{i+1})}\alpha_i$ et montrons que la propriété est vraie au rang n. Appliquant (1.18) avec le bon indice, il vient

$$\theta_n \leq (1+h_{n-1}L)\theta_{n-1} + \alpha_{n-1} \leq (1+h_{n-1}L)(e^{L(t_{n-1}-t_0)}\theta_0 + \sum_{i=0}^{n-2} e^{L(t_{n-1}-t_{i+1})}\alpha_i) + \alpha_{n-1}$$

Utilisant à nouveau l'inégalité élémentaire $(1+x) \leq e^x$, on obtient

$$1 + h_{n-1}L \leq e^{L\,h_{n-1}} = e^{L\,(t_n-t_{n-1})}.$$

Donc $(1+h_{n-1}L)e^{L(t_{n-1}-t_0)} \leq e^{L(t_n-t_0)}$ et $(1+h_{n-1}L)e^{L(t_{n-1}-t_{i+1})} \leq e^{L(t_n-t_{i+1})}$
Ainsi

$$\theta_n \leq e^{L(t_n-t_0)}\theta_0 + \sum_{i=0}^{n-1} e^{L(t_n-t_{i+1})}\alpha_i.$$

La récurrence est terminée. ■

Remarque. De façon plus piétonne, on peut directement itérer la majoration 1.18.

$$\theta_{n+1} \leq (1+h_nL)\theta_n + \alpha_n$$

$$\begin{aligned}\theta_{n+1} \leq\ & (1+h_nL)(1+h_{n-1}L)\dots(1+h_0L)\theta_0 \\ & +(1+h_nL)(1+h_{n-1}L)\dots(1+h_1L)\alpha_0 \\ & +(1+h_nL)(1+h_{n-1}L)\dots(1+h_2L)\alpha_1 \\ & +\dots \\ & +(1+h_nL)\alpha_{n-1} \\ & +\alpha_n\end{aligned}$$

Les différentes lignes correspondent à l'amplification des erreurs commises aux pas de temps successifs. En majorant $1+hL \leq \exp hL$, on obtient

$$\begin{aligned}\theta_{n+1} \leq\ & e^{h_nL}e^{h_{n-1}L}\dots e^{h_0L}\theta_0 \\ & +e^{h_nL}e^{h_{n-1}L}\dots e^{h_1L}\alpha_0 \\ & +e^{h_nL}e^{h_{n-1}L}\dots e^{h_2L}\alpha_1 \\ & +\dots \\ & +e^{h_nL}\alpha_{n-1} \\ & +\alpha_n\end{aligned}$$

Ce qui donne bien

$$\theta_{n+1} \leq e^{L(t_{n+1}-t_0)}\theta_0 + e^{L(t_{n+1}-t_1)}\alpha_0 + e^{L(t_{n+1}-t_2)}\alpha_1 \ldots + \alpha_n.$$

□

Montrons maintenant la proposition.

Théorème 1 *si* $\mathbf{y}''$ *existe et est continue, le schéma d'Euler est convergent, l'erreur globale* $\max_{0\leq n\leq N} \|\mathbf{e}_n\|$ *est* $O(h)$ *où* $h = \max_{0\leq n\leq N-1} h_n$. *On dit que* ***le schéma d'Euler est d'ordre 1****.*

Preuve. Il suffit de poser $\theta_n = \| \mathbf{e}_n \|$ et $\alpha_n = \|\boldsymbol{\epsilon}_n\|$; (1.14) montre que (1.18) est vraie. Or d'après la proposition précédente, l'erreur de consistance est d'ordre deux, donc $\|\boldsymbol{\epsilon}_n\| \leq Ch_n^2$. Alors (1.19) donne, puisque la condition initiale donne $\theta_0 = \mathbf{y}(t_0) - \mathbf{y}_0 = 0$:

$$\| \mathbf{e}_n \| \leq \sum_{i=0}^{n-1} e^{L(t_n - t_{i+1})} \underbrace{\|\boldsymbol{\epsilon}_i\|}_{\leq Ch_i^2 \leq Chh_i} \leq e^{L(b-a)} Ch \underbrace{\sum_{i=0}^{n-1} h_i}_{b-a} \leq C(b-a)e^{L(b-a)}h = O(h)$$

■

On montre facilement en prenant le cas particulier de l'équation $y' = y,\ y(0 = 1$ (voir l'exercice 4 du TD 1) qu'on ne peut pas avoir une meilleure majoration que $\| \mathbf{e}_n \| = O(h)$.

Remarque. Dans le cas du problème de Cauchy on l'on connaît $y(t_0) = y_0$, $\mathbf{e}_0 = 0$ et la formule 1.19 s'écrit :

$$\| \mathbf{e}_n \| \leq e^{L(t_n - t_1)}\|\boldsymbol{\epsilon}_0\| + e^{L(t_n - t_2)}\|\boldsymbol{\epsilon}_1\| + \ldots + e^{L(t_n - t_{n-1})}\|\boldsymbol{\epsilon}_{n-2}\| + \|\boldsymbol{\epsilon}_{n-1}\|.$$

Chaque terme s'interprète : $e^{L(t_n-t_1)}\|\boldsymbol{\epsilon}_0\|$ correspond à la propagation-amplification de l'erreur de consistance ϵ_0 commise au temps t_1 jusqu'au temps t_n, $e^{L(t_n-t_2)}\|\boldsymbol{\epsilon}_1\|$ correspond à la propagation-amplification de l'erreur de consistance ϵ_1 commise au temps t_2 jusqu'au temps $t_n \ldots e^{L(t_n-t_{n-1})}\|\boldsymbol{\epsilon}_{n-2}\|$ correspond à la propagation-amplification de l'erreur de consistance ϵ_{n-2} commise au temps t_{n-1} jusqu'au temps t_n et enfin $\|\boldsymbol{\epsilon}_{n-1}\|$ est la dernière erreur de consistance commise à l'instant t_n. Ainsi la formule 1.19 décrit la façon dont *les erreurs se propagent et s'accumulent*. Plus une erreur est ancienne plus elle a le temps de s'amplifier exponentiellement. La figure 1.3 qui suit éclaire la preuve, attention les notations sont différentes car elle est tirée du magnifique poly en ligne de G. Wanner `http://www.unige.ch/~wanner/Numi.html` déjà cité dans la bibliographie).

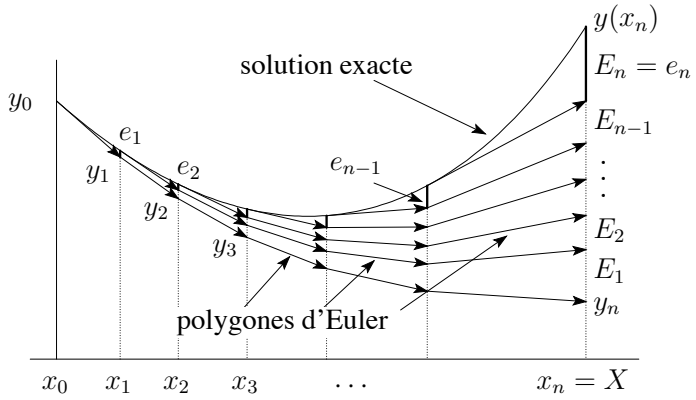


FIG. III.3: "Lady Windermere's Fan", Estimation de l'erreur globale

FIGURE 1.3 – l'éventail de Lady Windermere d'après Hairer-Wanner

□

Remarque. Si on suppose que la condition initiale $y(t_0) = y_0$ est vérifiée seulement de manière approchée, à cause de la précision finie des machines par exemple ou bien pour des raisons physiques de précision de mesure, la majoration (1.19) donne

$$\|\mathbf{e}_n\| \leq e^{L(t_n-t_0)}\|y(t_0) - y_0\| + e^{L(t_n-t_1)}\|\boldsymbol{\epsilon}_0\| + e^{L(t_n-t_2)}\|\boldsymbol{\epsilon}_1\| + \ldots + e^{L(t_n-t_{n-1})}\|\boldsymbol{\epsilon}_{n-2}\| + \|\boldsymbol{\epsilon}_{n-1}\|.$$

Donc l'erreur globale est toujours majorée par $\|\mathbf{e}_n\| \leq C\,(\|y(t_0) - y_0\| + h)$ de sorte que si $y(t_0) - y_0 \to 0$ le schéma converge. Le schéma est dit *stable* vis à vis des perturbations de la condition initiale. □

Le théorème 1 se généralise immédiatement aux *schémas à un pas* i.e. de la forme

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h_n\boldsymbol{\Phi}(t_n, \mathbf{y}_n, h_n) \quad \text{pour}\, 0 \leq n \leq N-1, \tag{1.20}$$

où $\boldsymbol{\Phi}$ vérifie, pour une constante M :

$$\|\boldsymbol{\Phi}(t, \mathbf{y}^{(1)}, h) - \boldsymbol{\Phi}(t, \mathbf{y}^{(2)}, h)\| \leq M\|\mathbf{y}^{(1)} - \mathbf{y}^{(2)}\| \,\text{pour tout}\, (t, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}) \in [a,b]\times\mathbb{R}^m \times \mathbb{R}^m \,\text{et}\, h \geq 0 \,\text{assez petit} \tag{1.21}$$

dont l'erreur de consistance est naturellement définie par :

$$\boldsymbol{\epsilon}(h) = \mathbf{y}(t+h) - \mathbf{y}(t) - h\,\boldsymbol{\Phi}(t, \mathbf{y}(t), h). \tag{1.22}$$

On peut énoncer le théorème de convergence des schémas à un pas.

Théorème 2 *si un schéma à un pas a une erreur de consistance $\mathcal{O}(h^{p+1})$ le schéma est convergent, l'erreur globale $\max_{0\leq n\leq N}\|\mathbf{e}_n\|$ est $\mathcal{O}(h^p)$ où $h = \max_{0\leq n\leq N-1} h_n$. On dit que* ***le schéma est d'ordre p****.*

La preuve est identique à celle du schéma d'Euler.

1.2.1.3 Les premières méthode de Runge-Kutta

Pour obtenir une méthode plus précise il faut améliorer l'erreur de consistance (1.17) et on peut remarquer en faisant un développement de Taylor que $\frac{\mathbf{y}(t+h)-\mathbf{y}(t)}{h} = \mathbf{y}'(t+\frac{h}{2}) + O(h^2)$ (au lieu de $\frac{\mathbf{y}(t+h)-\mathbf{y}(t)}{h} = \mathbf{y}'(t) + O(h)$). Le schéma dit *du point milieu* (Runge, 1895) s'écrit ainsi :

$$\begin{array}{l} \mathbf{u}_2 = \mathbf{y}_0 + \frac{h}{2}\mathbf{f}(t_0, y_0) \\ \mathbf{y}(t_0+h) \approx \mathbf{y}_1 = \mathbf{y_0} + hf(t_0 + \frac{h}{2}, \mathbf{u}_2) \end{array} \tag{1.23}$$

Remarque. On peut comprendre cette méthode en partant de $y(t+h) = y(t) + \int_t^{t+h} y'(s)ds$. On écrit

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s))ds \tag{1.24}$$

et on approche l'intégrale par la méthode du point milieu :

$$\int_t^{t+h} f(s, y(s))ds \approx hf(t + h/2, y(t + h/2)).$$

On obtient ainsi

$$y(t+h) \approx y(t) + hf(t + h/2, y(t + h/2)).$$

Mais on ne connaît pas $y(t + h/2)$. On effectue alors une *prédiction* :

$$y(t + h/2) \approx u_2 = y(t) + \frac{h}{2}f(t, y(t))$$

par la méthode d'Euler suivie d'une *correction* :

$$y(t+h) \approx y(t) + h\, f(t + h/2, u_2).$$

□

Pour démontrer qu'on obtient ainsi une méthode d'ordre 2, il faut comparer les développements de Taylor $\mathbf{y}(t+h)$ (solution exacte) et $\mathbf{y}_1$ (solution approchée). Or, on peut calculer $\mathbf{y}''(t) = \frac{d}{dt}\mathbf{f}(t, \mathbf{y}(t))$ en utilisant la différentielle de f.

$$d\mathbf{f} = \frac{\partial \mathbf{f}}{\partial t}dt + \frac{\partial \mathbf{f}}{\partial y}dy$$

donc "en divisant par dt"

$$\frac{d}{dt}\mathbf{f}(t, \mathbf{y}(t)) = \frac{\partial \mathbf{f}}{\partial t} + \frac{\partial \mathbf{f}}{\partial y}\frac{dy}{dt}$$

$$\mathbf{y}''(t) = \frac{d}{dt}\mathbf{f}(t, \mathbf{y}(t)) = \frac{\partial \mathbf{f}}{\partial t} + \frac{\partial \mathbf{f}}{\partial y}\mathbf{y}'(t)$$

En réutilisant le fait que $\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t))$ on obtient donc que $\mathbf{y}''(t) = \frac{\partial \mathbf{f}}{\partial t} + \frac{\partial \mathbf{f}}{\partial y}\,\mathbf{f}(t, \mathbf{y}(t))$. Ainsi en notant $\mathbf{f}_t = \frac{\partial \mathbf{f}}{\partial t}$ et $\mathbf{f}_\mathbf{y} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}$ (matrice jacobienne $m \times m$), on obtient

$$\mathbf{y}(t+h) = \mathbf{y}(t) + h\mathbf{y}'(t) + \tfrac{h^2}{2}\mathbf{y}''(t) + O(h^3) = \mathbf{y} + h\mathbf{f} + \tfrac{h^2}{2}(\mathbf{f}_t + \mathbf{f}_\mathbf{y}\mathbf{f}) + O(h^3) \tag{1.25}$$

D'autre part :

$$\mathbf{y}_1 = \mathbf{y} + h\mathbf{f}(t + \tfrac{h}{2}, \mathbf{y} + \tfrac{h}{2}\mathbf{f}(t, \mathbf{y})) = \mathbf{y} + h(\mathbf{f}(t, \mathbf{y}) + \tfrac{h^2}{2}(\mathbf{f}_t + \mathbf{f}_\mathbf{y}\mathbf{f}) + O(h^3) \tag{1.26}$$

On a ici effectué un développement de Taylor de la fonction *à deux variables*

$$\mathbf{f}(t+h, y+k) = \mathbf{f}(t, y) + h\frac{\partial \mathbf{f}}{\partial t}(t, y) + k\frac{\partial \mathbf{f}}{\partial y}(t, y) + O(h^2 + k^2)$$

où $k := \frac{h}{2}\mathbf{f}(t, \mathbf{y})$, si bien que finalement l'erreur de consistance introduite en (1.17) vérifie, en posant $\Phi(t, y; h) := f(t + h/2, y + h/2 f(t, y))$,

$$\boldsymbol{\epsilon}_n = \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - h_n\boldsymbol{\Phi}(t_n, \mathbf{y}(t_n)) = O(h_n^3) \tag{1.27}$$

donc on a une erreur de consistance en $O(h^3)$. De la même manière qu'au paragraphe précédent, on a un schéma à un pas de la forme (1.15). Il est facile (exercice !) de vérifier que $\Phi(t, y, h)$ est Lipschitzienne par rapport à la variable y et en suivant la même démarche que pour le schéma d'Euler, on démontre que le schéma du point milieu est convergent et que l'erreur globale est $O(h^2)$. On dit que **le schéma du point-milieu est d'ordre 2.**

Variante : on peut également approcher l'intégrale

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s))ds \tag{1.28}$$

par la méthode du trapèze :

$$\int_t^{t+h} f(s, y(s))ds \approx \frac{h}{2}\left(f(t, y(t)) + f(t+h, y(t+h))\right)$$

On obtient ainsi

$$y(t+h) \approx y(t) + \frac{h}{2}\left(f(t, y(t)) + f(t+h, y(t+h))\right).$$

Mais on ne connaît pas $y(t+h)$. On effectue alors une *prédiction* :

$$y(t_0 + h) \approx u_2 = y_0 + hf(t_0, y_0)$$

par la méthode d'Euler suivie d'une *correction*. Le schéma s'écrit alors

$$y(t_0 + h) \approx y_1 = y_0 + \frac{h}{2}\left(f(t_0, y_0) + f(t_0 + h, u2)\right).$$

Cette méthode dite du *trapèze explicite* est du même ordre que la méthode du point milieu. Par un développement de Taylor, cf TD, on estime l'erreur de consistance

$$y(t_0 + h) - y_1 = O(h^3),$$

si bien que le schéma est aussi d'ordre 2.

1.2.1.4 un schéma d'ordre 3 : le schéma de Heun

Présentons maintenant une méthode d'ordre 3 : la méthode de Heun. Elle repose sur la formule d'intégration de Gauss-Radau suivante :

$$\int_0^1 g(t)\,dt \approx \frac{1}{4}g(0) + \frac{3}{4}g(2/3)$$

qui est exacte pour les polynômes de degré inférieur ou égal à 2, ainsi qu'on le vérifie aisément. On en déduit

$$y(t+h) \approx y(t) + h\left(\frac{1}{4}f(t,y(t)) + \frac{3}{4}f(t+\frac{2h}{3}, y(t+\frac{2h}{3})\right).$$

Pour obtenir une erreur de consistance d'ordre 4, il suffit de "prédire" la valeur de $y(t+\frac{2h}{3})$ à l'ordre 3, car l'erreur commise sera multipliée par le facteur h devant $f(\cdot, y(t+\frac{2h}{3}))$. Faisons cela avec la méthode du point milieu avec h remplacé par $\frac{2h}{3}$. Cea donne (Heun 1900)

$$u_2 = y_0 + \frac{h}{3}f(t_0, y_0)$$

$$u_3 = y_0 + \frac{2h}{3}f(t_0 + \frac{h}{3}, u_2)$$

$$y_1 = y_0 + h\left(\frac{1}{4}f(t_0,y_0) + \frac{3}{4}f(t_0+\frac{2h}{3}, u_3)\right).$$

Par une preuve similaire à la précédente, on voit que **le schéma de Heun est d'ordre 3**.[9] La figure 1.4 qui suit illustre géométriquement les différents schémas. La figure 1.5 compare les principaux schémas à un pas et illustre l'ordre de convergence en $O(h)$ pour Euler, $O(h^2)$ pour Runge, $O(h^3)$ pour Heun, $O(h^4)$ pour Runge-Kutta 4 (voir alinéa suivant). On a utilisé un pas de temps initial $h = 1$. L'équation différentielle résolue est $y' = y$ dont la solution exacte est $y(t) = y_0 \exp t$. La condition initiale est $y_0 = 1$ et l'intervalle de temps est $[0, 7]$.

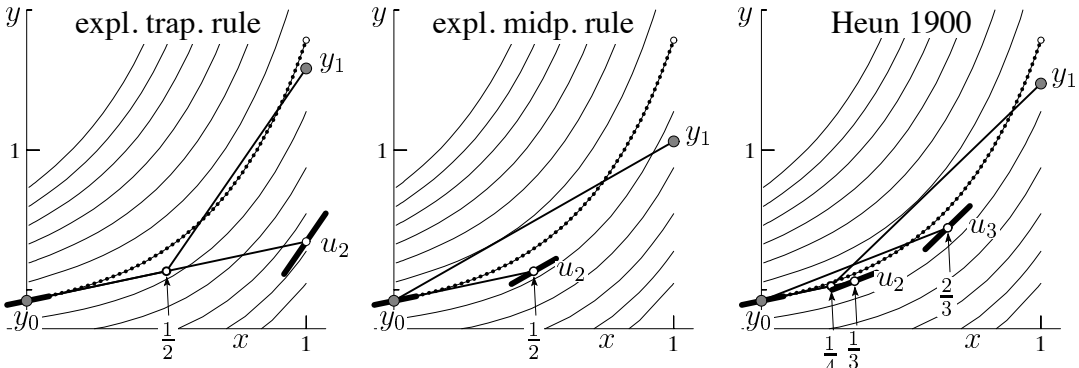


FIG. III.5: Méthodes de Runge-Kutta pour $y' = x^2 + y^2$, $y_0 = 0.46$, $h = 1$; pointillé: solution exacte.

FIGURE 1.4 – Illustration graphique des méthodes de Runge et Heun, d'après Wanner

9. G. Wanner raconte que le premier programme qui a tourné sur le premier ordinateur (aux USA) fut une équation différentielle résolue par la méthode de Heun.

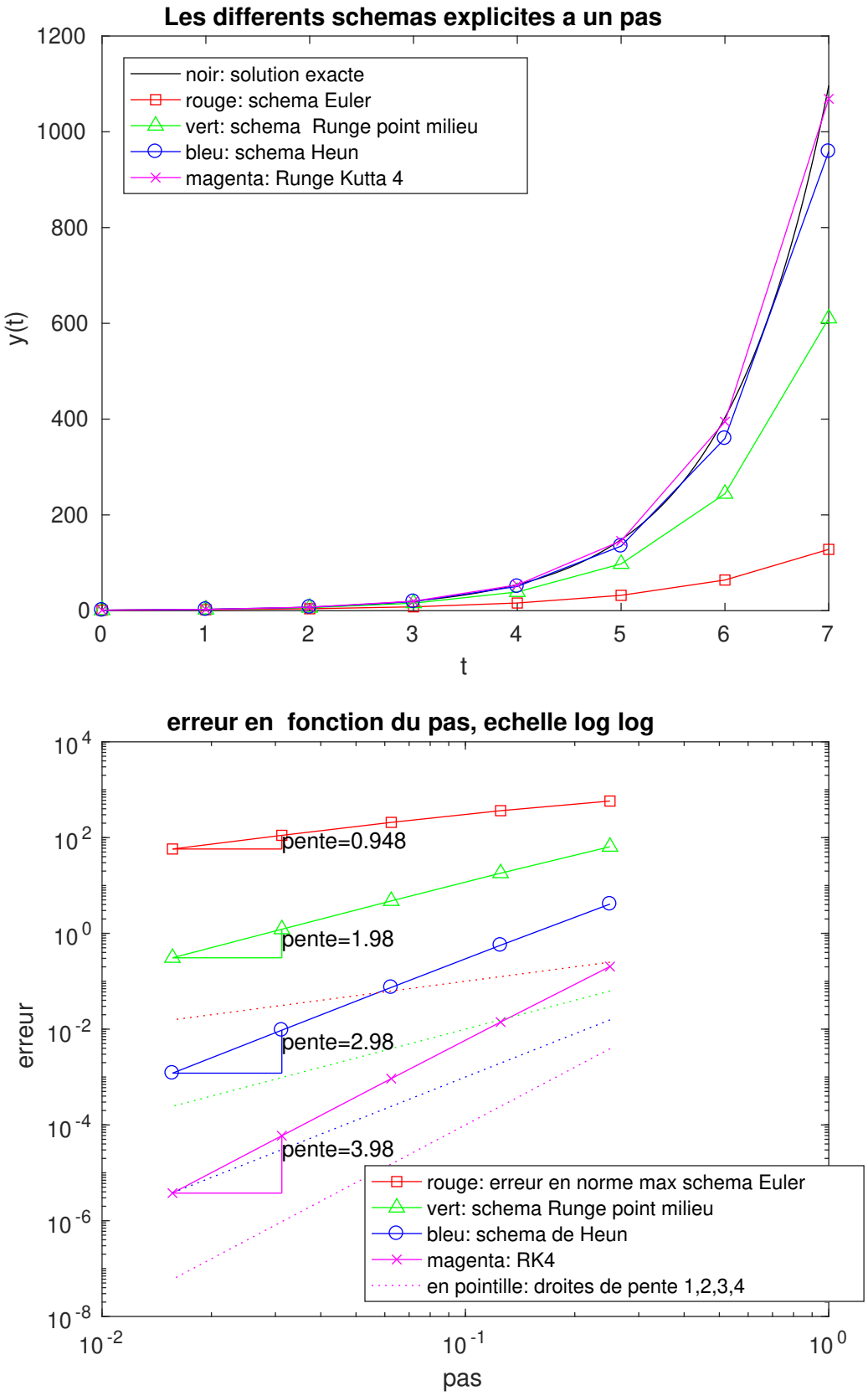


FIGURE 1.5 – Ordre de convergence de divers schémas à un pas.

1.2.1.5 Méthode de Runge-Kutta d'ordre 4.

En utilisant le même principe, on peut approcher l'intégrale

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s))ds \tag{1.29}$$

par la méthode de Simpson (exacte pour les polynômes de degré inférieur ou égal à 3) :

$$\int_t^{t+h} f(s, y(s))ds \approx \frac{h}{6}\left(f(t, y(t)) + 4f(t+h/2, y(t+h/2) + f(t+h, y(t+h))\right).$$

Il faut alors *prédire* $y(t+h/2)$ et $y((t+h)$. La méthode de Runge-Kutta dite RK4 est ainsi décrite par les formules suivantes qui donnent le moyen de calculer $\mathbf{y}_{n+1}$, noté $\mathbf{y}_1$ ici, à partir de $\mathbf{y}_n$ noté $\mathbf{y}_0$:

$$\begin{array}{l}
\mathbf{u}_2 = y_0 + \frac{h}{2}\,\mathbf{f}(t_0, \mathbf{y}_0) \\
\mathbf{u}_3 = y_0 + \frac{h}{2}\,\mathbf{f}(t_0 + \frac{h}{2}, u_2) \\
\mathbf{u}_4 = y_0 + h\,\mathbf{f}(t_0 + \frac{h}{2}, u_3) \\
\mathbf{y}(t_0+h) \approx \mathbf{y}_1 = \mathbf{y_0} + h\left(\frac{1}{6}\mathbf{f}(t_0, u_1) + \frac{2}{6}\mathbf{f}(t_0 + \frac{h}{2}, u_2) + \frac{2}{6}\mathbf{f}(t_0 + \frac{h}{2}, u_3) + \frac{1}{6}\mathbf{f}(t_0+h, u_4)\right)
\end{array}, \tag{1.30}$$

si bien qu'avec la notation de (1.15) et avec $t = t_0$, $\mathbf{y} = \mathbf{y}_0$:

$$\mathbf{\Phi}(t_0, \mathbf{y}_0; h) = \frac{1}{6}\left(\mathbf{f}(t_0, \mathbf{y}_0) + 2\mathbf{f}(t_0 + \frac{h}{2}, u_2) + 2\mathbf{f}(t_0 + \frac{h}{2}, u_3) + \mathbf{f}(t_0+h, u_4)\right) \tag{1.31}$$

C'est un schéma à un pas d'ordre 4 (erreur globale en $O(h^4)$ car on peut vérifier que l'erreur de consistance est en $O(h^5)$). On pourrait faire des calculs, assez fastidieux quand même, qui donnent toutes les formules de ce type d'ordre 2, 3 et 4. Pour plus de détails on peut renvoyer aux livres M. Crouzeix, A.L. Mignot, 1984, Analyse numérique des équations différentielles : Masson et Hairer, S.P. Nørsett, G. Wanner,1993, Solving ordinary differential equations I. Nonstiff problems : Springer-Verlag. Ces estimations d'erreur nécessitent souvent l'usage de la *formule de Taylor* à des ordres assez élevés. Les logiciels de calcul formel sont alors très précieux.

1.2.1.6 ($\star$) Notions sur les méthodes de Runge-Kutta plus générales.

Les méthodes de Runge-Kutta *explicites* à s étages sont de la forme :

$$\begin{array}{l}
\mathbf{k}_1 = \mathbf{f}(t, \mathbf{y}) \\
\mathbf{k}_2 = \mathbf{f}(t + c_2 h, \mathbf{y} + h a_{2,1}\mathbf{k}_1) \\
\mathbf{k}_3 = \mathbf{f}(t + c_3 h, \mathbf{y} + h(a_{3,1}\mathbf{k}_1 + a_{3,2}\mathbf{k}_2)) \\
\cdots \\
\mathbf{k}_s = \mathbf{f}(t + c_s h, \mathbf{y} + h(a_{s,1}\mathbf{k}_1 + a_{s,2}\mathbf{k}_2 + ... + a_{s,s-1}\mathbf{k}_{s-1})) \\
\mathbf{y}(t+h) \approx \mathbf{y}_1 = \mathbf{y} + h(b_1\mathbf{k}_1 + b_2\mathbf{k}_2 + ... + b_s\mathbf{k}_s)
\end{array} \tag{1.32}$$

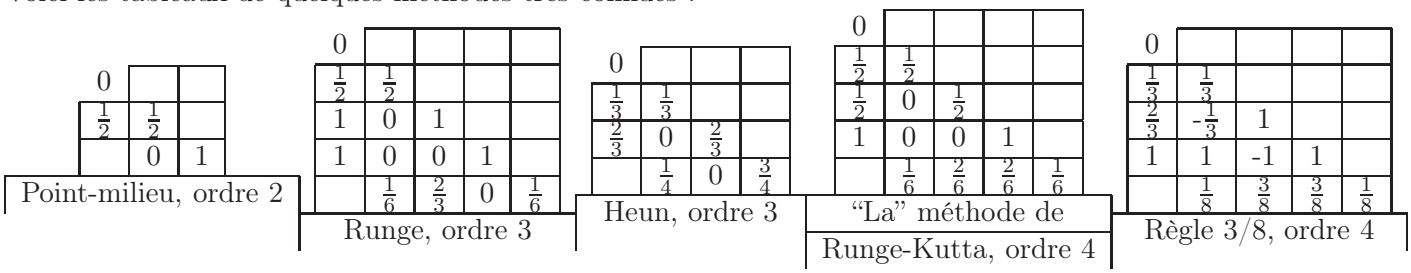
si bien qu'avec la notation de (1.15) et avec $t = t_n$ $\mathbf{y} = \mathbf{y}_n$:

$$\mathbf{\Phi}(t_n, \mathbf{y}_n; h) = b_1\mathbf{k}_1 + b_2\mathbf{k}_2 + ... + b_s\mathbf{k}_s. \tag{1.33}$$

Il s'agit encore d'une méthode à un pas et on vérifie aussi que la fonction $\mathbf{\Phi}(t, \mathbf{y}; h)$ est Lipschitzienne par rapport à la variable $\mathbf{y}$.
Il est d'usage de disposer les coefficients d'une telle formule dans un tableau de la forme :

$$\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{2,1} & & & & \\
c_3 & a_{3,1} & a_{3,2} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s-1} & \\
\hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s
\end{array}$$

Voici les tableaux de quelques méthodes très connues :



Point-milieu, ordre 2 | Runge, ordre 3 | Heun, ordre 3 | "La" méthode de Runge-Kutta, ordre 4 | Règle 3/8, ordre 4

Remarque. On respecte toujours les contraintes : $c_1 = 0$, $c_i = \sum_{j=1}^{s-1} a_{i,j}$ et $1 = \sum_{j=1}^{s} b_j$. Ceci afin d'assurer au minimum $k_i = f(t_0 + c_i h, y(t_0 + c_i h)) + O(h^2)$ et d'intégrer exactement l'équation différentielle $y' = 1$. □

Il existe aussi des méthodes de Runge-Kutta *implicites*. Par exemple le *schéma d'Euler implicite* :

$$\mathbf{y}(t+h) \approx \mathbf{y}_1 = \mathbf{y} + h\,\mathbf{f}(t+h, \mathbf{y}_1).$$

Le terme « implicite » décrit le fait que $\mathbf{y}_1$ n'est pas donné explicitement. En vertu du théorème des fonctions implicites, l'équation $\mathbf{y}_1 - \mathbf{y} - h\,\mathbf{f}(t+h, \mathbf{y}_1) = 0$ définit $\mathbf{y}_1$. En effet considérons la fonction de plusieurs variables $F(\mathbf{y}, h, \mathbf{y}_1) := \mathbf{y}_1 - \mathbf{y} - h\,\mathbf{f}(t+h, \mathbf{y}_1)$. On a $F(\mathbf{y}_0, 0, \mathbf{y}_0) = 0$ et aussi $\frac{\partial F}{\partial \mathbf{y_1}}(\mathbf{y}_0, 0, \mathbf{y}_0) = Id$ qui est inversible donc pour pour h suffisamment petit on peut expliciter $\mathbf{y}_1 = \varphi(h, \mathbf{y}_0)$ comme une fonction de h et $\mathbf{y}_0$.

D'un point de vue pratique, $\mathbf{y}_1$ est un point fixe de l'application

$$\mathbf{y_1} \mapsto g(y_1) := \mathbf{y} + h\,\mathbf{f}(t+h, \mathbf{y}_1)$$

qui est contractante dés que $hL < 1$ ainsi qu'on le vérifie aisément :

$$|g(y_1) - g(z_1)| = h\,|(\mathbf{f}(t+h, \mathbf{y}_1) - \mathbf{f}(t+h, \mathbf{z}_1))| \leq hL\,|y_1 - z_1|$$

De manière analogue *la règle du trapèze implicite* s'écrit

$$\mathbf{y}(t+h) \approx \mathbf{y}_1 = \mathbf{y} + \frac{h}{2}(\mathbf{f}(t, \mathbf{y}) + \mathbf{f}(t+h, \mathbf{y}_1))$$

On peut vérifier facilement que l'application $\mathbf{y_1} \mapsto g(y_1) := \mathbf{y_0} + \frac{h}{2}(\mathbf{f}(t, \mathbf{y_0}) + \mathbf{f}(t+h, \mathbf{y}_1))$ est Lipschitzienne :

$$|g(y_1) - g(z_1)| = \frac{h}{2}\,|(\mathbf{f}(t+h, \mathbf{y}_1) - \mathbf{f}(t+h, \mathbf{z}_1))| \leq \frac{hL}{2}|y_1 - z_1|$$

et contractante dès que $hL < 2$.
On peut donc calculer $\mathbf{y_1}$ par la méthode des approximations successives. En pratique on effectue quelques itérations de point fixe.
La forme générale des schémas implicites est :

$$\mathbf{k}_i = \mathbf{f}(t + c_i h, \mathbf{y} + h \textstyle\sum_{j=1}^{j=s} a_{i,j}\mathbf{k}_j),\ i = 1, \ldots s$$
$$\mathbf{y}(t+h) \approx \mathbf{y}_1 = \mathbf{y} + h \textstyle\sum_{i=1}^{s} b_i \mathbf{k}_i$$

le tableau associé étant

$$\begin{array}{c|cccc} c_1 & a_{1,1} & a_{1,2} & \cdots & a_{1,s} \\ c_2 & a_{2,1} & a_{2,2} & \cdots & a_{2,s} \\ \vdots & \vdots & \vdots & \ddots & \\ c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

(1.34)

Voici quelques tableaux de tels schéma. Ces schémas sont utiles pour les problèmes "raides" et dans ces cas implicites on peut atteindre des ordre $2s$ avec s étapes (par problèmes raides on entend des équations différentielles dont les solutions présentent des variations rapides, voir section 1.2.3).

1 | 1
$\quad$ | 1

Euler implicite

$\frac{1}{2}$ | $\frac{1}{2}$
$\quad$ | 1

Point milieu implicite

0 | 0 | 0
$\frac{2}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$
$\quad$ | $\frac{1}{4}$ | $\frac{3}{4}$

Schéma de Hammer & Hollingsworth

1.2.1.7 (⋆) Notions sur les estimations d'erreurs utilisées dans les codes adaptatifs

Le pas de la subdivision $h_n := t_{n+1} - t_n$ n'est pas forcément *constant* au cours des itérations et peut être adapté en fonction d'une estimation de l'erreur. Dans la mesure où c'est possible, si cette estimation d'erreur est négligeable on augmente le pas, si par contre l'estimation est trop grande on le diminue. Pour cela il est indispensable de disposer d'une estimation de l'erreur locale.

En général pour estimer l'erreur, on utilise *deux* schémas numériques *simultanément*. Un schéma de Runge-Kutta est alors utilisé combiné à un autre, d'ordre plus élevé, afin de permettre une estimation de l'erreur par soustraction des deux valeurs calculées par les schémas, y_{n+1} et $\widehat{y_{n+1}}$. On estime $e_{n+1} := y(t_n + h_n) - y_{n+1} \approx \widehat{y_{n+1}} - y_{n+1}$.[10] On augmente ou diminue alors h_n selon la taille de e_{n+1}. Cela donne des algorithmes adaptatifs où les pas h_n sont automatiquement choisis pour essayer de garantir une précision donnée. On présente ces méthodes avec des tableaux de la forme :

$$
\begin{array}{c|cccccc}
0 & & & & & & \\
c_2 & a_{2,1} & & & & & \\
c_3 & a_{3,1} & a_{3,2} & & & & \\
\vdots & \vdots & \vdots & \ddots & & & \\
c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s-1} & & \\
\hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s & \\
\hline
 & \widehat{b}_1 & \widehat{b}_2 & \cdots & \widehat{b}_{s-1} & \widehat{b}_s & \widehat{b}_{s+1}
\end{array}
$$

où il est entendu que $\mathbf{y}_1 = \mathbf{y} + h\sum_{i=1}^{s} b_i \mathbf{k}_i$ avec $\mathbf{k}_1 = \mathbf{f}(t, \mathbf{y})$ et $\mathbf{k}_i = \mathbf{f}(t + c_i h, \mathbf{y} + + h\sum_{j=1}^{j=i-1} a_{i,j}\mathbf{k}_j)$, $2 \le i \le s$ et l'estimation d'erreur est $\widehat{\mathbf{y}_1} - \mathbf{y}_1$ où $\widehat{\mathbf{y}_1} = \mathbf{y} + h(\sum_{i=1}^{s} \widehat{b}_i \mathbf{k}_i + \widehat{b}_{s+1}\mathbf{f}(t+h, \mathbf{y}_1)$. Il faut citer ici[11] la méthode de Merson qui est d'ordre 4 et dont l'estimation d'erreur est d'ordre 3 en général, la méthode de Bogacki et Schampine qui est d'ordre 3 avec une estimation d'erreur d'ordre 2 et la méthode de Dormand-Prince qui est d'ordre 5 avec une estimation d'erreur d'ordre 4. Ces deux dernières méthodes sont utilisées dans `MATLAB`, respectivement dans les procédure `ode23` et `ode45` ; on peut vérifier les coefficients des tableaux qui suivent avec les commandes `type ode23` et `type ode45`, qui permettent de voir le code source de Matlab.

0					
$\frac{1}{3}$	$\frac{1}{3}$				
$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$			
$\frac{1}{2}$	$\frac{1}{8}$	0	$\frac{3}{8}$		
1	$\frac{1}{2}$	0	$-\frac{3}{2}$	2	
	$\frac{1}{6}$	0	0	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{1}{10}$	0	$\frac{3}{10}$	$\frac{2}{5}$	$\frac{1}{5}$

méthode de Merson "34"

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{3}{4}$	0	$\frac{3}{4}$		
	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{4}{9}$	
	$\frac{7}{24}$	$\frac{6}{24}$	$\frac{8}{24}$	$\frac{3}{24}$

méthode de Bogacki, Shampine "23"

10. On parle de schémas emboîtés (embedded formulas) car les deux schémas ont beaucoup de coefficients en commun ce qui a l'avantage d'économiser le nombre d'évaluations de $\mathbf{f}$.

11. Les références sont :

R.H. Merson, 1957, An operational method for the study of integration processes : Proc. Symp. Data Processing, Weapons Research Establishment, Salisbury, Australia, p 110-1 – 110-25 ;

P. Bogacki, L.F. Shampine, 1989, A 3(2) pair of Runge-Kutta formulas : Applied Mathematics Letters, **2**, 1 – 9 ;

J.R. Dormand, P.J. Prince, 1980, A family of embedded Runge-Kutta formulae : J. Comp. Appl. Math., **6**, 19 – 26.

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{55}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

Méthode de Dormand-Prince “45”

1.2.2 Les méthodes multipas

Pour augmenter la précision du schéma d'Euler, le processus employé dans les méthodes de Runge-Kutta n'est pas le seul possible. On peut aussi, après une période de démarrage, utiliser les valeurs approchées de $\mathbf{y}$, $\mathbf{y}_{n-1}$, $\mathbf{y}_{n-2}$, ..., $\mathbf{y}_{n-k+1}$ aux pas qui précèdent $\mathbf{y}_n$. Les premiers procédés de ce genre qu'on range dans la classe des méthodes multipas [12] sont antérieurs aux méthodes de Runge-Kutta et sont dus à Adams (et publiés par Bashforth [13]).

1.2.2.1 Méthodes d'Adams et de différentiation rétrograde (BDF).

Les méthodes d'Adams sont obtenues en approchant l'intégrale du second membre de $\mathbf{y}(t_{n+1}) = \mathbf{y}(t_n) + \int_{t_n}^{t_{n+1}} \mathbf{f}(t, \mathbf{y}(t))dt$ par **l'intégrale du polynôme d'interpolation** de $t \to \mathbf{f}(t, \mathbf{y}(t))$ aux points t_n, t_{n-1}, ..., t_{n-k+1} (méthodes explicites) dans le cas des *méthodes dites d'Adams-Bashforth* et aux points t_{n+1}, t_n, ..., t_{n-k+1} (méthodes implicites) dans le cas des *méthodes dites d'Adams-Moulton*. On note $\mathbf{f}_n = \mathbf{f}(t_n, \mathbf{y}_n)$ où $\mathbf{y}_n$ est l'approximation ainsi obtenue pour $\mathbf{y}(t_n)$. Dans le cas où le pas $h = t_{i+1} - t_i$ est constant les premières méthodes d'Adams-Bashforth sont avec comme ordre de convergence respectivement $1, 2, 3, 4$ (donc erreurs locales en $O(h^2)$, $O(h^3)$, $O(h^4)$, $O(h^5)$ respectivement) :

$$\begin{array}{l} k=1: \ \mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}_n \ \text{(Euler explicite)}, \\ k=2: \ \mathbf{y}_{n+1} = \mathbf{y}_n + h(\frac{3}{2}\mathbf{f}_n - \frac{1}{2}\mathbf{f}_{n-1}), \\ k=3: \ \mathbf{y}_{n+1} = \mathbf{y}_n + h(\frac{23}{12}\mathbf{f}_n - \frac{16}{12}\mathbf{f}_{n-1} + \frac{5}{12}\mathbf{f}_{n-2}), \\ k=4: \ \mathbf{y}_{n+1} = \mathbf{y}_n + h(\frac{55}{24}\mathbf{f}_n - \frac{59}{24}\mathbf{f}_{n-1} + \frac{37}{24}\mathbf{f}_{n-2} - \frac{9}{24}\mathbf{f}_{n-3}). \end{array} \tag{1.35}$$

Toujours à pas constant, les premières méthodes d'Adams-Moulton sont, avec les mêmes ordres de convergence :

$$\begin{array}{l} k=0: \ \mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}_{n+1} \ \text{(Euler implicite)}, \\ k=1: \ \mathbf{y}_{n+1} = \mathbf{y}_n + h(\frac{1}{2}\mathbf{f}_{n+1} + \frac{1}{2}\mathbf{f}_n) \ \text{(règle du trapèze implicite,)} \\ k=2: \ \mathbf{y}_{n+1} = \mathbf{y}_n + h(\frac{5}{12}\mathbf{f}_{n+1} + \frac{8}{12}\mathbf{f}_n - \frac{1}{12}\mathbf{f}_{n-1}), \\ k=3: \ \mathbf{y}_{n+1} = \mathbf{y}_n + h(\frac{9}{24}\mathbf{f}_{n+1} + \frac{19}{24}\mathbf{f}_n - \frac{5}{24}\mathbf{f}_{n-1} + \frac{1}{24}\mathbf{f}_{n-2}). \end{array} \tag{1.36}$$

À partir de ces méthodes, une *méthode prédicteur-correcteur* est construite de la manière suivante :
P (prédiction) : on utilise une formule de type Adams-Bashforth pour faire une prédiction $\widehat{\mathbf{y}}_{n+1}$ de $\mathbf{y}_{n+1}$;
E (évaluation) : on évalue la fonction $\mathbf{f}$ avec cette approximation $\widehat{\mathbf{f}}_{n+1} = \mathbf{f}(t_{n+1}, \widehat{\mathbf{y}}_{n+1})$;
C (correction) : on porte cette approximation dans une formule d'Adams-Moulton ce qui donne $\mathbf{y}_{n+1}$;
E (évaluation) : pour continuer on évalue $\mathbf{f}_{n+1} = \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})$.
Cela s'appelle un schéma PECE. C'est la procédure la plus courante mais il existe aussi des schémas PECECE, des schémas PEC ...

Remarque. Les coefficients du second membre présentent la particularité de sommer à un. En effet on avance de h entre t_n et t_{n+1}. En particulier, les solutions de l'équation différentielle triviale $y' = 1$ sont données par $y_{n+1} = y_n + h$ qui correspond bien à $y(t) = t + \text{const}$. □

12. On dit aussi méthodes à pas liés.

13. La référence est F. Bashforth, 1883, An attempt to test the theories of capillary action by comparing the theoretical and measured forms of drops of fluid. With an explanation of the method of integration employed in constructing the tables which give the theoretical form of such drops, by C. Adams : Cambridge Univ. Press.

Les *méthodes de Nyström* (explicites) et *de Milne-Simpson* (implicites) sont construites de la même manière mais à partir de $\mathbf{y}(t_{n+1}) = \mathbf{y}(t_{n-1}) + \int_{t_{n-1}}^{t_{n+1}} \mathbf{f}(t, \mathbf{y}(t))dt$. Toujours à pas h constant, les premières méthodes de Nyström sont ($k = 2$ est identique à $k = 1$) :

$$\begin{array}{l} k = 1 : \ \mathbf{y}_{n+1} = \mathbf{y}_{n-1} + 2h\mathbf{f}_n \text{ (schéma saute-mouton)}, \\ k = 3 : \ \mathbf{y}_{n+1} = \mathbf{y}_{n-1} + h(\frac{7}{3}\mathbf{f}_n - \frac{2}{3}\mathbf{f}_{n-1} + \frac{1}{3}\mathbf{f}_{n-2}), \end{array} \tag{1.37}$$

et les premières méthodes de Milne-Simpson sont :

$$\begin{array}{l} k = 0 : \ \mathbf{y}_{n+1} = \mathbf{y}_{n-1} + 2h\mathbf{f}_{n+1}, \\ k = 1 : \ \mathbf{y}_{n+1} = \mathbf{y}_{n-1} + 2h\mathbf{f}_n, \\ k = 2 : \ \mathbf{y}_{n+1} = \mathbf{y}_{n-1} + h(\frac{1}{3}\mathbf{f}_{n+1} + \frac{4}{3}\mathbf{f}_n + \frac{1}{3}\mathbf{f}_{n-1}), \\ k = 4 : \ \mathbf{y}_{n+1} = \mathbf{y}_{n-1} + h(\frac{29}{90}\mathbf{f}_{n+1} + \frac{124}{90}\mathbf{f}_n + \frac{24}{90}\mathbf{f}_{n-1} + \frac{4}{90}\mathbf{f}_{n-2} - \frac{1}{90}\mathbf{f}_{n-3}. \end{array} \tag{1.38}$$

Dans ces cas aussi on peut définir des schémas prédicteurs-correcteurs et bien sûr on peut étendre à d'autres intégrales que $\int_{t_{n-1}}^{t_{n+1}} \mathbf{f}(t, \mathbf{y}(t))dt$.

Remarque. Les coefficients du second membre présentent la particularité de sommer à 2. En effet on avance de $2h$ entre t_{n-1} et t_{n+1}. En particulier, les solutions de l'équation différentielle triviale $y' = 1$ sont données par $y_{n+1} = y_{n-1} + 2h$ qui correspond bien à $y(t) = t + \text{const}$. □

Les *méthodes de différentation rétrograde* appelées en anglais BDF (backward differentiation formula) sont construites d'une autre manière. **On dérive le polynôme d'interpolation q** de $t \to \mathbf{y}(t)$ aux points t_{n+1} (en ce point $\mathbf{y}_{n+1}$ n'est pas (encore) connu), t_n, ..., t_{n-k+1} et on écrit : $\mathbf{q}'(t_{n+1}) = \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})$.

On obtient des *formules implicites* dont les 6 premières sont stables et les autres sont *instables* (voir section suivante), ce que nous admettons aussi. Ces 6 formules implicites de différentiation rétrograde dans le cas de pas constant sont :

$$\begin{array}{l} k = 1 : \ \mathbf{y}_{n+1} - \mathbf{y}_n = h\mathbf{f}_{n+1}, \ (\text{Euler implicite}) \\ k = 2 : \ \frac{3}{2}\mathbf{y}_{n+1} - 2\mathbf{y}_n + \frac{1}{2}\mathbf{y}_{n-1} = h\mathbf{f}_{n+1}, (\text{BDF2}) \\ k = 3 : \ \frac{11}{6}\mathbf{y}_{n+1} - 3\mathbf{y}_n + \frac{3}{2}\mathbf{y}_{n-1} - \frac{1}{3}\mathbf{y}_{n-2} = h\mathbf{f}_{n+1}, \\ k = 4 : \ \frac{25}{12}\mathbf{y}_{n+1} - 4\mathbf{y}_n + 3\mathbf{y}_{n-1} - \frac{4}{3}\mathbf{y}_{n-2} + \frac{1}{4}\mathbf{y}_{n-3} = h\mathbf{f}_{n+1}, \\ k = 5 : \ \frac{137}{60}\mathbf{y}_{n+1} - 5\mathbf{y}_n + 5\mathbf{y}_{n-1} - \frac{10}{3}\mathbf{y}_{n-2} + \frac{5}{4}\mathbf{y}_{n-3} - \frac{1}{5}\mathbf{y}_{n-4} = h\mathbf{f}_{n+1}, \\ k = 6 : \ \frac{147}{60}\mathbf{y}_{n+1} - 6\mathbf{y}_n + \frac{15}{2}\mathbf{y}_{n-1} - \frac{20}{3}\mathbf{y}_{n-2} + \frac{15}{4}\mathbf{y}_{n-3} - \frac{6}{5}\mathbf{y}_{n-4} + \frac{1}{6}\mathbf{y}_{n-4} = h\mathbf{f}_{n+1}. \end{array} \tag{1.39}$$

Remarque. Les coefficients du premier membre présentent la particularité de sommer à zéro. En effet pour l'équation différentielle triviale $y' = 0$, les constantes doivent être solution du schéma. □

1.2.2.2 Un exemple d'instabilité

Contrairement aux méthodes à un pas, dans ce cas des méthodes multipas, il ne suffit pas qu'une méthode soit consistante, c'est à dire d'erreur locale tendant vers zéro avec le pas (plus vite que $O(h)$), pour qu'elle soit convergente. Voici une formule [14] qui est d'ordre 3 au moins mais diverge (le pas étant constant, $t_n = t_0 + nh$, on note $\mathbf{y}_n$ l'approximation de $\mathbf{y}(t_n)$, et $\mathbf{f}_n = \mathbf{f}(t_n, \mathbf{y}_n)$) :

$$\mathbf{y}_{n+1} + 4\mathbf{y}_n - 5\mathbf{y}_{n-1} = h(4\mathbf{f}_n + 2\mathbf{f}_{n-1}). \tag{1.40}$$

Montrons que l'erreur de consistance est en $O(h^4)$.
Pour cela, portons la solution exacte dans le schéma et examinons la différence

$$\epsilon(h_n) := \mathbf{y}(t_{n+1}) + 4\mathbf{y}(t_n) - 5\mathbf{y}(t_{n-1}) - h\left(4\mathbf{f}(t_n, \mathbf{y}(t_n)) + 2\mathbf{f}(t_{n-1}, \mathbf{y}(t_{n-1}))\right). \tag{1.41}$$

14. On l'obtient en cherchant une formule explicite à 3 pas d'ordre maximum.

En effectuant un développement de Taylor au point t_n et en utilisant que $y' = f(t, y)$ on obtient :

$$\epsilon(h) = \mathbf{y}(t_n) + hy'(t_n) + \frac{h^2}{2}\, y''(t_n) + \frac{h^3}{6}\, y'''(t_n) \tag{1.42}$$

$$+4\mathbf{y}(t_n) - 5\left(\mathbf{y}(t_n) - hy'(t_n) + \, y''(t_n) - \frac{h^3}{6}\, y'''(t_n)\right) \tag{1.43}$$

$$-h\left(4y'(t_n) + 2y'(t_{n-1})\right) + \mathcal{O}(h^4) \tag{1.44}$$

$$\epsilon(h) = \mathbf{y}(t_n) + hy'(t_n) + \frac{h^2}{2}\, y''(t_n) + \frac{h^3}{6}\, y'''(t_n) \tag{1.45}$$

$$+4\mathbf{y}(t_n) - 5\left(\mathbf{y}(t_n) - hy'(t_n) + \frac{h^2}{2}\, y''(t_n) - \frac{h^3}{6}\, y'''(t_n)\right) \tag{1.46}$$

$$-4hy'(t_n) - 2h\left(y'(t_n) - hy''(t_n) + \frac{h^2}{2}\, y'''(t_n)\right) + \mathcal{O}(h^4) \tag{1.47}$$

Après simplification il reste $\epsilon(h) = \mathcal{O}(h^4)$. Donc l'erreur de consistance (locale) est $\mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1} = O(h^4)$ et la méthode serait au moins d'ordre 3 (si elle était stable, l'erreur globale serait en $O(h^3)$). Mais cette méthode n'est pas stable ainsi qu'on va le voir plus bas. Voici en effet ce qui se passe lorsqu'on applique la méthode à l'équation différentielle triviale $y'(t) = 0$, $y(0) = 1$, dont la solution exacte est $y(t) = 1$, avec un pas h constant. En supposant que les valeurs de démarrage sont exactes, les relations à satisfaire sont :

$$y_0 = 1,\ y_1 = 1,\ y_{n+1} + 4y_n - 5y_{n-1} = 0 \text{ pour } n \geq 1 \tag{1.48}$$

Or la solution générale de la relation de récurrence $y_{n+1} + 4y_n - 5y_{n-1} = 0$ pour $n \geq 1$ est $y_n = \alpha\lambda_1^n + \beta\lambda_2^n$ où $\lambda_1 = 1$ et $\lambda_2 = -5$ sont les racines de l'équation caractéristique $\zeta^2 + 4\zeta - 5 = 0$. Pour trouver la solution de valeurs initiales y_0 et y_1 il suffit de résoudre le système linéaire 2×2 en α, β $\begin{cases} \alpha + \beta = y_0 \\ \lambda_1\alpha + \lambda_2\beta = y_1 \end{cases}$ qui a une solution unique le déterminant $\lambda_2 - \lambda_1 = -6$ étant toujours non nul. On trouve immédiatement $\alpha = 1$ et $\beta = 0$, ce qui donne

$$y_n = \lambda_1^n = 1 \tag{1.49}$$

Mais si on commet une petite erreur d'arrondi en prenant par exemple $y_0 = 1$ et $y_1 = 1 + \epsilon$ on obtient alors

$$y_n = \alpha\lambda_1^n + \beta\lambda_2^n = (1 + \frac{\epsilon}{6}) - \frac{\epsilon}{6}\,(-5)^n \tag{1.50}$$

Le premier terme est très voisin de 1 lorsque $\epsilon \ll 1$, ce qui est favorable, mais le second est non borné quand $n \to \infty$ et oscille de plus en plus violemment à mesure que n augmente. On dit que le schéma est instable. Ainsi il n'y a pas convergence vers la solution $y = 1$ et cela provient du fait que le polynôme $\rho(\zeta) = \zeta^2 + 4\zeta - 5$ a pour racine $\zeta = 1$ et $\zeta = -5$: *cette deuxième racine de module* > 1 *propage les "petites erreurs de consistance" commises à chaque pas de manière explosive.* On pourrait objecter que si on prend exactement $y_0 = 1$ et $y_1 = 1$ il n'y pas pas de problème. Cependant si on change la condition initiale en $y_0 = 0.1$ et $y_1 = 0.1$, on a vu dans le chapitre 1 que $0.1 = 1/10$ n'admet pas d'écriture binaire finie donc on ne peut assurer *exactement* $y_0 = 0.1$ et $y_1 = 0.1$ On constate numériquement que y_n diverge violemment au bout de quelques dizaines d'itérations en codant

```
x=0.1;
y=0.1;
z=zeros(30,1);
for i=1:30
    z(i)=-4*y+5*x
    x=y;
    y=z(i);
end
```

par exemple $y_{28} \approx -700$ et $y_{29} \approx 3450$! Ce schéma est donc inutilisable en pratique sur une équation non triviale, car il est impossible d'assurer une précision infinie et les erreurs sont amplifiées de manière exponentielle par le facteur λ_2^n.

1.2.2.3 Notions sur le résultat général

On considère une méthode multipas, à pas constant h, de la forme :

$$\alpha_k \mathbf{y}_{n+k} + \alpha_{k-1}\mathbf{y}_{n+k-1} + ... + \alpha_0 \mathbf{y}_n = h(\beta_k \mathbf{f}_{n+k} + ... + \beta_0 \mathbf{f}_n) \tag{1.51}$$

où $\alpha_k \neq 0, |\alpha_0| + |\beta_0| > 0$.
La méthode est *explicite* si $\beta_k = 0$ et *implicite* si non.

Définition 2 *Le schéma est dit consistant si l'erreur de consistance* $\epsilon(h)$

$$\epsilon(h) := \sum_{i=0}^{k} (\alpha_i \mathbf{y}(t+ih) - h\beta_i \mathbf{f}(t+ih, \mathbf{y}(t+ih)) \tag{1.52}$$

est $o(h)$ *quand* $h \to 0$. *On dira que ce* schéma est d'ordre p *si l'erreur de consistance* $\epsilon(h)$ *est en* $O(h^{p+1})$ *pour toute fonction* $t \to \mathbf{y}(t)$ *suffisamment régulière.*

On associe au schéma (1.51) les polynômes :

$$\begin{array}{l} \rho(\zeta) = \alpha_k \zeta^k + \alpha_{k-1}\zeta^{k-1} + ... + \alpha_0, \\ \sigma(\zeta) = \beta_k \zeta^k + \beta_{k-1}\zeta^{k-1} + ... + \beta_0. \end{array} \tag{1.53}$$

Proposition 2 *un schéma est consistant si et seulement si*

$$\rho(1) = 0,\ \rho'(1) = \sigma(1). \tag{1.54}$$

Preuve. Il suffit de remplacer $\mathbf{y}(t+ih)$ et $\mathbf{y}'(t+ih) = \mathbf{f}(t+ih, \mathbf{y}(t+ih))$ par leurs développement de Taylor dans (1.52)) :

$$\begin{aligned} \epsilon(h) &= \sum_{i=0}^{k} (\alpha_i \mathbf{y}(t+ih) - h\beta_i \mathbf{f}(t+ih, \mathbf{y}(t+ih)) && (1.55) \\ &= \sum_{i=0}^{k} \alpha_i \mathbf{y}(t+ih) - h\beta_i \mathbf{y}'(t+ih) && (1.56) \\ &= \sum_{i=0}^{k} \alpha_i \left(\mathbf{y}(t) + ih\, \mathbf{y}'(t)\right) - h\beta_i \mathbf{y}'(t) + o(h) && (1.57) \\ &= \left(\sum_{i=0}^{k} \alpha_i\right) \mathbf{y}(t) + h\left(\sum_i (i\alpha_i - \beta_i)\right) \mathbf{y}'(t) + o(h) && (1.58) \end{aligned}$$

On en déduit que le schéma est consistant si et seulement si $\sum_i \alpha_i = 0$ et $\sum_i i\alpha_i = \sum_i \beta_i$ ce qui donne bien $\rho(1) = 0,\ \rho'(1) = \sigma(1)$. ■

Remarque. La condition $\rho(1) = 0$ revient à vérifier que $y = Cte$ est solution du schéma lorsque $f \equiv 0$. La condition $\sum_i i\alpha_i = \sum_i \beta_i$ revient à imposer en plus que $y(t) = t$ est solution du schéma lorsque $f \equiv 1$. □

Définition 3 *On dit que le* schéma *(1.51)* est stable *si la solution générale de la relation de récurrence linéaire :*

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + ... + \alpha_0 y_n = 0, \tag{1.59}$$

est bornée en fonction des données initiales i.e s'il existe une constante $C > 0$ *telle qu'étant données des valeurs de "démarrage"* y_0, y_1, ..., y_{k-1} *quelconques la suite* $\{y_n\}_{n\in\mathbb{N}}$ *solution de (1.59) vérifie* $|y_n| \leq C \max_{0\leq i\leq k-1} |y_i|$ *que soit* $n \in \mathbb{N}$.

Remarque. Cela impose en particulier que les solutions du schéma appliqué à l'équation $y' = 0$ restent bornées. Ce qui est bien la moindre des choses à demander. □

On démontre la proposition

Proposition 3 *Le schéma est stable si et seulement si*

$$\text{le polynôme } \rho(\zeta) \text{ a toutes ses racines de module } \leq 1, \text{ les racines de modules 1 étant simples.} \tag{1.60}$$

Preuve. Montrons d'abord que la condition (1.60) est nécessaire. C'est un résultat d'algèbre linéaire sur les suites récurrentes linéaires que la solution générale (y_n) de

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + ... + \alpha_0 y_n = 0, \tag{1.61}$$

est une combinaison linéaire de ζ^n si ζ est racine simple de $\rho(\zeta) = 0$, de $\zeta^n, n\zeta^n$ si ζ est racine double de $\rho(\zeta) = 0$, de $\zeta^n, n\zeta^n, \dots n^{l-1}\zeta^n$ si ζ est racine de multiplicité l de $\rho(\zeta) = 0$. Cela signifie que

$$y_n = p_1(n)\zeta_1^n + p_2(n)\zeta_2^n + ... + p_l(n)\zeta_l^n, \tag{1.62}$$

où ζ_1, ζ_2, ..., ζ_l sont les racines distinctes de ρ, la multiplicité de ζ_j étant m_j et où p_1, p_2, ..., p_l sont des polynômes, le degré de p_j étant au plus $m_j - 1$. En particulier pour que les solutions y_n restent bornées $\forall n$ il faut que d'une part toutes les racines ζ_j soient de module inférieur ou égal à un et qu'aucune racine multiple ne soit de module un.

Montrons maintenant que la condition (1.60) est suffisante. Sans perte de généralité, quitte à diviser tous les coefficients par α_k on peut supposer que $\alpha_k = 1$. Nous supposons aussi pour simplifier que $y(t)$ est scalaire. On introduit alors le vecteur de $\mathbb{R}^k$ $Y_n := \begin{pmatrix} y_n \\ y_{n+1} \\ \vdots \\ y_{n+k-1} \end{pmatrix}$ et la matrice $k \times k$

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & & & 0 & 1 \\ -\alpha_0 & \cdots & \cdots & -\alpha_{k-2} & -\alpha_{k-1} \end{pmatrix} \tag{1.63}$$

La relation de récurrence linéaire

$$y_{n+k} = -\alpha_0 y_n - \ldots - \alpha_{k-1} y_{n+k-1}$$

se traduit alors ainsi : $\begin{pmatrix} y_{n+1} \\ y_{n+2} \\ \vdots \\ y_{n+k} \end{pmatrix} = \mathbf{A} \begin{pmatrix} y_n \\ y_{n+1} \\ \vdots \\ y_{n+k-1} \end{pmatrix}$ c'est à dire $Y_{n+1} = AY_n$. Ainsi

$$Y_n = A^n Y_0. \tag{1.64}$$

Il suffit alord de *choisir* une norme sur $\mathbb{R}^k$ telle que $\|\mathbf{A}\| \leq 1$. Pour cela montrons le lemme d'algèbre linéaire suivant.

Lemme 2 *Soit A une matrice $k \times k$ dont toutes les valeurs propres sont de module inférieur ou égal à un et dont les éventuelles valeurs propres multiples sont de module strictement inférieur à un. Il existe une norme sur $\mathbb{R}^k$ telle que la norme matricielle subordonnée vérifie $\|A\| \leq 1$.*

Preuve du lemme. Effectuons la preuve pour $k = 3$ pour alléger. Si A est diagonalisable, alors elle peut s'écrire $A = P \begin{pmatrix} \zeta_1 & 0 & 0 \\ 0 & \zeta_2 & 0 \\ 0 & 0 & \zeta_3 \end{pmatrix} P^{-1}$. Sinon $\zeta_2 = \zeta_3$ est une racine double du polynôme caractéristique et on peut mettre A sous forme réduite de Jordan. $A = P \begin{pmatrix} \zeta_1 & 0 & 0 \\ 0 & \zeta_2 & 1 \\ 0 & 0 & \zeta_2 \end{pmatrix} P^{-1}$ ou bien $\zeta_1 = \zeta_2 = \zeta_3$ est racine triple et la forme de Jordan est $A = P \begin{pmatrix} \zeta_1 & 1 & 0 \\ 0 & \zeta_1 & 1 \\ 0 & 0 & \zeta_1 \end{pmatrix} P^{-1}$.

Dans le cas d'une racine double, par hypothèse $|\zeta_2| < 1$. Quitte à multiplier la 3-ième colonne de P par le facteur $1-|\zeta_2|$, on peut mettre A sous la forme de Jordan modifiée : $A = P \begin{pmatrix} \zeta_1 & 0 & 0 \\ 0 & \zeta_2 & 1-|\zeta_2| \\ 0 & 0 & \zeta_2 \end{pmatrix} P^{-1}$. Dans le cas d'une racine tripe, par hypothèse $|\zeta_1| < 1$. Quitte à multiplier la 2-ème colonne de P par le facteur $1-|\zeta_1|$ et la 3-ème colonne par le facteur $(1-|\zeta_1|)^2$, on peut mettre A sous la forme de Jordan modifiée : $A = P \begin{pmatrix} \zeta_1 & 1-|\zeta_1| & 0 \\ 0 & \zeta_1 & 1-|\zeta_1| \\ 0 & 0 & \zeta_1 \end{pmatrix} P^{-1}$.

Dans tous les cas on peut donc écrire : $A = PJP^{-1}$ avec $J = \begin{pmatrix} \zeta_1 & 0 & 0 \\ 0 & \zeta_2 & 0 \\ 0 & 0 & \zeta_3 \end{pmatrix}$ ou $J = \begin{pmatrix} \zeta_1 & 0 & 0 \\ 0 & \zeta_2 & 1-|\zeta_2| \\ 0 & 0 & \zeta_2 \end{pmatrix}$ ou $J = \begin{pmatrix} \zeta_1 & 1-|\zeta_1| & 0 \\ 0 & \zeta_1 & 1-|\zeta_1| \\ 0 & 0 & \zeta_1 \end{pmatrix}$

Choisissons maintenant la norme suivante sur $\mathbb{R}^k$: $\|x\| := \|P^{-1}x\|_\infty$. Majorons

$$\|Ax\| = \|P^{-1}Ax\|_\infty = \|JP^{-1}x\|_\infty \leq \|J\|_\infty \|P^{-1}x\|_\infty = \|J\|_\infty \|x\|. \tag{1.65}$$

Or on sait d'après le cours d'analyse numérique matricielle que la norme ∞ d'une matrice M se calcule en sommant les modules des coefficients en ligne $\|M\|_\infty = \max_i \sum_j |m_{i,j}|$ donc $\|J\|_\infty = \max_j |\zeta_j|$ dans le cas diagonalisable et $\|J\|_\infty = 1$ sinon et avec (1.65) on obtient $\|Ax\| \leq \|x\|$ donc $\|A\| \leq 1$. □

Considérons la matrice (où on a pris $k = 3$ pour alléger l'écriture)

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\alpha_0 & -\alpha_1 & -\alpha_2 \end{pmatrix}$$

Le polynôme caractéristique de A se calcule en développant suivant la dernière ligne :

$$\begin{vmatrix} -\zeta & 1 & 0 \\ 0 & -\zeta & 1 \\ -\alpha_0 & -\alpha_1 & -\alpha_2 - \zeta \end{vmatrix} = -\alpha_0 - \alpha_1\zeta - \alpha_2\zeta^2 - \zeta^3 = -\rho(\zeta).$$

Le polynôme caractéristique de la matrice A est précisément $\rho(\zeta)$.[15] Par hypothèse de stabilité (1.60) les valeurs propres de A satisfont les conditions du lemme 2. Il existe donc une norme matricielle subordonnée à une norme telle que $\|A\| \leq 1$. Ainsi de la relation (1.64) on déduit

$$\|Y_n\| \leq \|A^n\| \|Y_0\| \leq \|A\|^n \|Y_0\| \leq \|Y_0\|.$$

Comme les normes sur $\mathbb{R}^k$ sont équivalentes on déduit qu'il existe une constante $C > 0$ telle que

$$\|Y_n\|_\infty \leq C \, \|Y_0\|_\infty$$

ce qui donne exactement $|y_n| \leq C \max_{0 \leq i \leq k-1} |y_i|$ que soit $n \in \mathbb{N}$. ■

Exemple. Les méthodes d'Adams sont toutes stables. En effet le polynôme $\rho(\zeta) = \zeta^k - \zeta^{k-1} = \zeta^{k-1}(1-\zeta)$.
Exercice. Montrez que les méthodes de Nystrom et Milne-Simpson le sont également.

Précisons maintenant la notion de convergence pour les schémas multipas.

Définition 4 *Soit $T > 0$ une durée* fixée, *une subdivision $t_0 < t_1 < \ldots < t_N = t_0 + T$ de pas constant h. Etant données k valeurs de départ y_{0h}, y_{1h}, $\ldots y_{(k-1)h}$, on dit que le schéma (1.51) est* convergent *si on a $\max_n |y(t_n) - y_n| \to_{h \to 0} 0$ lorsque les k valeurs de départ vérifient $y(t_i) - y_{ih} \to_{h \to 0} 0$, $i = 0 \ldots k-1$.*
De plus, on dit que le schéma est convergent d'ordre p *si l'erreur globale $\max_{t_0 \leq t_n \leq t_0+T} |y(t_n) - y_n| = \mathcal{O}(h^p)$ lorsque les valeurs de départ vérifient $y(t_i) - y_{ih} = \mathcal{O}(h^p)$, $i = 0 \ldots k-1$.*

15. De ce fait, la matrice A est appelée matrice compagnon du polynôme $\rho(\zeta)$.

Le résultat fondamental suivant est dû à Germund Dahlquist (1956). On a l'équivalence :

Théorème 3 *Un schéma multipas est convergent si et seulement si il est stable et consistant. De plus si l'erreur de consistance est $\mathcal{O}(h^{p+1})$, il est convergent d'ordre p.*

Preuve. Montrons que la condition est suffisante. Nous donnons la preuve dans le cas des schémas explicites. Sans perte de généralité, quitte à diviser tous les coefficients par α_k on peut supposer que $\alpha_k = 1$. Nous supposons aussi pour simplifier que $y(t)$ est scalaire. Comme précédemment, on introduit alors le vecteur de $\mathbb{R}^k$ $Y_n := \begin{pmatrix} y_n \\ y_{n+1} \\ \vdots \\ y_{n+k-1} \end{pmatrix}$ et la matrice $k \times k$

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & & & 0 & 1 \\ -\alpha_0 & \cdots & \cdots & -\alpha_{k-2} & -\alpha_{k-1} \end{pmatrix} \tag{1.66}$$

de sorte que le schéma multipas s'écrit alors

$$\mathbf{Y}_{n+1} = \mathbf{A}\mathbf{Y_n} + h\Phi(t_n, Y_n, h) \text{ avec } \Phi(t_n, Y_n, h) = \begin{pmatrix} 0 \\ \vdots \\ \beta_{k-1}f_{n+k-1} + \beta_{k-2}f_{n+k-2} + ... + \beta_0 f_n \end{pmatrix}. \tag{1.67}$$

On procède ensuite de la même manière que dans le cas des méthodes à un pas. Le schéma est consistant donc la solution exacte vérifie

$$y(t_{n+k}) = -\alpha_{k-1}y(t_{n+k-1}) - \ldots - \alpha_0 y(t_n) + h\left(\beta_{k-1}f(t_{n+k-1}, y(t_{n+k-1})) + \ldots \beta_0 f(t_n, y(t_n)) + \varepsilon(h).\right.$$

ou $\varepsilon(h)$ désigne l'erreur de consistance (1.52). Donc les vecteurs associés à la solution exacte $\mathbf{Y}_n^e \equiv \begin{pmatrix} y(t_n) \\ \vdots \\ y(t_{n+k-2}) \\ y(t_{n+k-1}) \end{pmatrix}$ vérifient, pour un schéma d'ordre p :

$$\mathbf{Y}_{n+1}^e = \mathbf{A}\mathbf{Y}_n^e + h\mathbf{\Phi}(t_n, \mathbf{Y}_n^e, h) + \underbrace{\begin{pmatrix} 0 \\ 0 \\ \vdots \\ \varepsilon(h)) \end{pmatrix}}_{\text{erreur de consistance}} \tag{1.68}$$

Puisque f est Lipschitzienne (de constante de Lipschitz L), on montre aisément que $\mathbf{\Phi}$ aussi :
$\|\mathbf{\Phi}(t, \mathbf{Y}, h) - \mathbf{\Phi}(t, \mathbf{Z}, h)\| \leq M\|\mathbf{Y} - \mathbf{Z}\|$ $\forall \mathbf{Y}$ et $\mathbf{Z}$ vecteurs de $\mathbb{R}^k$, pour $h \leq 1$ et $M = k \max_{0 \leq i \leq k-1} |\beta_i| L$.
En soustrayant membre à membre (1.67) à (1.68), on obtient :

$$\mathbf{E}_{n+1} = \mathbf{A}\mathbf{E}_n + h(\mathbf{\Phi}(t, \mathbf{Y}_n^e, h) - \mathbf{\Phi}(t, \mathbf{Y}_n, h)) + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \varepsilon(h) \end{pmatrix}$$

Pour une norme sur $\mathbb{R}^k$ et la norme matricielle subordonnée, on peut majorer l'erreur globale $\|\mathbf{E}_n\| \equiv \|\mathbf{Y}_n^e - \mathbf{Y}_n\|$:

$$\|\mathbf{E}_{n+1}\| \leq (\|\mathbf{A}\| + hM)\|\mathbf{E}_n\| + \|\varepsilon(h)\|. \tag{1.69}$$

Comme dans la preuve de la proposition 3 on voit que le polynôme caractéristique de la matrice A est précisément $\rho(\zeta) = \zeta^k + \alpha_{k-1}\zeta^{k-1} + ... + \alpha_0$. Par hypothèse de consistance 1 est racine du polynôme caractéristique. Par hypothèse de stabilité les valeurs propres de A sont toutes de module inférieur ou égal à un et celles qui sont de module un sont simples, on peut donc appliquer le lemme 2 et *choisir* une norme sur $\mathbb{R}^k$ telle que $\|\mathbf{A}\| = 1$.

On a ainsi démontré que la suite des erreurs globales vérifie

$$\|\mathbf{E}_{n+1}\| \leq (1 + hM)\|\mathbf{E}_n\| + \|\varepsilon(h)\|. \tag{1.70}$$

On procède alors de la même manière que pour les méthodes à un pas en utilisant le lemme 1 où l'on doit tenir compte aussi des erreurs qu'on peut faire sur les valeurs de démarrage mais vu (1.19) (terme θ_0) cela est possible.

Réciproquement montrons que la convergence d'un schéma implique sa stabilité et sa consistance. Commençons par démontrer la condition de stabilité. Pour cela considérons l'équation différentielle particulière : $y' = 0$, avec la condition initiale $y(0) = 0$ dont l'unique solution est la solution $y(t) \equiv 0$. Supposons qu'il existe ζ racine du polynôme ρ telle que $|\zeta| > 1$. On suppose qu'on applique le schéma multipas avec un pas constant $h > 0$. Soient les valeurs de démarrage $y_0 = h, y_1 = h\zeta, \dots y_{k-1} = h\zeta^{k-1}$. La suite $y_n = h\zeta^n$ est alors solution du schéma. Or $y_n = h\zeta^{t/h}$ n'est pas bornée quand $h \to 0$ donc y_n ne converge pas vers 0 bien que $y_0 = h, y_1 = h\zeta, \dots y_{k-1} = h\zeta^{k-1}$ convergent bien vers 0 quand $h \to 0$. Le schéma ne converge donc pas. Cela prouve que toutes les racines de ρ sont de module inférieur ou égal à 1. Supposons maintenant qu'il existe ζ racine multiple du polynôme ρ telle que $|\zeta| = 1$. Soient les valeurs de démarrage $y_0 = 0, y_1 = \sqrt{h}\,\zeta, \dots y_{k-1} = \sqrt{h}\,(k-1)\zeta^{k-1}$. La suite $y_n = \sqrt{h}n\zeta^n$ est alors solution du schéma. Or $y_n = \frac{t}{\sqrt{h}}\zeta^{t/h}$ n'est pas bornée quand $h \to 0$. Le schéma ne converge donc pas. Cela prouve que toutes les racines multiples de ρ sont de module strictement inférieur à 1. La condition de stabilité (1.60) est donc nécessaire à la convergence.

Pour montrer la consistance nous allons considéder successivement deux équations différentielles très simples. Tout d'abord $y' = 0$, $y(0) = 1$ dont la solution exacte est $y(t) \equiv 1$. Le schéma donne la relation de récurrence :

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + ... + \alpha_0 y_n = 0.$$

Le schéma étant convergent, pour chaque n on doit avoir $y_n \to 1$ quand h vers 0. En passant à la limite dans la relation, on obtient

$$\alpha_k + \alpha_{k-1} + ... + \alpha_0 = 0$$

qui traduit précisément $\rho(1) = 0$.
Enfin considérons le problème de Cauchy $y' = 1$, $y(0) = 0$ dont la solution exacte est $y(t) = t$. Choisissons un pas constant h. Le schéma donne la relation de récurrence :

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + ... + \alpha_0 y_n = h(\beta_k + \beta_{k-1} + ... + \beta_0).$$

Prenons la suite $y_j = C\,jh$ et portons la suite dans la relation de récurrence.

$$C(\alpha_k(n+k)h + \alpha_{k-1}(n+k-1) + ... + \alpha_0 nh) = h(\beta_k + \beta_{k-1} + ... + \beta_0) = h\sigma(1)$$

$$Cnh(\alpha_k + \alpha_{k-1} + ... + \alpha_0) + Ch(\sum_{j=0}^{k} j\alpha_j) = h\sigma(1)$$

Nous avons montré que $\rho(1) = 0$ donc il reste en simplifiant par h :

$$C(\sum_{j=0}^{k} j\alpha_j) = C\rho'(1) = \sigma(1).$$

Donc la suite $y_n = C\,nh$ est solution si on prend $C = \frac{\sigma(1)}{\rho'(1)}$ qui est bien définie car 1 est racine simple de ρ comme le schéma est stable. Donc $y_n = Ct_n$ mais on doit avoir convergence de y_n vers $y(t_n) = t_n$ donc $C = 1$ donc $\rho'(1) = \sigma(1)$ et on a bien prouvé la consistance du schéma (1.54). ■

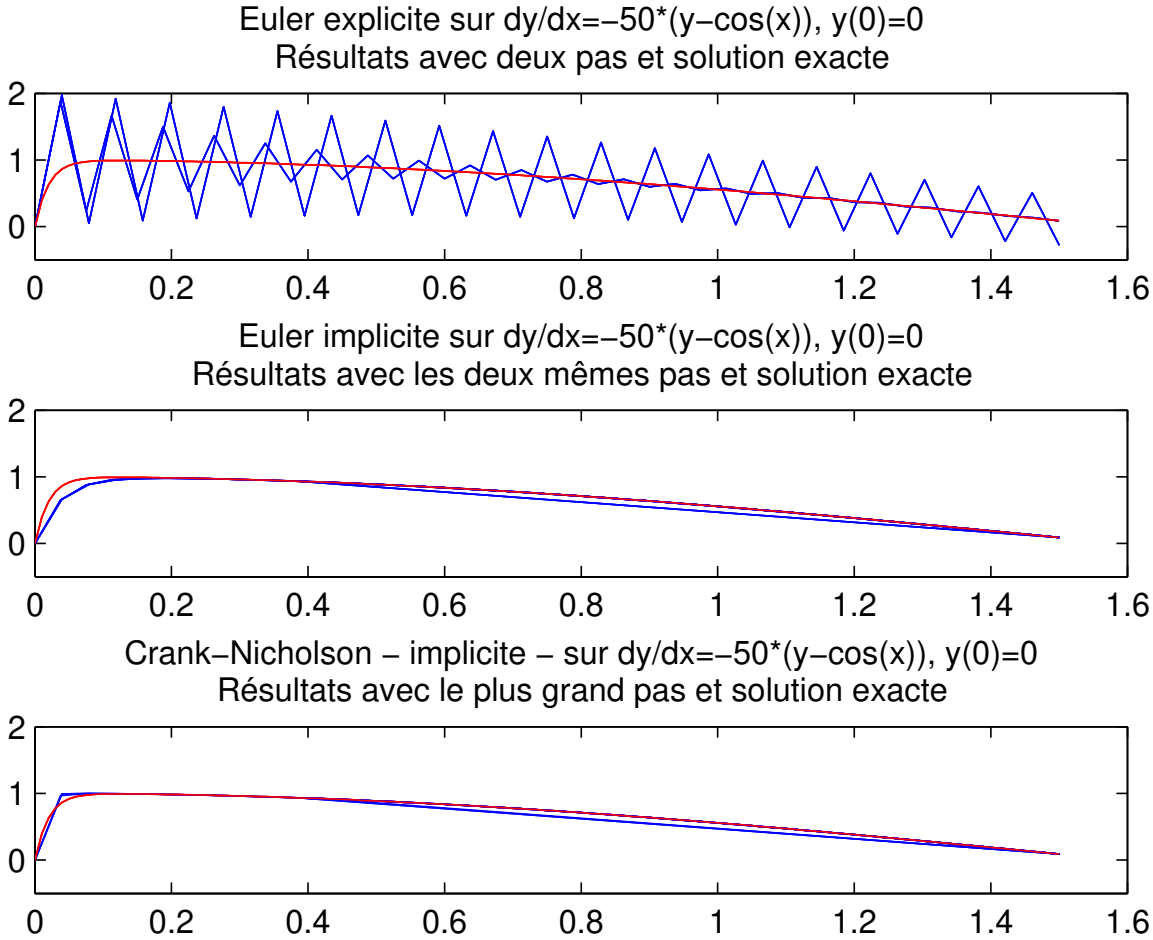
1.2.3 Notions sur les problèmes raides (stiff en anglais)

Les équations différentielles dites « raides » sont celles qui contiennent des second membres $f(t,y)$ qui peuvent varier brusquement en fonction de y. Cela signifie que la constante de Lipschitz L telle que $|f(t,y)-f(t,z)| \leq L|y-z|$ peut être très grande devant l'unité.

On va traiter ici un exemple très élémentaire mais, on l'espère éclairant. Considérons l'équation différentielle $y'(t) = -50(y(t)-\cos(t)) \equiv f(t,y(t))$, $y(0)=0$, où la constante de Lipschitz vaut $50 >> 1$. Dans le code `MATLAB` qui suit, on a appliqué à cette équation différentielle la méthode d'Euler (explicite), la méthode d'Euler implicite et la règle du trapèze implicite ou schéma de Crank-Nicolson $y_{n+1} = y_n + \frac{h_n}{2}(f(t_n,y_n)+f(t_{n+1},y_{n+1}))$.
Au moyen de la méthode de la variation de constante on peut calculer la solution (exacte)

$$y(x) = \frac{2500\cos(x) + 50\sin(x) - 2500e^{-50x}}{2501};$$

il y a une transition assez "raide" de $y(0)=0$ à $y(x) \cong \cos(x)$ entre $x=0$ et $x \lesssim 0.8$. On a obtenu les résultats affichés dans la figure qui suit en prenant des pas de temps $h = 1.974/50$ et $h = 1.875/50$ très voisins de $2/50$.



La méthode d'Euler explicite avec un pas constant $h>0$ sur une équation différentielle $y'(t) = f(t,y(t))$, $y(0)=y_0$ s'écrit :

$$y_{n+1} = y_n + hf(nh, y_n),\ n \geq 0, \tag{1.71}$$

et si l'équation différentielle est $y' = \lambda y$ où λ est une constante (solution en $e^{\lambda t}y_0$) :

$$y_{n+1} = R^{EE}(\lambda h) y_n \text{ donc } y_n = (R^{EE}(\lambda h))^n y_0 \text{ avec } R^{EE}(z) = 1 + z. \tag{1.72}$$

La solution numérique "n'explose pas" seulement si $|R(\lambda h)| \leq 1$ ce qui équivaut à $-1 \leq 1 + \lambda h \leq 1$ donc à, si $\lambda < 0$, $h \leq \frac{2}{|\lambda|}$. Dans l'exemple précédent on conçoit que la valeur de λ est -50, les pas adoptés sont trop proches du seuil $h = \frac{2}{|\lambda|} = 2/50$ et on voit des oscillations parasites.

La méthode d'Euler implicite avec un pas constant $h > 0$ sur une équation différentielle $y'(t) = f(t, y(t))$, $y(0) = y_0$ s'écrit :

$$y_{n+1} = y_n + hf((n+1)h, y_{n+1}),\ n \geq 0, \tag{1.73}$$

et si l'équation différentielle est $y' = \lambda y$:

$$y_{n+1} = R^{EI}(\lambda h) y_n \text{ donc } y_n = (R^{EI}(\lambda h))^n y_0 \text{ avec } R^{EI}(z) = \frac{1}{1-z}, \tag{1.74}$$

Pour $\lambda < 0$ et $h > 0$, $R^{EI}(\lambda h) = \frac{1}{1+|\lambda|h} < 1$ la solution numérique n'explose jamais et on constate de bons résultats ; il n'y a plus d'oscillations parasites. Cependant la transition brusque entre $t = 0$ et $t =\lesssim 0.8$ n'est pas reproduite avec une grande précision.

Avec la règle du trapèze, il n'y a plus d'oscillations parasites et le transition est mieux capturée parce que le schéma est d'ordre 2 :

$$y_{n+1} = R^{CN}(\lambda h) y_n \text{ donc } y_n = (R^{CN}(\lambda h))^n y_0 \text{ avec } R^{CN}(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}}. \tag{1.75}$$

En effet pour $\lambda < 0$ et $h > 0$, $R^{CN}(\lambda h) = \frac{1-|\lambda|h/2}{1+|\lambda|h/2} < 1$.

1.2.3.1 ($\star$) Compléments sur la stabilité.

D'une manière générale l'application d'une méthode à un pas à l'équation différentielle $y' = \lambda y$ se traduit en une récurrence de la forme :

$$y_{n+1} = R(h\lambda) y_n. \tag{1.76}$$

Le cas où $\Re\mathrm{e}(\lambda) < 0$ est le plus discriminant car la solution exacte $y(t) = C^{te} \exp(\lambda t)$ doit tendre vers zéro très rapidement donc, d'un point de vue comportement qualitatif, il est souhaitable que la solution numérique reste au moins bornée. Comme $y_{n+1}k = R(\lambda h)\, y_n$, il faut que $R(\lambda h) \leq 1$, et ce même lorsque le pas h n'est pas forcément assez petit pour que la solution numérique soit une bonne approximation. Cela conduit à définir le *domaine de stabilité du schéma* $S = \{z \in \mathbb{C} : |R(z)| \leq 1\}$. On dit qu'un schéma à un pas est A-stable (absolument stable) si $\mathbb{C}_- \equiv \{z \in \mathbb{C} : \Re\mathrm{e}(z) \leq 0\} \subseteq S$. Ainsi avec un schéma A-stable, pour $\Re\mathrm{e}(\lambda) < 0$ tous les pas h donnent des résultats "corrects". Le schéma d'Euler explicite n'est pas A-stable, par contre les schémas d'Euler implicite et de Crank-Nicholson le sont. Il y a d'autres notions de stabilité comme la L-stabilité qui est la A-stabilité avec en plus $\lim_{z\to\infty} R(z) = 0$. On peut montrer que les schémas de Runge-Kutta explicites ne sont pas A-stables. Par contre il y a des méthodes de Runge-Kutta implicites A-stables et même L-stables et des extensions de ces méthodes qui utilisent la dérivée $\frac{\partial \mathbf{f}}{\partial \mathbf{y}}$, appelées méthodes de type Rosenbrock sont souvent utilisées pour la résolution numérique de ces problèmes raides.

La notion de A-stabilité pour les méthodes multipas $\alpha_k \mathbf{y}_{n+k} + \alpha_{k-1}\mathbf{y}_{n+k-1} + ... + \alpha_0 \mathbf{y}_n = h(\beta_k \mathbf{f}_{n+k} + ... + \beta_0 \mathbf{f}_n)$ s'introduit de la manière suivante. On applique la méthode à une équation différentielle $y' = \lambda y$ et cela conduit à considérer, avec $\mu = \lambda h$, le polynôme $(\alpha_k - \mu\beta_k)\zeta^k + ... + (\alpha_0 - \mu\beta_0) = 0$ dont on note $\zeta_j(\mu)$ les racines. Le domaine de stabilité de la méthode est $S = \{\mu \in \mathbb{C};\ \begin{array}{l}\text{toutes les racines } \zeta_j(\mu) \text{ satisfont } |\zeta_j(\mu)| \leq 1,\\ \text{les racines multiples satisfaisant } |\zeta_j(\mu)| < 1\end{array}\ \}$ et la méthode est dite A-stable si $\mathbb{C}_- \subseteq S$. La seconde barrière de Dahlquist dit qu'une méthode multipas A-stable est forcément d'ordre $p \leq 2$. Enfin, si on définit, pour $0 < \alpha < \frac{\pi}{2}$, la notion de $A(\alpha)$-stabilité par le fait qu'une méthode est dite $A(\alpha)$-stable si $S_\alpha \equiv \{\mu \in \mathbb{C};\ |\arg(-\mu)| < \alpha,\ \mu \neq 0\} \subseteq S$, alors on peut établir que les méthodes de différentiation rétrograde à $k = 1, 2, ..., 6$ étapes sont $A(\alpha)$-stables avec les valeurs de α suivantes (on a aussi $\{\mu \in \mathbb{C};\ \Re\mathrm{e}(\mu) < -D\} \subseteq S$ avec les valeurs de D indiquées dans le tableau) :

k	1	2	3	4	5	6
α	90°	90°	86.03°	73.35°	51.84°	17.84°
D	0	0	0.083	0.667	2.327	6.075

Pratiquement on peut déceler qu'un problème est raide par le fait qu'une méthode numérique "ordinaire" (Runge-Kutta explicite, Adams) adaptative aboutit à de petits pas. Pour toutes ces considérations et des compléments, il faut se reporter à la bibliographie et en particulier au livre E. Hairer, G. Wanner, 1996, Solving ordinary differential equations II : Stiff and Differential-Algebraic problems : Springer-Verlag, Berlin.

1.2.3.2 Epilogue.

Terminons par quelques exemples. Voici les commandes MATLAB pour résoudre le problème proies prédateurs de Lokta Volterra (r représente une population de lapins (rabbits en anglais) et f représente une population de renards (fox en anglais) ; la solution est périodique, la période étant fonction des conditions initiales ; pour plus de détails voir le livre de Moler, exercice 7.15) :

$$\begin{array}{c} \frac{dr}{dt} = 2r - \alpha r f \\ \frac{df}{dt} = -f + \alpha r f \\ r(0) = r_0,\ f(0) = f_0 \end{array} \tag{1.77}$$

les données étant $\alpha = 0.01$, $r_0 = 300$, $f_0 = 150$:

```
f=@(t,y) [2*y(1)-0.01*y(1)*y(2); -y(2)+0.01*y(1)*y(2)]
ode23(f,[0 20],[300 150])
```

La première instruction définit le système différentiel, la deuxième le résout et affiche automatiquement le résultat. Pour conserver le résultat dans une structure afin de le reéchantillonner et par exemple le dessiner d'une autre façon on peut utiliser les instructions :

```
sol=ode23(f,[0 20],[300 150])
xx=0:0.01:20; yy=deval(sol,xx); figure(2), plot(xx,yy(1,:),xx,yy(2,:))
```

Pour plus détails il faut se reporter à la documentation `MATLAB`. Il faut quand même signaler que les "solveurs" proposés pour les problèmes ordinaires (non raides) sont `ode23` (méthode de type Runge Kutta de précision modeste mais rapide), `ode45` (méthode de type Runge Kutta précise) et `ode113` (méthode d'Adams-Bashforth-Moulton PECE d'ordre variable plus précise encore). De plus on peut modifier des options pour effectuer certaines opérations. Par exemple avec la fonction, écrite dans un fichier nommé `evenment.m` :

```
function [valeur,fin,direction] = evenment(t,y)
valeur=y(1)-300; fin=1; direction=1;
```

les instructions :

```
options=odeset('Events',@evenment);
[t,y,te,ye] = ode23(f,[0 20],[300 150],options);
te
ye
```

fournissent la période 4.9981 (`te(2)`) qu'on peut aussi déterminer approximativement sur les graphiques.
Les "solveurs" proposés, adaptés aux problèmes raides sont `ode15s`, `ode23s`, `ode23t` et `ode23tb` et pour montrer l'intérêt de ces codes, on peut proposer les manipulations suivantes, tiré du livre de Moler (paragraphe 7.9), avec une équation qui modélise la combustion d'une allumette :

$$\frac{dy}{dt} = y^2 - y^3,\ 0 \le t \le \frac{2}{\delta},\ y(0) = \delta$$

```
delta=0.0001;
f=@(t,y) y^2-y^3
options=odeset('RelTol',1e-4);
figure(3)
ode45(f,[0 2/delta],delta,options);
pause
delta=0.0001;
figure(4)
```

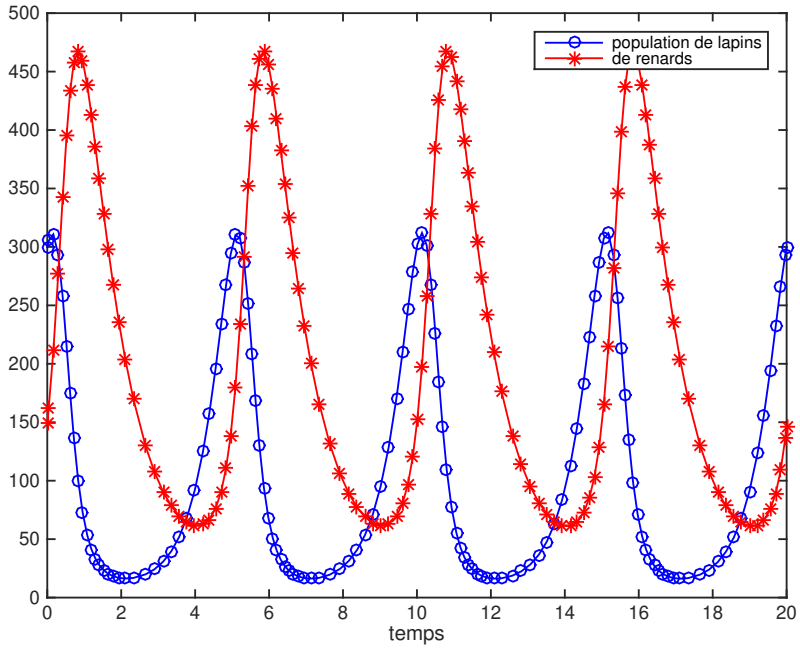


FIGURE 1.6 – Population proie-prédateurs.

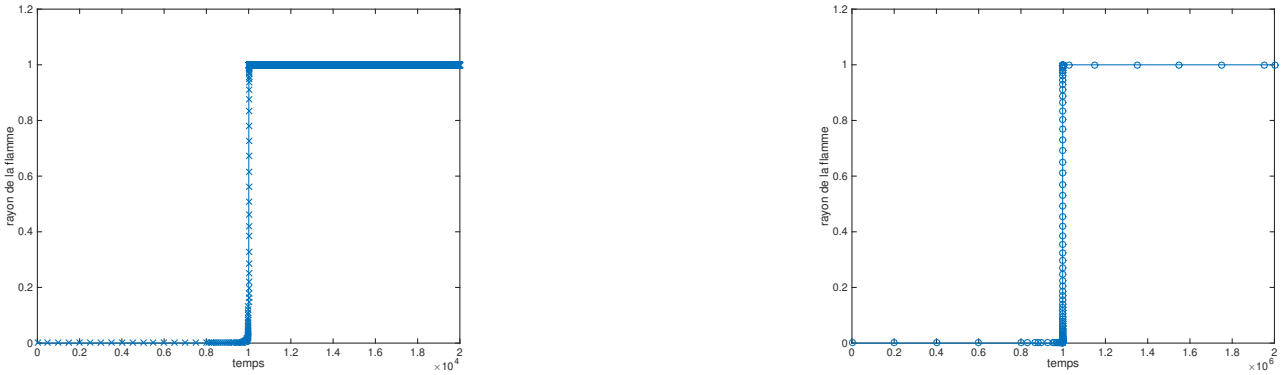


FIGURE 1.7 – Allumage d'une allumette, gauche : schéma standard, droite :un schéma pour équations raides.

`ode23s(f,[0 2/delta],delta,options);` On obtient apparemment les mêmes résultats.

Mais si on regarde de plus près, on constate que la simulation avec le solveur `ode23s` spécialement conçu pour les équations raides nécessite seulement 120 pas de temps pour une durée de 20000 alors que le solveur générique `ode45` a besoin de 12000 pas de temps pour la même simulation ! De plus la solution calculée a tendance à osciller autour de la valeur finale 1 alors qu'elle devrait se stabiliser à $y = 1$. De fait si on "zoome" sur la figure de gauche :

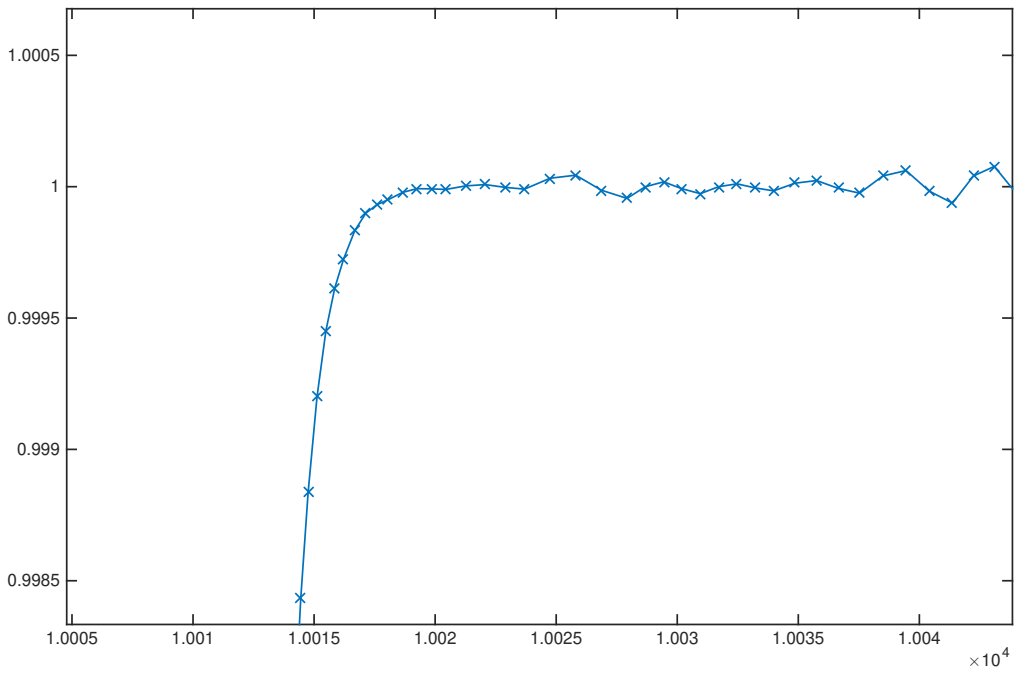


FIGURE 1.8 – Allumage d'une allumette, avec un schéma standard.

Chapitre 2

Equations aux dérivées partielles.

2.1 Introduction.

Il existe une très grande variété d'EDP (équations aux dérivées partielles) : Maxwell, Navier-Stokes, chaleur, KdV, Schrödinger...C'est un domaine très vaste des mathématiques pures et appliquées. Ce cours est une brève introduction élémentaire.
Notations : dans la suite $u : (x,t) \in \mathbb{R}^d \times \mathbb{R} \to u(x,t) \in \mathbb{R}$ désigne une fonction numériques de plusieurs variables x, t, x variable d'espace (pouvant être éventuellement multidimensionnelle), t est la variable de temps. On notera en général :

$$u_t = \frac{\partial u}{\partial t},\ u_x = \frac{\partial u}{\partial x},\ u_{xt} = \frac{\partial^2 u}{\partial x \partial t},\ u_{xx} = \frac{\partial^2 u}{\partial x^2},\ u_{xx} = \frac{\partial^2 u}{\partial x^2},\ u_{xxx} = \frac{\partial^3 u}{\partial x^3}$$

etc...

Définition 5 *Une équation aux dérivées partielles (EDP) en la fonction scalaire $u(x,y,z,t)$ est une équation de la forme*

$$F(t,x,y,z,u,u_x,u_y,u_z,u_t \dots D^m u) = 0 \tag{2.1}$$

où $F := F(t,x,Du,D^2u,\dots,D^mu)$ est une fonction de plusieurs variables.

L'ordre de l'EDP est l'ordre m maximum des dérivées apparaissant dans l'équation. On dit que u est solution (classique) de l'EDP dans un domaine $Q \subset \mathbb{R}^d \times \mathbb{R}^+$ si $u \in \mathcal{C}^m(Q)$ et que u ainsi que ses dérivées partielles satisfont l'équation en tout point de Q.

Les EDP sont utilisées pour modéliser une grande variété de phénomènes physiques, biologiques, ...dans des domaines scientifiques très divers. En général pour déterminer une solution, il faut également donner des conditions initiales $u(\cdot, t=0)$, des conditions limites au bord du domaine spatial. Comme la solution de l'EDP décrit en principe la solution d'un problème concret, on souhaite de plus que des petites erreurs sur les données n'engendrent pas de grosses différences sur la solution u. Etudier une EDP c'est donc trouver les bonnes données initiales et aux limites qui assurent

— *l'existence* de solutions à l'EDP,
— montrer qu'avec ces données *la solution est unique* dans une certaine classe de fonction,
— que la solution dépend continûment des données.

Si ces trois critères sont satisfaits, on dit que le problème est *bien-posé.*

Dans de très rares cas, nous pourrons trouver une expression analytique de u. Parfois, on pourra trouver une expression sous forme intégrale (convolution), ou somme de série de Fourier par exemple.
Le plus souvent il faudra calculer une approximation de u par un *schéma numérique*, comme dans le cas des équations différentielles.

Exemple. l'équation eikonale de l'optique

$$u_x^2 + u_y^2 = 1$$

une équation de transport

$$u_t + a(x,t)\, u_x = 0$$

l'équation des ondes

$$u_{tt} - c^2 u_{xx} = 0$$

L'équation de la chaleur

$$u_t - \nu\, u_{xx} = 0$$

L'équation de Poisson

$$u_{xx} + u_{yy} = f$$

L'équation de Burgers

$$u_t + u\, u_x = 0$$

l'équation de Korteweg de Vries

$$u_t + 6u\, u_x + u_{xxx} = 0$$

Une propriété qui simplifie grandement l'étude est la *linéarité.* On dit que l'edp est linéaire si $u \mapsto F(t, x, y, z, u, u_x, u_y, u_z, u_t \dots D^m u)$ est linéaire. Dans ce cas u et ses dérivées apparaissent seulement à la puissance un et e les coefficients de u et ses dérivées dépendent seulement des variables indépendantes $x, y, \dots, t$.

Exercice. Parmi les exemples précédents quelles sont les EDP linéaires ?

Lorsque l'EDP est linéaire, on peut l'écrire sous la forme $Lu = g$ où L est une fonctionnelle linéaire définie sur un espace de fonction approprié.

2.2 EDP linéaires du premier ordre.

L'EDP linéaire générale du premier ordre en les variables x, y s'écrit

$$a(x,y)u_x + b(x,y)u_y + c(x,y)u = g(x,y). \tag{2.2}$$

L'edp (2.2) peut s'écrire $Lu = g$ avec $Lu = au_x + bu_y + cu$. La linéarité de l'edp permet d'ajouter des solutions : si $Lu = g_1$ et $Lv = g_2$ alors

$$L(\alpha u + \beta v) = \alpha g_1 + \beta g_2$$

pour des scalaires quelconques α et β. C'est le *principe de superposition.*

2.2.1 Cas des coefficients constants.

Commençons par étudier une équation linéaire simple. Soit $c > 0$ une constante.

$$u_t + c\, u_x = 0 \tag{2.3}$$

Remarquons que $u_t + c\, u_x = (c, 1) \cdot \nabla u$ où le gradient $\nabla u = (u_x, u_t)$. Il est alors naturel de considérer les droites d'équation $x = ct + x_0$ sur lesquelles u est constante :

$$\frac{d}{dt}\{u(x_0 + ct, t)\} = c\, u_x + u_t = 0$$

Donc $u(x_0 + ct, t) = u(x_0, 0)$. Nous pouvons alors énoncer le théorème.

Théorème 4 *Soit $f(x)$ une fonction $\mathcal{C}^1$ sur $\mathbb{R}$. Il existe une unique solution $\mathcal{C}^1$ au problème de Cauchy*

$$u_t + c\, u_x = 0, \quad u(x, 0) = f(x). \tag{2.4}$$

Elle est donnée explicitement par la formule $u(x, t) = f(x - ct)$.

Nous allons voir que le problème est bien posé en étudiant la dépendance de la solution par rapport à la donnée initiale f. Soient $f(x)$ et $g(x)$ deux données initiales et $u(x,t)$, $v(x,t)$ les solutions correspondantes. Par *linéarité* de l'EDP, on peut écrire

$$u(x,t) - v(x,t) = f(x-ct) - g(x-ct)$$

et cela implique

$$\max_{x,t} |u(x,t) - v(x,t)| = \max_{x,t} |f(x-ct) - g(x-ct)| = \max_{x} |f(x) - g(x)|$$

$$\|u - v\|_{L^\infty} \leq \|f - g\|_{L^\infty}.$$

On a bien une *dépendance continue* des solutions vis à vis de la donnée initiale (pour la norme infinie ici). Le problème est donc bien posé.

Remarque. quelle que soit l'unité adoptée pour u, si x est une longueur et t un temps, c est homogène à une longueur/temps donc à une vitesse. La grandeur c est appelée célérité ou vitesse de propagation. La solution $u(x,t)$ est une onde ou un signal qui se propage à vitesse c vers la droite lorsque $c > 0$, (resp. vers la gauche si $c < 0$). Les courbes $x - ct = Cte$ sont appelées les *caractéristiques* parce qu'elle portent l'information (la valeur) de u. Voir la figure 2.2.1. La méthode qui consiste à construire la solution en analysant son comportement le long des courbes caractéristiques est appelée *méthode des caractéristiques* et permet d'étudier de nombreuses EDP du premier ordre. □

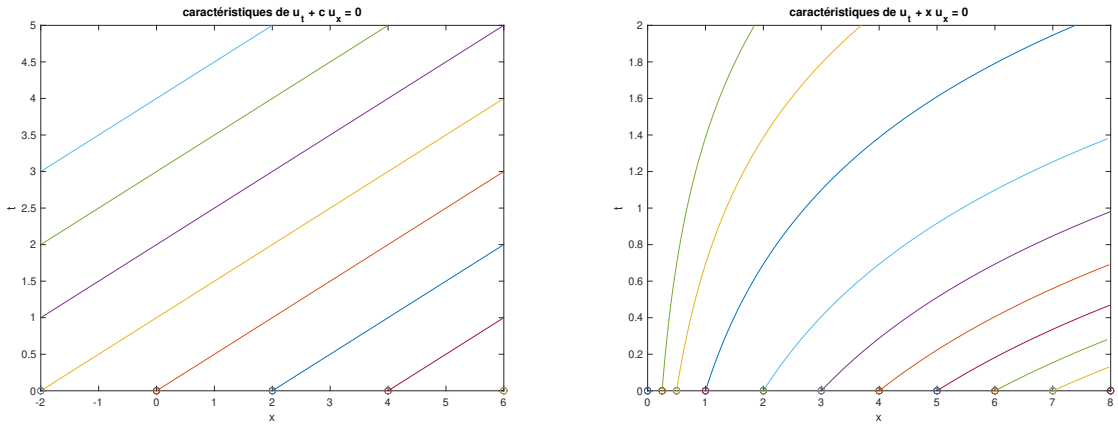


FIGURE 2.1 – courbes caractéristiques des EDP $u_t + cu_x = 0$ (gauche) et $u_t + xu_x = 0$ (droite).

2.2.2 Méthode des caractéristiques.

Reprenons l'équation de transport $u_t + c\,u_x = 0$ mais en ne supposant plus que la vitesse de propagation est constante.

Remarque. La vitesse $c = c(x)$ peut dépendre de la position, lorsque par exemple le milieu modélisé n'est pas homogène. □

Considérons l'équation de transport linéaire suivante :

$$u_t + c(x)\,u_x = 0, \quad u(x,0) = f(x). \tag{2.5}$$

Il est naturel d'introduire les courbes caractéristiques, définies par l'EDO

$$\dot{\xi} = c(\xi). \tag{2.6}$$

Si la fonction $c(x)$ est $\mathcal{C}^1$ on peut appliquer le théorème de Cauchy-Lipschitz et on est assuré de l'existence (locale au moins) de courbes caractéristiques. Dans ce cas la fonction d'une variable $v : t \mapsto u(\xi(t),t)$ vérifie une ODE très simple : $\dot{v} = u_t + c(x)\,u_x = 0$ donc $v(t) = v(0)$ est constante. Ce qui se traduit par

$u(\xi(t),t) = u(\xi(0),0)$. Ainsi si la caractéristique est issue de $\xi(0) = x_0$ à $t = 0$ on a $u(\xi(t),t) = f(x_0)$.
Si on veut calculer $u(x,t)$ en un point donné, il suffit donc de déterminer la courbe caractéristique passant par ce point (x,t) et de remonter le temps pour trouver sa valeur en $t = 0$ que nous noterons $x_0 = p(x,t)$ et que nous appellerons *le pied* de la caractéristique. Cela n'est pas toujours possible car l'existence de solution globale en temps n'est pas garantie, l'EDO n'étant pas linéaire. Mais dans le cas où c'est possible, le théorème de Cauchy-Lipschitz garantit l'unicité de la courbe caractéristique donc si $p(x,t)$ existe, il est défini sans ambigüité. Dans ce cas, la solution du problème de Cauchy (2.5) s'exprime :

$$u(x,t) = f(p(x,t))$$

Nous pouvons énoncer le théorème.

Théorème 5 *Soit $c(x)$ une fonction $\mathcal{C}^1$ sur $\mathbb{R}$. Soit $t \geq 0$. On suppose qu'il est possible de remonter la caractéristique passant par x,t jusqu'à un point $p(x,t)$ sur l'axe des x. Alors il existe une unique solution $\mathcal{C}^1$ au problème de Cauchy* (2.5)*, qui est donnée par*

$$u(x,t) = f(p(x,t)).$$

Etudions la dépendance de la solution par rapport à la donnée initiale. Soient $f(x)$ et $g(x)$ deux données initiales et $u(x,t)$, $v(x,t)$ les solutions correspondantes, on peut écrire

$$u(x,t) - v(x,t) = f(p(x,t)) - g(p(x,t))$$

et cela implique

$$\max_{x,t} |u(x,t) - v(x,t)| = \max_{x} |f(x) - g(x)|.$$

Le problème est donc bien posé

Exemple.

$$u_t + x\, u_x = 0, \quad u(x,0) = f(x).$$

dans ce cas les caractéristiques vérifient $\dot{\xi} = \xi$ donc $\xi(t) = x_0 e^t$. On trouve facilement le pied de chaque caractéristique : $p(x,t) = x\, e^{-t}$ et la solution est donnée par $u(x,t) = f(x\, e^{-t})$. Les caractéristiques sont représentées dans la figure 2.2.1.

2.2.3 Loi de conservation non linéaire : premières difficultés.

2.2.3.1 Principe d'une loi de conservation.

soit $u(x,t)$ la densité d'une quantité (par exemple la masse linéique d'un fluide en $kg \cdot m^{-1}$). La quantité totale (par exemple la masse) présente dans le segment $[a,b]$ à l'instant t est donc :

$$\int_a^b u(x,t)\, dx.$$

Si on sait par ailleurs que le *flux* de la quantité qui traverse le point x est donné par $F(u(x,t))$, comptée positivement lorsque la quantité traverse dans le sens des x croissants, de la gauche vers la droite. $F(u(x,t))$ est par exemple la masse de fluide qui traverse le point x par unité de temps, en $kg \cdot s^{-1}$). La conservation de la masse impose que

$$\frac{d}{dt}\left\{\int_a^b u(x,t)\, dx\right\} = F(u(a,t)) - F(u(b,t)).$$

Remarquez que le flux entrant (resp. sortant) est bien $F(u(a,t))$ (resp. $F(u(b,t))$. En supposant que la fonction $u(x,t)$ est $\mathcal{C}^1$, on peut dériver sous l'intégrale :

$$\int_a^b u_t(x,t)\, dx = F(u(a,t)) - F(u(b,t)).$$

On utilise ensuite le théorème fondamental du calcul infinitésimal :

$$F(u(a,t)) - F(u(b,t)) = -\int_a^b F(u(x,t))_x dx.$$

Ainsi

$$\int_a^b u_t(x,t) + F(u(x,t))_x \, dx = 0.$$

Comme le segment $[a,b]$ est quelconque, si la fonction u ainsi que la fonction F sont $\mathcal{C}^1$, cela impose que

$$u_t + F(u)_x = u_t + F'(u)\, u_x = 0. \tag{2.7}$$

L'EDP (2.7) est appelée une *loi de conservation.* C'est en géneral une EDP *non linéaire.*

Exemple. (Trafic routier.) Supposons que $u(x,t)$ représente la densité de voiture au point x circulant sur une route de gauche à droite. La vitesse à laquelle les véhicules circulent dépend évidemment de la densité de véhicules. Soit β la densité maximale de véhicule. La vitesse est donnée par

$$k \cdot (\beta - u)$$

où k est une constante de proportionnalité. Le flux de véhicule qui traverse au point x est alors :

$$F(u) = k\, u(\beta - u).$$

Pour simplifier prenons $k = 1$ dans la suite. Le flux maximal est atteint lorsque $u = \beta/2$. Si nous comparons avec l'EDP (2.5) la célérité c dépend maintenant de la solution u :

$$c(u) = F'(u).$$

Revenons au cas général. Utilisons à nouveau *la méthode des caractéristiques.* Ce sont les courbes solutions de l'EDO :

$$\frac{d\xi}{dt} = c(u(\xi,t)). \tag{2.8}$$

Comme précédemment u est *constante* le long des caractéristiques.

$$\frac{d}{dt}u(\xi(t),t) = u_x \frac{d\xi}{dt} + u_t = c(u)u_x + u_t = 0.$$

Notons $\xi(t,x_0)$ la courbe caractéristique issue du point $(x_0,0)$ sur l'axe des x. Alors $u(\xi(t,x_0),t) = u(x_0,0)$. Revenant alors à l'EDO (2.8), dont le second membre est en fait *constant*, puisque $u(\xi(t),t)$ est constant. La courbe caractéristique est en fait la droite :

$$x = \xi(t,x_0) = x_0 + c(u(x_0,0))\, t.$$

En utilisant la condition initiale $u(x_0,0) = f(x_0)$,

$$\xi(t,x_0) = x_0 + c(f(x_0))\, t$$

Retournons à l'exemple du trafic routier. Dans ce cas $c(u) = F'(u) = \beta - 2u$. Remarquez que $c(u) < 0$ quand $u > \beta/2$. Attention $c(u)$ n'est pas la vitesse individuelle des véhicules, puisque les véhicules roulent dans le même sens de gauche à droite. $c(u)$ est une vitesse de propagation d'onde. Par exemple, quand les véhicules s'arrêtent à un feu, il y a une onde de densité croissante qui remonte vers l'arrière de la file de véhicule. Pour fixer les idées, supposons que la densité initiale des véhicules est donnée par :

$$f(x) = u(x,0) = \begin{cases} 0 & \text{si} \quad x \le 0 \\ \beta x^2(3-2x) & \text{si} \quad 0 \le x \le 1 \\ \beta & \text{si} \quad 1 \le x \end{cases}$$

Pour étudier l'évolution de la densité de véhicules, on doit résoudre l'EDP

$$u_t + F(u)_x = u_t + c(u)\, u_x = 0 \quad u(x,0) = f(x) \tag{2.9}$$

avec $c(u) = F'(u) = \beta - 2u$. La condition $f(x) = \beta$ correspond à un bouchon, les voitures sont à l'arrêt. La route est vide pour $x \le 0$ et dans la région de transition $0 \le x \le 1$ la densité des voitures augmente de 0 à la capacité maximale β. La valeur $\beta/2$ ets atteinte pour $x = 1/2$. On peut alors tracer les caractéristiques

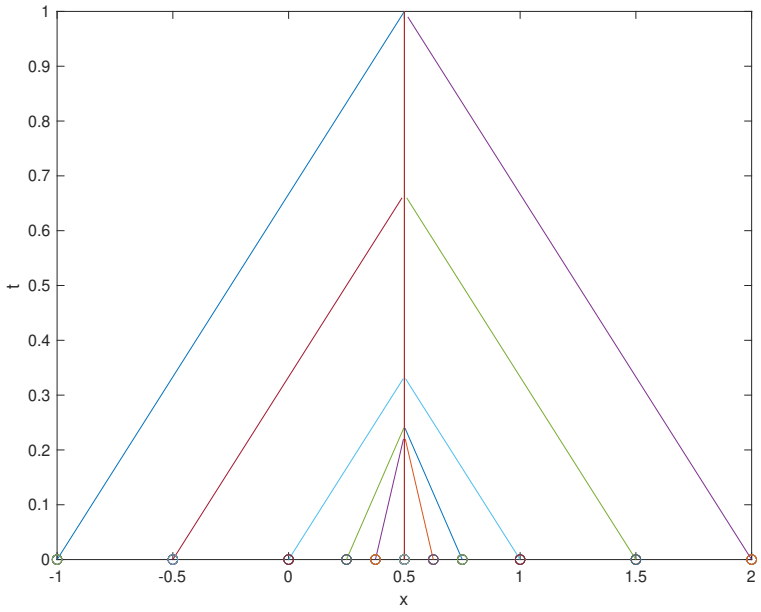


FIGURE 2.2 – caractéristiques de l'EDP $u_t + (\beta - 2u)u_x,\ \beta = 1.5$

dans le plan x, t qui sont des droites de pente positive si $0 \leq x \leq 1/2$, de pente négative si $1/2 \leq x \leq 1$, tandis que la caractéristique issue de $x = 1/2$ est verticale (on prend x en abscisse et t en ordonnée). Les pentes de ces droites varient donc continûment entre β pour $x \leq 0$ et $-\beta$ pour $x \geq 1$. Voir figure 2.2

Comme u est constante sur les caractéristiques, $u = 0$ sur la droite $x = \beta t$, $u = \beta/2$ sur la droite verticale $x = 1/2$ et $u = \beta$ sur la droite $x = 1 - \beta t$. A l'instant précise $t = 1/(2\beta)$ les caractéristiques se coupent au point $(1/2, 1/(2\beta))$. En ce point la densité n'est pas définie car elle devrait prendre 3 valeurs distinctes, ce n'est plus une fonction usuelle ! En fait la courbe solution ne peut pas être prolongée au sens classique car pour $t \geq t^\star$, où l'instant $t^\star \leq 1/(2\beta)$ une discontinuité apparaît. Pour mieux saisir le phénomène, on peut visualiser comment évolue le profil initial en fonction du temps sur la figure suivante 2.3. On voir sur

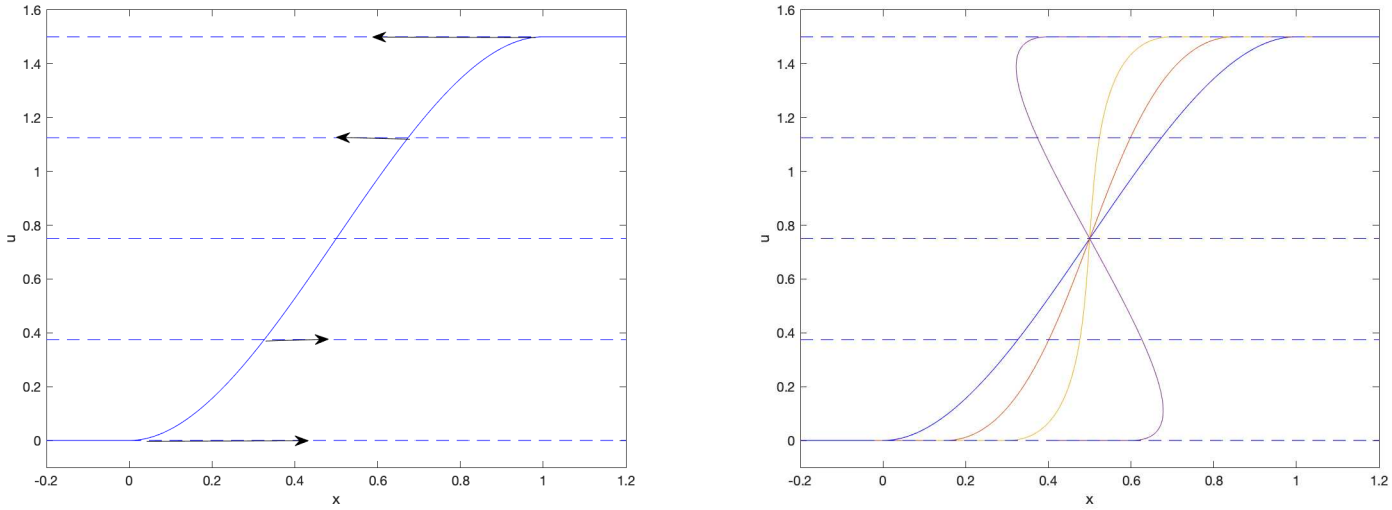


FIGURE 2.3 – gauche : profil initial et direction d'évolution. droite : "solution" à t=0, t=0.1, t=0.14, t=0.4 ($\beta = 1.5$)

la figure 2.3 que le profil ne correspond plus à une fonction pour $t = 0.4$. Pour prolonger la solution au delà de l'instant où les caractéristiques se croisent, il faut introduire une solution discontinue et définir la notion de« solution faible » car la solution étant discontinue, la dérivée u_x n'est plus définie au sens usuel. Il y a

développement d'une singularité même si la donnée initiale est lisse. Dans le cas présent la solution pour $t \geq t^\star$ est représentée sur la figure 2.4. La solution correspond à un « bouchon » qui n'évolue plus.

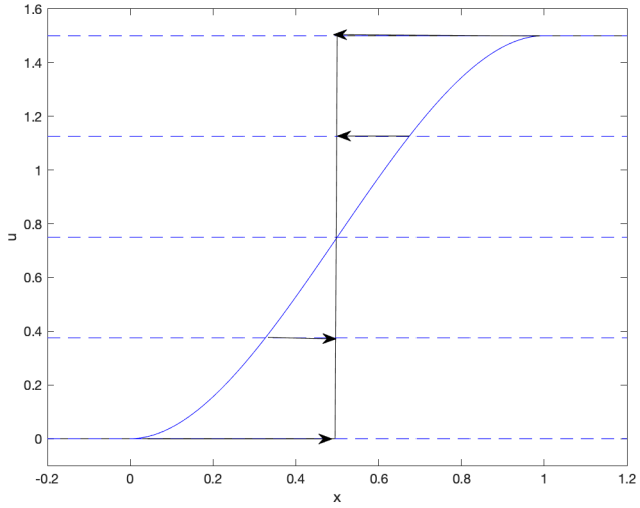


FIGURE 2.4 – solution lisse à $t = 0$ et solution discontinue pour $t \geq t^\star$.

La théorie des lois de conservations non linéaire dépasse le cadre d'un cours de L3. Pour des compléments, vous pouvez consulter l'ouvrage de référence de P. Lax [13].

2.2.4 Méthode des différences finies.

2.2.4.1 Principe de discrétisation.

Revenons sur les façons de discrétiser la dérivée d'une fonction g de classe $\mathcal{C}^2$.
Différence finies avant :

$$g'(x) = \frac{g(x+h) - g(x)}{h} + O(h)$$

Différence finies arrière :

$$g'(x) = \frac{g(x) - g(x-h)}{h} + O(h)$$

Différence finies centrées :

$$g'(x) = \frac{g(x+h) - g(x-h)}{2\,h} + O(h^2)$$

et enfin pour la dérivée seconde

$$g''(x) = \frac{g(x+h) - 2g(x) + g(x-h)}{h^2} + O(h^2).$$

Soit c un réel constant *positif* pour fixer les idées. On considère de nouveau l'EDP

$$u_t + c\,u_x = 0. \tag{2.10}$$

Appliquons ces discrétisations aux équations aux dérivées partielles. Comme il faut discrétiser des dérivées par rapport au temps et aussi par rapport à la variable d'espace x, il faut introduire un pas de temps δt et aussi un pas d'espace δx. Pour j et n entiers, on note $x_j = j\delta x$ et $t_n = n\delta t$. Ainsi les (x_j, t_n) définissent une grille de points ou un *maillage* dans le plan (x, t). Comme pour les EDO, on cherche à approcher $u(x_j, t_n)$. On notera $u_j^n \approx u(x_j, t_n)$ la valeur approchée calculée par le schéma considéré.

En discrétisant u_t et u_x par les différences finies *avant* , on obtient le schéma explicite suivant :

$$\frac{u_j^{n+1} - u_j^n}{\delta t} + c\,\frac{u_{j+1}^n - u_j^n}{\delta x} = 0.$$

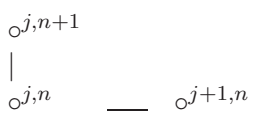
Le schéma peut s'écrire ainsi :

$$u_j^{n+1} = (1+r)\,u_j^n - r\,u_{j+1}^n.$$

où on a noté

$$r = \frac{c\,\delta t}{\delta x}$$

appelé *nombre de Courant.*[1] Le stencil de calcul est donc



On calcule l'*erreur de consistance* du schéma comme pour les schémas d'EDO en portant une solution exacte de l'EDP (2.10) dans le schéma numérique :

$$\epsilon(\delta t, \delta x) := u(x_j, t^{n+1}) - (1+r)\,u(x_j, t^n) + r\,u(x_{j+1}, t^n).$$

On obtient)en effectuant des développements de Taylor (Cf TD) :

$$\epsilon(\delta t, \delta x) = O(\delta t^2) + O(\delta t \cdot \delta x)$$

On suppose maintenant $c > 0$. Prenons la donnée initiale $u(x, t=0) = f(x)$ définie par

$$f(x) = \begin{cases} 0 & \text{si} \quad x \leq -1 \\ x+1 & \text{si} \quad -1 \leq x \leq 0 \\ 1 & \text{si} \quad 0 \leq x \end{cases}$$

Le schéma donne $u_j^n = 1, \quad \forall n \geq 0,\ \forall j \geq 0$. Le schéma ne peut pas converger car $u(x_j, t_n) = f(x_j - ct_n) = 0$ quand $x_j \leq ct_n - 1$. Le schéma ne va pas chercher l'information dans la bonne direction !

2.2.4.2 Décentrage amont ou « upwinding ».

Pour cette raison, on va utiliser une discrétisation spatiale « amont ». Lorsque $c > 0$, cela correspond à une différence finie arrière :

$$u_x(x,t) = \frac{u(x,t) - u(x-\delta x, t)}{\delta x} + \mathcal{O}(\delta x).$$

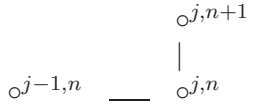
Le schéma « amont » ou « upwind » s'écrit alors :

$$\frac{u_j^{n+1} - u_j^n}{\delta t} + c\,\frac{u_j^n - u_{j-1}^n}{\delta x} = 0.$$

Le schéma peut s'écrire ainsi :

$$u_j^{n+1} = r\,u_{j-1}^n + (1-r)\,u_j^n. \tag{2.11}$$

Le stencil de calcul est maintenant



L'*erreur de consistance* du schéma est encore

$$\epsilon(\delta t, \delta x) = O(\delta t^2) + O(\delta t \cdot \delta x).$$

Cette fois le schéma va bien chercher l'information du bon côté. Cependant cela ne suffit pas. Regardons de plus près. Avec le schéma amont (2.11) la valeur u_j^n depend des valeurs à $t = 0$ suivantes $u_{j-n}^0, u_{j-(n-1)}^0, \ldots u_j^0$ qui sont situées dans l'intervalle $[x_j - n\delta x, x_j]$. La valeur exacte $u(x_j, t_n)$ devrait être $u(x_j - c\,t_n, 0)$. Cependant si le nombre de Courant $r > 1$ alors $x^\star = x_j - c\,t_n < x_j - n\delta x = x_{j-n}$. Le *domaine de dépendance* du schéma $u_{j-n}^0, u_{j-(n-1)}^0, \ldots u_j^0$ ne contient pas $x^\star$. La solution discrète u_j^n ne peut pas recevoir la bonne valeur car le

1. Richard Courant, 1888-1972, fondateur du Courant Institute of Mathematical Sciences, NYU.

schéma ne propage pas suffisamment pas vite l'information depuis l'amont. En revanche si $r = \frac{c\,\delta t}{\delta x} \leq 1$, le schéma a des chances de converger. La condition $\frac{c\,\delta t}{\delta x} \leq 1$ est appelée condition CFL. [2]

Remarque. Lorsque $r = CFL < 1$, on peut voir en particulier que u_j^{n+1} est une combinaison convexe de u_{j-1}^n et u_j^n donc le schéma vérifie le principe du maximum discret :

$$\inf_j u_j^0 \leq u_j^n \leq \sup_j u_j^0$$

□

Quand $CFL = \frac{c\,\delta t}{\delta x} < 1$, on peut montrer et nous admettrons dans ce cours que le schéma décentré amont est convergent quand $\delta t, \delta x \to 0$. L'erreur de consistance étant d'ordre 2, c'est un schéma d'ordre 1. Il n'est donc pas très précis, ainsi qu'on le constate sur la figure 2.6.

2.2.4.3 Stabilité au sens de Von Neumann.

On peut essayer d'augmenter la précision en utilisant le *schéma centré* suivant.

$$\frac{u_j^{n+1} - u_j^n}{\delta t} + c\,\frac{u_{j+1}^n - u_{j-1}^n}{2\delta x} = 0.$$

Le schéma s'écrit :

$$u_j^{n+1} = -\frac{r}{2}\,u_{j-1}^n + u_j^n + \frac{r}{2}\,u_{j+1}^n \tag{2.12}$$

Le domaine de dépendance du schéma est maintenant $u_{j-n}^0, u_{j-(n-1)}^0, \dots u_j^0, u_{j+1}^0 \dots u_{j+n}^0$ et convient aussi bien pour $c > 0$ que pour $c < 0$.
Si $r < 1$ le domaine de dépendance contient bien $x^\star$, cependant on constate sur la figure 2.6 que le schéma ne converge pas. On va prouver qu'il est *instable*. Pour cela on prend la donnée initiale oscillante $f(x) = \exp(ikx)$. Le nombre $k \in \mathbb{R}$ est la fréquence spatiale. Le schéma donne

$$u_j^n = G(k)^n u_j^0.$$

Le facteur d'amplification

$$G(k) = 1 + ir\sin(k\delta x)$$

est de module $1 + r^2\sin(k\delta x)^2$ strictement supérieur à 1 si $0 < k\delta x < \pi$ même si r est petit. Le schéma amplifie les oscillations de façon non bornées. On dit que le schéma centré (2.12) est *instable* au sens de Von Neumann.
Pour le schéma amont (2.11), le facteur d'amplification

$$G(k) = 1 - r + r\exp(-ik\delta x).$$

Lorsque $r \leq 1$, le module de G est inférieur à un car $|G| \leq (1-r) + r = 1$ donc *le schéma amont est stable.* Cela explique sa convergence en vertu du théorème de Lax que nous admettrons : *un schéma est convergent ssi il est stable et consistant.*

2.2.4.4 D'autres schémas.

Il est possible de stabiliser le schéma centré en le modifiant légèrement. C'est le schéma de Lax-Friedrichs. On reprend la même EDP d'advection linéaire (2.10), mais on ne suppose plus que $c > 0$. On considère le schéma explicite suivant :

$$\frac{u_j^{n+1} - \frac{1}{2}[u_{j+1}^n + u_{j-1}^n]}{\delta t} + c\,\frac{u_{j+1}^n - u_{j-1}^n}{2\delta x} = 0.$$

$$u_j^{n+1} = (\frac{1}{2} + \frac{c}{2\rho})\,u_{j-1}^n + (\frac{1}{2} - \frac{c}{2\rho})\,u_{j+1}^n.$$

2. d'après un article célèbre de Courant-Friedrichs-Lewy.

où on a noté $\rho = \delta x/\delta t$. L'*erreur de consistance* du schéma est

$$\epsilon(\delta t, \delta x) = \mathcal{O}(\delta t^2) + \mathcal{O}(\delta x^2)$$

Etudions la stabilité du schéma. Soit $k \in \mathbb{R}$. On prend la donnée initiale $f(x) = \exp(ikx)$. Le nombre k correspond à une fréquence spatiale. Le schéma donne

$$u_j^n = G(k)^n \cdot u_j^0.$$

où le facteur d'amplification $G(k) = (\frac{1}{2} - \frac{c}{2\rho})\exp(ik\delta x) + (\frac{1}{2} + \frac{c}{2\rho})\exp(-ik\delta x)$.
On vérifie aisément que $|G(k| \leq 1$ ssi $-1 \leq c/\rho \leq 1$. Lorsque cette condition appelée condition de Courant-Friedrichs-Lewy est vérifiée, on dit que le schéma de Lax-Friedrichs est *stable*. Dans ce cas $\|u_\cdot^n\|_\infty \leq \max_j |u_j^0|$. Le schéma est convergent et d'ordre 1.

Si on veut une schéma plus précis, on modifie encore le schéma de la façon suivante. Schéma de Lax-Wendroff. On reprend la même EDP d'advection linéaire (2.10), où $c \in \mathbb{R}$. On considère le schéma explicite suivant :

$$\frac{u_j^{n+1} - u_j^n}{\delta t} + c\,\frac{u_{j+1}^n - u_{j-1}^n}{2\delta x} - \frac{1}{2}c^2\delta t\,\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\delta x)^2} = 0.$$

Le schéma peut s'écrire ainsi :

$$u_j^{n+1} = \frac{1}{2}(r^2 - r)\,u_{j+1}^n + (1 - r^2)\,u_j^n + \frac{1}{2}(r^2 + r)\,u_{j-1}^n.$$

où on a noté $r = c\,\delta t/\delta x$ (appelé nombre de Courant)[3]. On calcule de même l'*erreur de consistance* du schéma :

$$\epsilon(\delta t, \delta x) = \mathcal{O}(\delta t^3) + \mathcal{O}(\delta x^3)$$

Etudions la stabilité du schéma. Soit $k \in \mathbb{R}$. On prend la donnée initiale $f(x) = \exp(ikx)$. Le gain du schéma est : $|G(k)| = |1 - r^2 + \frac{1}{2}(r^2 - r)\exp(ik\,\delta x) + \frac{1}{2}(r^2 + r)\exp(-ik\,\delta x)|$.
et on peut vérifier (Cf TD) que $|G(k| \leq 1$ si $-1 \leq r \leq 1$. Lorsque cette condition, appelée condition de Courant-Friedrichs-Lewy, est vérifiée, on dit que le schéma de Lax-Wendroff est *stable*. Il est donc convergent et d'ordre 2.

Cependant, comme les coefficients du schéma ne sont plus positifs, même lorsque $-1 \leq r \leq 1$ la propriété $\|u_\cdot^n\|_\infty \leq \max_j |u_j^0|$ n'est plus vérifiée. On dit que le schéma n'est plus monotone. Des petites oscillations parasites apparaissent au voisinage des discontinuités, cf figure 2.5.

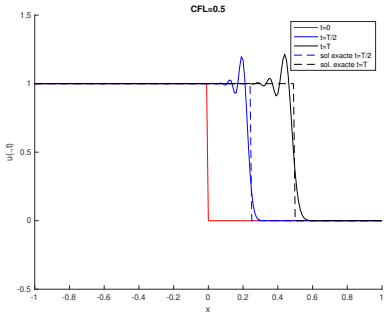


FIGURE 2.5 – Oscillations parasites avec le schéma de Lax-Wendroff.

On peut vérifier sur la figure 2.6 que le schéma centré diverge, que les schémas upwind et de Lax-Friedrichs convergent pour un $CFL < 1$ mais sont d'ordre 1 seulement (pas très précis), et que le schéma de Lax-Wendroff, bien que plus précis, étant d'ordre 2, peut présenter des petites oscillations parasites. On voit enfin sur la figure 2.7 que tous les schémas divergent lorsque la $CFL > 1$.

3. Richard Courant, 1888-1972, fondateur du Courant Institute of Mathematical Sciences, NYU.

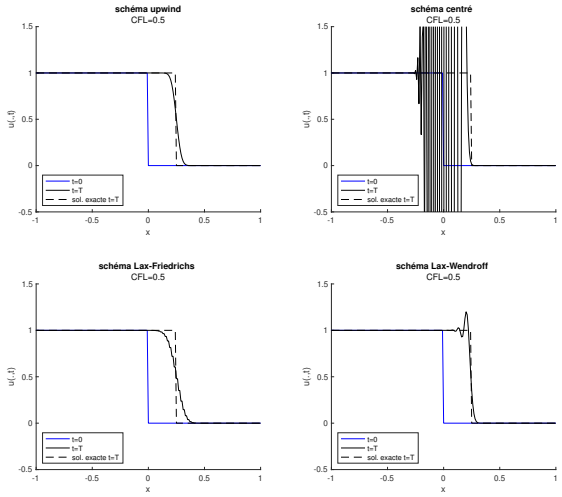


FIGURE 2.6 – Comparaison des différents schémas explicites.

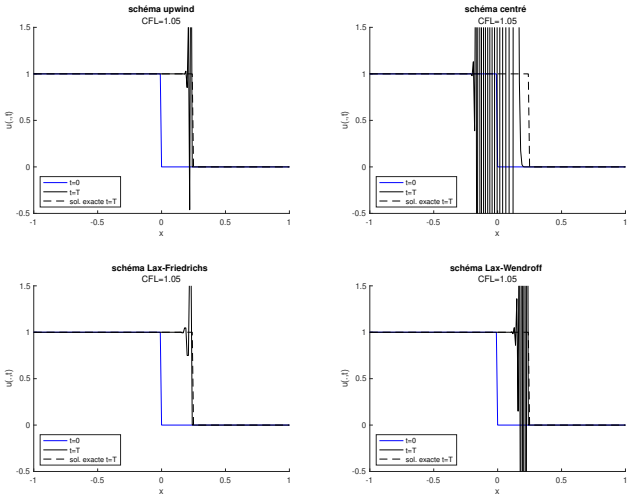


FIGURE 2.7 – Divergence des schémas lorsque $CFL > 1$.

2.3 Equations de diffusion.

2.3.1 Obtention de l'équation de la chaleur.

Soit $u(x,t)$ la température au point x à l'instant t dans un milieu unidimensionnel pour simplifier, par exemple une tige. La quantité totale de chaleur emmagasinée dans le segment $[a,b]$ à l'instant t est donc :

$$\int_a^b \rho c\, u(x,t)\, dx$$

où ρ désigne la densité du milieu en $kg \cdot m^{-1}$ et c sa capacité calorifique en $J \cdot kg^{-1}K^{-1}$.

Soit $F(u(x,t))$ le *flux* de chaleur qui traverse le point x, compté positivement lorsque la quantité traverse dans le sens des x croissants, de la gauche vers la droite. $F(u(x,t))$ en $J \cdot s^{-1}$). La conservation de chaleur impose que

$$\frac{d}{dt}\left\{\int_a^b \rho c\, u(x,t)\, dx\right\} = F(u(a,t)) - F(u(b,t)).$$

Remarquez que le flux entrant (resp. sortant) est bien $F(u(a,t))$ (resp. $F(u(b,t))$. En supposant que la fonction $u(x,t)$ est $\mathcal{C}^1$, on peut dériver sous l'intégrale :

$$\int_a^b \rho c\, u_t(x,t)\, dx = F(u(a,t)) - F(u(b,t)).$$

On utilise ensuite le théorème fondamental du calcul infinitésimal :

$$F(u(a,t)) - F(u(b,t)) = -\int_a^b F(u(x,t))_x dx.$$

Ainsi

$$\int_a^b \rho c\, u_t(x,t) + F(u(x,t))_x\, dx = 0.$$

Comme le segment $[a,b]$ est quelconque, si la fonction u ainsi que la fonction F sont $\mathcal{C}^1$, cela impose que

$$\rho c\, u_t + F(u)_x = 0. \tag{2.13}$$

Maintenant la loi de Fourier-Fick dit que le flux de chaleur traversant x est donné par

$$F(u) = -k u_x$$

où k est la conductivité thermique du milieu en $J \cdot m \cdot s^{-1} \cdot K^{-1}$. Ce principe exprime que la chaleur va des zones chaudes vers les zones froides et que le flux de chaleur est proportionnel au gradient de température. En portant l'expression de $F(u)$ dans (2.13) on obtient

$$u_t - \mu u_{xx} = 0. \tag{2.14}$$

où la constant $\mu = \frac{k}{\rho c} > 0$ est appelée coefficient de diffusion. Cette équation est appelée équation de la chaleur ou équation de diffusion. C'est une équation *linéaire* et elle vérifie donc le *principe de supeposition*(Cf section 2.2). Comme il y a une dérivée par rapport au temps, c'est une *équation d'évolution*. Il faut donc prescrire une donnée initiale

$$u(x, t=0) = f(x).$$

2.3.2 ($\star$) Solution par convolution avec noyau gaussien sur l'espace entier.

2.3.3 ($\star$) Solution par série de Fourier en domaine borné.

Cf TD 6 Effet régularisant et convergence vers l'état stationnaire.

2.3.4 Discrétisation par différences finies.

Comme dans la section 2.2.4.1, Introduisons un pas de temps δt et aussi un pas d'espace δx. Pour j et n entiers, on note $x_j = j\delta x$ et $t_n = n\delta t$. Ainsi les (x_j, t_n) définissent une grille de points ou un *maillage* dans le plan (x,t). On discrétise naturellement

$$u_t(x,t) = \frac{u(x,t+\delta t) - u(x,t)}{\delta t} + \mathcal{O}(\delta t). \quad \text{et} \quad u_{xx}(x,t) = \frac{u(x+\delta x,t) - 2u(x,t) + u(x-\delta x,t)}{\delta x^2} + \mathcal{O}(\delta x^2).$$

On obtient le schéma suivant :

$$u_j^{n+1} = r\, u_{j-1}^n + (1-2r)\, u_j^n + r\, u_{j+1}^n. \tag{2.15}$$

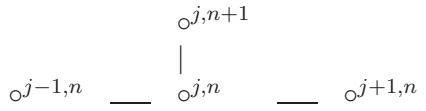
où u_j^n désigne a valeur approchée de $u(x_j, t_n)$ calculée par le schéma et

$$r = \frac{\mu\, \delta t}{\delta x^2}.$$

L'erreur de consistance du schéma se calcule aisément (Cf TD) :

$$\epsilon(\delta t, \delta x) = \mathcal{O}(\delta t^2) + \mathcal{O}(\delta t \cdot \delta x^2).$$

C'est un schéma *explicite* : pour calculer les valeurs u_j^{n+1} au temps t_{n+1}, il suffit de connaître les valeurs de u_j^n au temps précédent t_n. Le stencil de calcul est très simple :



Remarque. Lorsque $r \leq 1/2$, on peut voir en particulier que u_j^{n+1} est une moyenne pondérée des valeurs u_{j-1}^n u_j^n et $u_{j=1}^n$ donc le schéma vérifie le principe du maximum discret :

$$\inf_j u_j^0 \leq u_j^n \leq \sup_j u_j^0$$

□

Nous allons voir que la condition

$$r = \frac{\mu\,\delta t}{\delta x^2} < \frac{1}{2}$$

est en fait une condition nécessaire et suffisante de stabilité. Etudions la stabilité du schéma par la méthode de Von Neumann. Pour cela on prend la donnée initiale $f(x) = \exp(ikx)$. Le nombre k est la fréquence spatiale. Le schéma donne

$$u_j^n = G(k)^n u_j^0$$

où le facteur d'amplification de la fréquence k est donné par le nombre réel

$$G(k) = 1 - 2r + 2r\cos k\delta x = 1 - 4r\sin^2(k\delta x/2).$$

On a évidemment $1 - 4r \leq G(k) \leq 1$. Mais pour garantir $|G(k)| \leq 1$ pour toute fréquence k il faut et il suffit que

$$r = \frac{\mu\,\delta t}{\delta x^2} \leq 1/2. \tag{2.16}$$

C'est la *condition de stabilité du schéma explicite* (2.15).

Remarque. Cette condition est beaucoup plus exigeante que la condition CFL de l'équation de transport vue à la section 2.11. En effet elle impose un pas de temps δt de l'ordre de δx^2, ce qui demande un pas de temps très petit. Cela peut rendre les calculs numériques trop coûteux. [4] □

Sous la condition de stabilité 2.16 le schéma (2.15) est stable et consistant, le théorème de Lax permet d'affirmer qu'il est convergent. Comme l'erreur de consistance est en $\mathcal{O}(\delta t^2 + \delta t \cdot \delta x^2) = \mathcal{O}(\delta t \cdot (\delta t + \delta x^2))$, on peut démontrer le schéma est d'ordre 1 en temps et d'ordre 2 en espace.
Pour éviter la condition de stabilité (2.16), on est conduit à utiliser des schémas *implicites.* Pour cela on estime

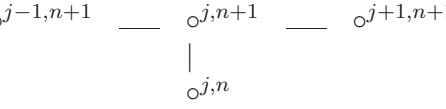
$$u_{xx}(x_j, t_{n+1}) = \frac{u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1}}{(\delta x)^2} + \mathcal{O}(\delta x^2),$$

$$u_t(x_j, t_{n+1}) = \frac{u_j^{n+1} - u_j^n}{\delta t} + \mathcal{O}(\delta t).$$

Ce qui revient à estimer u_t par une différence finie *arrière.*
On obtient le schéma :

$$-r\,u_{j-1}^{n+1} + (1+2r)\,u_j^{n+1} - r\,u_{j+1}^{n+1} = u_j^n \tag{2.17}$$

Le stencil de calcul est alors :



4. surtout en dimension supérieure à un.

On montre aisément que l'erreur de consistance est encore

$$\epsilon(\delta t, \delta x) = \mathcal{O}(\delta t^2) + \mathcal{O}(\delta t \cdot \delta x^2).$$

Il faut alors résoudre une système linéaire tridiagonal pour calculer u_j^{n+1}. Prenons un exemple pour fixer les idées. Considérons un maillage de 5 points $x_j, j = 0, \ldots, 4$. On se donne les valeurs initiales $u_j^0, j = 0, \ldots, 4$. Le schéma donne le système linéaire à 5 inconnues $u_j^1, j = 0, \ldots, 4$.

$$\left\{ \begin{array}{rrrrrcl} -ru_0^1 & +(1+2r)u_1^1 & -ru_2^1 & & & = & u_1^0 \\ & -ru_1^1 & +(1+2r)u_2^1 & -ru_3^1 & & = & u_2^0 \\ & & -ru_2^1 & +(1+2r)u_3^1 & -ru_4^1 & = & u_3^0 \end{array} \right.$$

C'est un système *sous-déterminé*. Une façon d'avoir le même nombre d'inconnues que d'équations est d'imposer des conditions limites en x_0 et x_4. Il y a de nombreuses possibilités. On peut prescrire la valeur des inconnues au bord (condition de Dirichlet), ou bien imposer les flux au bord (condition de Neumann). La façon la plus simple et neutre est d'imposer une condition limite *périodique* : $u_{-1}^n = u_4^n$, $u_5^n = u_0^n$. On obtient alors le système carré :

$$\left\{ \begin{array}{rrrrrcl} +(1+2r)u_0^1 & -ru_1^1 & & & -ru_4^1 & = & u_0^0 \\ -ru_0^1 & +(1+2r)u_1^1 & -ru_2^1 & & & = & u_1^0 \\ & -ru_1^1 & +(1+2r)u_2^1 & -ru_3^1 & & = & u_2^0 \\ & & -ru_2^1 & +(1+2r)u_3^1 & -ru_4^1 & = & u_3^0 \\ -ru_0^1 & & & -ru_3^1 & +(1+2r)u_4^1 & = & u_4^0 \end{array} \right.$$

Dans le cas général, si on considère un maillage de $N+1$ points $x_j, j = 0, \ldots, N$. Si on note U^n le vecteur$(u_0^n, u_1^n, \ldots u_N^n)^T$, le schéma se traduit par le système linéaire suivant : $AU^{n+1} = U^n$ avec la matrice tridiagonale périodique

$$\mathbf{A} = \begin{pmatrix} 1+2r & -r & & \cdots & -r \\ -r & 1+2r & -r & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & -r & 1+2r & -r \\ -r & \cdots & \cdots & -r & 1+2r \end{pmatrix} \tag{2.18}$$

On montre aisément que la matrice A est symétrique définie positive :

$$x^T A x = \sum_{1 \le i < n} (1+2r)x_i^2 - 2r x_i\, x_{i+1} + (1+2r)x_n^2 - 2r x_n x_1$$

comme $-x_i^2 - x_{i+1}^2 \le 2x_i\, x_{i+1} \le x_i^2 + x_{i+1}^2$ on voit que

$$(1+4r)(x_1^2 + \ldots x_n^2) \ge x^T A x \ge x_1^2 + \ldots x_n^2.$$

La matrice A est donc *inversible et bien conditionnée* et on peut calculer U^{n+1} en résolvant le système linéaire $AU^{n+1} = U^n$ à chaque itération.

Remarque. La matrice A est *creuse* et possède une structure bande ce qui facilite la résolution du système. □

On peut enfin étudier la stabilité du schéma par la méthode de Von Neumann. Pour cela prenons $u_j^0 = \exp(ikx_j)$. Montrons par récurrence sur n que $u_j^n = G(k)^n \exp(ikx_j)$. Pour $n = 0$ c'est vrai. Supposons HR_n et vérifions HR_{n+1}. Le schéma se traduit par l'égalité :

$$-rG(k)^{n+1}\exp(ikx_j)\exp(-ik\delta x) + (1+2r)G(k)^{n+1}\exp(ikx_j) - rG(k)^{n+1}\exp(ikx_j)\exp(ik\delta x) = G(k)^n \exp(ikx_j).$$

Simplifions par $G(k)^n \exp(ikx_j)$ il vient :

$$-rG(k)\exp(-ik\delta x) + (1+2r)G(k) - rG(k)\exp(ik\delta x) = 1.$$

Cela donne

$$G(k)(1+2r) - r(\exp(ik\delta x) + \exp(-ik\delta x) = 1$$

$$G(k) = \frac{1}{1+2r(1-\cos(k\delta x))}.$$

Avec l'identité $1-\cos(k\delta x) = 2\sin^2(k\delta x/2)$ on obtient

$$G(k) = \frac{1}{1+4r\sin^2(k\delta x/2)} \leq 1 \quad \forall k.$$

On en déduit que ce schéma est inconditionnellement stable, il n'y plus de restriction du type (2.16) sur les pas de temps et d'espace.

Le schéma implicite est encore consistant et l'erreur de consistance est du même ordre que celle du schéma explicite. Etant stable et consistant, il est convergent et c'est un schéma d'ordre un en temps.

Pour obtenir un schéma d'*ordre deux*, on utilise comme pour les équations différentielles ordinaires, un schéma des trapèzes Cf section 1.2.1.3. Cela revient à effectuer la moyenne des deux schémas explicites et implicites :

$$\frac{u_j^{n+1}-u_j^n}{\delta t} = \mu\,\frac{1}{2}\left\{\frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{\delta x^2} + \frac{u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1}}{\delta x^2.}\right\}$$

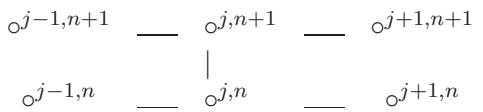
Le schéma s'écrit alors :

$$-r\,u_{j-1}^{n+1} + (1+2r)\,u_j^{n+1} - r\,u_{j+1}^{n+1} = r\,u_{j-1}^n + (1-2r)\,u_j^n + r\,u_{j+1}^n. \tag{2.19}$$

Attention, le nombre r vaut maintenant

$$r = \frac{1}{2}\frac{\mu\delta t}{\delta x^2}.$$

Le stencil de calcul est alors :



C'est le schéma de Crank-Nicolson, dont on calcule aisément l'erreur de consistance :

$$\epsilon(\delta t, \delta x) = \mathcal{O}(\delta t^3) + \mathcal{O}(\delta t \cdot \delta x^2) = \mathcal{O}(\delta t \cdot (\delta t^2 + \delta x^2)).$$

On montre aisément par la méthode de Von Neumann qu'il est inconditionnellement stable (exercice). C'est un schéma d'ordre deux en temps et en espace.

Là encore, il faut prescrire des conditions limites. Si on choisit par exemple des conditions limites périodiques, Dans le cas général, si on considère un maillage de $N+1$ points $x_j, j = 0, \ldots, N$. Si on note U^n le vecteur$(u_0^n, u_1^n, \ldots u_N^n)^T$, le schéma se traduit par le système linéaire suivant : $AU^{n+1} = BU^n$ avec A et B les matrices tridiagonales périodiques

$$\mathbf{A} = \begin{pmatrix} 1+2r & -r & & \cdots & -r \\ -r & 1+2r & -r & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & -r & 1+2r & -r \\ -r & \cdots & \cdots & -r & 1+2r \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 1-2r & r & & \cdots & r \\ r & 1-2r & r & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & r & 1-2r & r \\ r & \cdots & \cdots & r & 1-2r \end{pmatrix} \tag{2.20}$$

On a déjà prouvé que A était inversible donc le schéma est bien posé.

2.3.5 Cas d'équilibre en dimension un.

Abordons maintenant un dernier type d'équation aux dérivéees partielles, correspondant aux phénomènes d'équilibre. Lorsque $t \to +\infty$, on constate que la température $u(x,t)$ converge vers un état stationnaire *indépendant de* t. On note encore $u(x)$ l'état stationnaire d'équilibre thermique. On a évidemment $u_t = 0$ donc $u(x)$ est solution de l'équation différentielle ordinaire :

$$-\mu \frac{d^2u}{dx^2} = 0$$

(avec des conditions limites convenables en domaine borné). Si on ajoute une source de chaleur, disons $f(x)$ l'équation devient

$$-\mu \frac{d^2u}{dx^2} = f.$$

Pour fixer les idées, supposons qu'on s'intéresse à l'équilibre thermique d'un barreau homogène $0 < x < L$. On prescrit la température des deux extrémités $u(x=0) = u_g,\ u(x=L) = u_d$. La température à l'équilibre est solution du problème aux limites :

$$-\mu \frac{d^2u}{dx^2}(x) = f(x), \quad 0 < x < L.$$

$$u(x=0) = u_g,\ u(x=L) = u_d.$$

On discrétise ce problème par la méthode des différences finies (quitte à changer f on peut prendre $\mu = 1$). On note $h = \delta x = L/N$ et on definit la subdivision $0 = x_0 < x_1 < x_2 < \dots x_N < x_{N+1} = L$. On pose $u_j = u(x_j)$ et $f_j = f(x_j)$. On utilise les différences finies centrées pour discrétiser la dérivée seconde et on obtient

$$\frac{-u_{j-1} + 2u_j - u_{j+1}}{h^2} = f_j. \tag{2.21}$$

On obtient encore un système tridiagonal symétrique défini positif $AU = F$ en les inconnues $U = (u_1, u_2, \dots, u_N)^T$ et de second membre $F = (f_1 + \frac{1}{h^2}u_g, u_2, \dots, f_N + \frac{1}{h^2}u_d)^T$ où la matrice est

$$\mathbf{A} = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \tag{2.22}$$

2.3.6 En dimension supérieure, équation de Poisson.

Si on s'intéresse maintenant à l'équilibre thermique d'un carré homogène $\Omega =]0, L[\times]0, L[$ on est conduit au problème au limite suivant :

$$-\mu \frac{\partial^2 u}{\partial x^2}(x,y) + \frac{\partial^2 u}{\partial y^2}(x,y) = f(x,y), \quad (x,y) \in \Omega,$$

$$u_{|\partial\Omega} = u_b \quad (x,y) \in \partial\Omega.$$

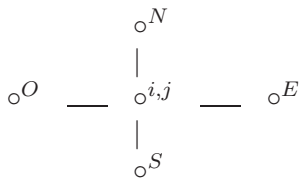
C'est l'équation de Poisson qui fait intervenit l'opérateur *laplacien*

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Les inconnues $u_{i,j} = u(x_i, y_j)$ sont définies sur la grille de points (ih, jh) où le pas $h := 1/(N+1)$ On discrétise alors cette équation aisément par la méthode des différences finies et on obtient le célèbre schéma à 5 points,

$$\frac{-u_{i,j-1} - u_{i-1,j} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1}}{h^2} = f_j.$$

dont le stencil de calcul est :



Où on a utilisé les 4 points cardinaux pour désigner $(i \pm 1, j \pm 1)$ les 4 points voisins de (i, j).

2.4 ($\star$) Epilogue : classification.

Nous avons vu ainsi trois types très différents d'équations aux dérivées partielles.

— L'équation de transport :

$$u_t + cu_x = 0,$$

— L'équation de diffusion :

$$u_t - \mu u_{xx} = 0,$$

— L'équation de Laplace (ou de Poisson)

$$u_{xx} + u_{yy} = 0.$$

Plus généralement considérons une EDP du second ordre linéaire à coefficients constants réels.

$$au_{xx} + 2bu_{xy} + cu_{yy} + du_x + eu_y + fu = 0. \tag{2.23}$$

où les réels $(a, b, c) \neq (0, 0, 0)$. Cherchons des solutions sous forme $\exp(x\xi + y\eta)$. Injectons cette expression dans l'EDP, on obtient l'équation caractéristique :

$$a\xi^2 + 2b\xi\eta + c\eta^2 + (d\xi + e\eta) + f = 0.$$

C'est l'équation d'une conique dans le plan (ξ, η). D'après la théorie des formes quadratiques on connaît la nature de la conique en étudiant la forme quadratique

$$a\xi^2 + 2b\xi\eta + c\eta^2 = (\xi, \eta) \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix}$$

Lorsque $ac - b^2 > 0$ on a l'équation d'une ellipse, Lorsque $ac - b^2 = 0$ on a l'équation d'une parabole, Lorsque $ac - b^2 < 0$ on a l'équation d'une hyperbole. Pour cette raison, lorsque $ac - b^2 > 0$, (resp. $ac - b^2 = 0$, $ac - b^2 < 0$) on dit que l'EDP est elliptique,(resp. parabolique, hyperbolique.) Nous allons montrer qu'avec un changement de variable affine on peut se ramener à trois formes canoniques d'EDP. La matrice

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

est symétrique réelle, elle est donc diagonalisable dans une base orthonormale

$$P = \begin{pmatrix} \alpha & \gamma \\ \beta & \delta \end{pmatrix}$$

Cela se traduit par

$$A = P \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} P^T$$

où λ et μ sont les valeurs propres de A. Posons maintenant :

$$X = \alpha x + \beta y \quad Y = \gamma x + \delta y.$$

$$U(X, Y) = u(x, y).$$

Par la règle de différentiation composée :

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial X}\,\frac{\partial X}{\partial x} + \frac{\partial}{\partial Y}\,\frac{\partial Y}{\partial x}, \quad \frac{\partial}{\partial y} = \frac{\partial}{\partial X}\,\frac{\partial X}{\partial y} + \frac{\partial}{\partial Y}\,\frac{\partial Y}{\partial y}$$

Ce qui donne

$$\frac{\partial}{\partial x} = \alpha\frac{\partial}{\partial X} + \gamma\frac{\partial}{\partial Y}, \quad \frac{\partial}{\partial y} = \beta\frac{\partial}{\partial X} + \delta\frac{\partial}{\partial Y}.$$

On calcule alors

$$u_{xx} = \alpha^2\, U_{XX} + 2\alpha\gamma\, U_{XY} + \gamma^2\, U_{YY},$$
$$u_{yy} = \beta^2\, U_{XX} + 2\beta\delta\, U_{XY} + \delta^2\, U_{YY},$$
$$u_{xy} = \alpha\beta\, U_{XX} + (\alpha\delta + \beta\gamma)\, U_{XY} + \gamma\delta\, U_{YY}.$$

L'EDP (2.23) devient dans les nouvelles coordonnées

$$\tilde{a}\, U_{XX} + 2\tilde{b}\, U_{XY} + \tilde{c}\, U_{YY} + \tilde{d}\, U_X + \tilde{e}\, U_Y + \tilde{f}\, U = 0.$$

Considérons les *termes du second ordre*.

$$\tilde{a} = a\alpha^2 + 2b\alpha\beta + c\beta^2,$$
$$\tilde{b} = a\alpha\gamma + 2b(\alpha\delta + \beta\gamma) + 2c\beta\delta,$$
$$\tilde{c} = a\gamma^2 + 2b\gamma\delta + c\delta^2.$$

On reconnaît

$$\tilde{A} = \begin{pmatrix} \tilde{a} & \tilde{b} \\ \tilde{b} & \tilde{c} \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} \alpha & \gamma \\ \beta & \delta \end{pmatrix}$$

La matrice

$$\tilde{A} = P^T A P = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}.$$

Dans les nouvelles variables, l'équation (2.23) s'écrit donc simplement

$$\lambda\, U_{XX} + \mu\, U_{YY} + \text{termes d'ordres inférieurs} = 0$$

Il suffit alors d'effectuer un dernier *scaling* ou changement d'échelle pour se ramener aux trois formes canoniques (on revient aux notations usuelles pour simplifier)

— elliptique si $ac - b^2 > 0$,

$$u_{xx} + u_{yy} + \text{termes d'ordres inférieurs} = 0,$$

— parabolique si $ac - b^2 = 0$,

$$u_{xx} + \text{termes d'ordres inférieurs} = 0,$$

— hyperbolique si $ac - b^2 < 0$,

$$u_{xx} - u_{yy} + \text{termes d'ordres inférieurs} = 0,$$

Le cas elliptique correspond à l'équation de Laplace, le cas parabolique correspond à l'équation de diffusion de la chaleur, le cas

$$u_{xx} - u_{yy} = 0$$

correspond à *l'équation des ondes ou des cordes vibrantes.* Cette dernière équation est peut être écrite sous forme d'une système de deux équations de transport :

$$\begin{aligned} u_x - u_y &= v \\ v_x + v_y &= 0. \end{aligned}$$

Cette classification est cependant insuffisante pour traiter tous les types d'EDP, par exemple les EDP dispersives (Schrödinger, KdV) ne rentrent pas dans ce cadre. Il n'y pas de théorie générale des EDP, c'est ce qui fait la richesse fascinante de ce domaine.

Bibliographie

[1] Gilbert Strang, Introduction to applied mathematics, Wellesley-Cambridge press, 1986.

[2] H.R. Schwarz, Numerical Analysis, A comprehensive introduction, Wiley, 1989.

[3] Cleve Moler, Numerical computing with Matlab, `http://www.mathworks.com/moler/`

[4] Le Mathematica computational knowledge engine `http://www.wolframalpha.com/`

[5] 3BLUE1BROWN SERIES Saison 4 Episode 1 Differential equations, studying the unsolvable `https://youtu.be/p_di4Zn4wz4`

[6] Arieh Iserles, A First course in the numerical analysis of differential equations, Cambridge University press, 2009.

[7] M. Crouzeix, A.L. Mignot, Analyse numérique des équations différentielles : Masson, Paris, 1984.

[8] M. Crouzeix, A.L. Mignot, Exercice d'analyse numérique des équations différentielles : Masson, Paris, 1986.

[9] J.P. Demailly, Analyse numérique et équations différentielles : EDP Sciences, Grenoble, 1991.

[10] E. Hairer, S.P. Norsett, G. Wanner, Solving Ordinary Differential Equations I : Nonstiff Problems, Springer, Berlin, 2009.

[11] N.J. Higham, Accuracy and stability of numerical algorithms : Siam, Philadelphia, 1996.

[12] Jeffery Cooper, Introduction to Partial Differential Equations with Matlab, Birkhäuser, 1998.

[13] Peter D. Lax, Hyperbolic systems of conservation laws and the mathematical theory of shock waves, SIAM Regional Conference Series in Applied Mathematics, 11, 1972.