Licence de Mathématiques - Sorbonne Université
année 2020-2021

UE LU3MA232

Analyse Numérique

Antoine Gloria

Laboratoire Jacques-Louis Lions

30 novembre 2020

Table des matières

1 Introduction **5**
- 1.1 Analyse numérique . . . 5
- 1.2 Ce qu'il faut savoir avant de commencer . . . 6
  - 1.2.1 Prérequis généraux . . . 6
  - 1.2.2 Topologie . . . 6
  - 1.2.3 Calcul différentiel . . . 7
  - 1.2.4 Equations différentielles ordinaires . . . 8

2 Notion de schéma numérique pour les EDO **9**
- 2.1 Equations différentielles ordinaires : rappels . . . 9
  - 2.1.1 Définition . . . 9
  - 2.1.2 Le problème de Cauchy . . . 10
  - 2.1.3 Approche historique par la méthode d'Euler . . . 13
- 2.2 La notion de schéma numérique . . . 17
- 2.3 Schémas numériques généraux . . . 25

3 Analyse numérique matricielle **29**
- 3.1 Motivation . . . 29
- 3.2 Méthodes directes . . . 30
  - 3.2.1 Notation et opérations élémentaires . . . 30
  - 3.2.2 Factorisation $A = LU$ et résolution de $AY = B$ . . . 30
  - 3.2.3 Méthode de Cholesky . . . 34
- 3.3 Méthodes itératives . . . 37
  - 3.3.1 Reformulation de $AY = B$ comme problème de minimisation . . . 37
  - 3.3.2 Méthode du gradient à pas fixe . . . 38
  - 3.3.3 Méthode du gradient conjugué . . . 41
- 3.4 Quelle méthode choisir en pratique ? . . . 44

4 Schémas à un pas : analyse générale **45**
- 4.1 Schémas explicites . . . 45
  - 4.1.1 Formulation du schéma . . . 45
  - 4.1.2 Stabilité . . . 47
  - 4.1.3 Consistance . . . 50
  - 4.1.4 Convergence . . . 52
  - 4.1.5 Ordre d'un schéma, estimation d'erreur . . . 53
- 4.2 Schémas implicites . . . 59

5 Méthode de Newton **67**
5.1 Présentation en dimension 1 . . . 67
5.2 La méthode de Newton dans $\mathbb{R}^n$ . . . 70
5.2.1 $g(x) = 0$ du point de vue théorique . . . 71
5.2.2 La méthode de Newton(-Raphson) . . . 72
5.2.3 Quelques illustrations en dimension 2 . . . 76
5.3 Pour en savoir plus : le théorème de Kantorovich . . . 79
5.4 Retour à la méthode d'Euler implicite . . . 83
5.5 Considérations de mise en œuvre pratique . . . 84
5.6 Les méthodes de quasi-Newton . . . 85

6 Schémas d'ordre élevé **89**
6.1 Schémas de type Taylor . . . 89
6.2 Schémas de Runge-Kutta. . . . 91
6.3 Méthodes d'Adams . . . 104

7 Intégration numérique **109**
7.1 Intégration de type interpolation de Lagrange . . . 109
7.1.1 Interpolation de Lagrange . . . 109
7.1.2 Formules de Newton-Côtes . . . 112
7.1.3 Formules composites . . . 113
7.2 Intégration de type interpolation de Gauss . . . 114
7.2.1 Polynômes de Legendre . . . 114
7.2.2 Méthode de Gauss-Legendre . . . 116
7.2.3 Formules composites . . . 118

8 Précision numérique **119**
8.1 Schémas d'ordre élevé et précision numérique . . . 119
8.2 Contrôle du pas de temps . . . 121

9 Propriétés asymptotiques de schémas numériques **127**
9.1 Stabilité absolue . . . 127
9.1.1 Domaine de stabilité . . . 130
9.1.2 Stabilité absolue de quelques schémas numériques . . . 132
9.2 Schémas numériques pour les systèmes hamiltoniens . . . 137

A Rappels de topologie **147**
A.1 Espaces métriques . . . 147
A.2 Normes d'application linéaire et norme subordonnée . . . 150
A.3 L'ensemble des matrices inversibles est ouvert . . . 151

B Rappels de calcul différentiel **153**

C Rappels sur les équations différentielles ordinaires **159**
C.1 Présentation générale . . . 159
C.1.1 Systèmes différentiels d'ordre 1 . . . 159
C.1.2 Le cas des équations différentielles autonomes . . . 160
C.2 Équations différentielles linéaires . . . 161

C.2.1 Définitions et propriétés générales 162
C.2.2 Systèmes différentiels linéaires à coefficients constants 164
C.2.3 Systèmes différentiels linéaires à coefficients variables 170
C.3 Existence et unicité dans le cas général 177
C.3.1 Résultats d'existence et d'unicité dans le cas général 177
C.4 existence locale d'une solution 182
C.4.1 Existence globale à l'aide des fonctions de Liapounov 189
Bibliographie 194
Index 196

Chapitre 1

Introduction

1.1 Analyse numérique

L'analyse numérique est une discipline à l'interface des mathématiques et de l'informatique. Elle s'intéresse tant au développement, à l'analyse, qu'à la mise en pratique des méthodes permettant de résoudre, par des calculs purement numériques, des problèmes d'analyse mathématique.

Plus formellement, l'analyse numérique est l'étude des algorithmes permettant de résoudre numériquement par discrétisation les problèmes de mathématiques continues (distinguées des mathématiques discrètes). Cela signifie qu'elle s'occupe principalement de répondre de façon numérique à des questions posées sur des modèles (formulés en langage mathématique) survenant dans les sciences physiques et l'ingénierie. Branche des mathématiques appliquées, son développement est étroitement lié à celui des outils informatiques.

Le calcul scientifique est omniprésent dans la vie quotidienne : votre téléphone n'est autre qu'un calculateur. Chaque texte, image, son, ou vidéo échangé ou partagé implique du calcul, de la compression ou décompression de données etc. L'analyse numérique est la branche des mathématiques qui étudie ces algorithmes et répond aux questions telles que

— L'algorithme est-il bien défini (ou va-t-il faire planter l'ordinateur) ?
— L'algorithme fait-il bien ce que pour quoi il a été créé (exemple : résout-il bien mon équation) ?
— L'algorithme est-il précis (penser à la qualité de compression d'une image) ?
— L'algorithme est-il efficace et rapide ?

L'objectif de ce cours est d'adopter ce schéma de pensée et d'introduire quelques méthodes d'analyse numérique pour répondre à des questions variées. Pour donner un fil conducteur au cours, nous nous focaliserons sur l'approximation numérique des solutions d'équations différentielles ordinaires. Outre son intérêt historique, la résolution numérique des équations différentielles ordinaires est le premier pas vers la résolution numérique des équations aux dérivées partielles (équations utilisées pour modéliser la plupart des phénomènes physiques, mécaniques, chimiques, économiques, biologiques voire bio-médicaux...). Notre étude des EDO nous amènera à définir le concept de schéma numérique, les notions de stabilité et de consistance, et d'aborder les questions de résolution numérique de grands systèmes linéaires, d'équations non linéaires, des questions d'approximation de fonctions et d'intégration numérique. Nous conclurons par la question de reproduire au niveau discret

des propriétés fondamentales des modèles continus associés (ici dans le cadre des schémas symplectiques pour les systèmes hamiltoniens).

De nombreux algorithmes et domaines de l'analyse numérique ne seront pas abordés dans ce cours, notamment les questions d'optimisation, pour lesquelles nous renvoyons par exemple au cours de calcul différentiel et optimisation.

1.2 Ce qu'il faut savoir avant de commencer

Les mathématiques sont une discipline cumulative : on y construit des édifices de plus en plus élevés sur des bases larges et solides. Ce cours ne fera pas exception à ce principe général des études en mathématiques. Aucun prérequis n'est cependant demandé en programmation.

1.2.1 Prérequis généraux

— Ensembles, applications.
— $\mathbb{R}$, $\mathbb{C}$, propriétés algébriques et topologiques.
— Algèbre linéaire en dimension finie, matrices, diagonalisation, trigonalisation (on ne saurait trop insister sur combien l'algèbre linéaire est fondamentale).
— Espaces euclidiens, espaces hermitiens.
— L'intégrale de Riemann suffira. Une intégrale fonction de sa borne supérieure donne une primitive de l'intégrande (par exemple si celle-ci est continue).

1.2.2 Topologie

— Espaces métriques, boules ouvertes, boules fermées.
— Topologie d'un espace métrique : ouverts, fermés. Intérieur et adhérence d'une partie d'un espace métrique.
— Convergence d'une suite dans un espace métrique.
— Applications continues entre deux espaces métriques. Applications uniformément continues entre deux espaces métriques.
— Compacts d'un espace métrique. L'image d'un compact par une application continue est compacte. Toute application continue d'un compact à valeurs dans un espace métrique est bornée. Tout application continue d'un compact métrique à valeurs dans un espace métrique est uniformément continue.
— Espaces vectoriels [1] normés, topologie d'espace métrique associée à une norme. Déclinaison des notions précédentes dans ce cas particulier.
— Suite de Cauchy dans un espace métrique, espace métrique complet, espace vectoriel normé complet ou de Banach.
— Toute application uniformément continue d'une partie d'un espace métrique à valeurs dans un espace métrique complet admet un unique prolongement continu à l'adhérence de la partie de départ.
— Dans un espace de Banach, toute série normalement convergente, c'est-à-dire dont la série des normes est convergente dans $\mathbb{R}_+$, est convergente, c'est-à-dire que la suite

1. On ne parlera ici que d'espaces vectoriels sur les corps $\mathbb{R}$ ou $\mathbb{C}$.

de ses sommes partielles converge dans l'espace de Banach. La limite de cette suite est appelée somme de la série.
— Toutes les normes sur un espace vectoriel de dimension finie sont équivalentes.
— Tous les espaces vectoriels normés de dimension finie sont complets, *i.e.*, de Banach.
— Normes usuelles sur $\mathbb{R}^m$, $\mathbb{C}^m$.
— Les compacts d'un espace vectoriel normé de dimension finie sont ses fermés bornés.
— L'espace des fonctions continues d'un intervalle fermé borné $\bar{I}$ à valeurs dans $\mathbb{R}^m$, noté $C^0(\bar{I};\mathbb{R}^m)$, muni de la norme $\|y\|_{C^0(\bar{I};\mathbb{R}^m)} = \max_{t\in\bar{I}} \|y(t)\|_{\mathbb{R}^m}$ (où l'on a pris n'importe quelle norme sur $\mathbb{R}^m$) est un espace de Banach (de dimension infinie). La convergence d'une suite de fonctions au sens de cet espace n'est autre que la convergence uniforme sur $\bar{I}$.

Quelques rappels sont donnés à l'annexe A.

1.2.3 Calcul différentiel

On se doute bien que pour parler d'EDO, il faut savoir différentier toutes sortes d'applications.

— Application différentiable d'un ouvert d'un espace vectoriel normé à valeurs dans un autre espace vectoriel normé (différentiabilité en un point, différentiabilité partout). Application continûment différentiable.
— Différentielle en un point d'une telle application comme application linéaire entre les deux espaces vectoriels.
— À toutes fins utiles, on rappelle qu'une dérivée partielle n'a rien de bien méchant. C'est juste une dérivée ordinaire par rapport à une variable quand on fixe toutes les autres variables.
— En dimension finie, après un choix de base dans chacun des deux espaces vectoriels de départ et d'arrivée, la matrice qui représente la différentielle d'une application dans ces bases est appelée sa *matrice jacobienne*. Ses coefficients sont les dérivées partielles des différentes composantes. Plus explicitement, soit $f : U \to F$ une application de U ouvert d'un espace vectoriel normé E de dimension k à valeurs dans un espace vectoriel normé F de dimension m. On suppose f différentiable au point $x_0 \in U$. Sa différentielle en x_0 est une application linéaire df_{x_0} de E dans F. Si l'on choisit une base $(u_j)_{j=1,\dots,k}$ de E et une base $(v_i)_{i=1,\dots,m}$ de F, et que l'on note (x_j) les coordonnées cartésiennes associées dans E et (y_i) les coordonnées cartésiennes associées dans F, alors l'application f est représentée par m applications coordonnées f_i de l'ouvert de $\mathbb{R}^k$ contenant les coordonnées des points de U, à valeurs dans $\mathbb{R}$, de telle sorte que

$$f(x) = \sum_{i=1}^{m} f_i(x_1, x_2, \dots, x_k) v_i, \text{ où } x = \sum_{j=1}^{k} x_j u_j.$$

La différentielle df_{x_0} de f en x_0 est alors représentée dans ces bases par la matrice jacobienne $\nabla f(x_0)$, matrice $m \times k$ dont les coefficients sont donnés par $(\nabla f(x_0))_{ij} = \frac{\partial f_i}{\partial x_j}(x_0)$, $i = 1, \dots, m$, $j = 1, \dots, k$. Cette représentation a lieu au sens usuel de

l'algèbre linéaire, c'est-à-dire que pour tout vecteur $h = \sum_{j=1}^{k} h_j u_j$ de E, on a

$$df_{x_0}h = \sum_{i=1}^{m} (df_{x_0}h)_i v_i \quad \text{avec} \quad (df_{x_0}h)_i = \sum_{j=1}^{k} (\nabla f(x_0))_{ij} h_j = \sum_{j=1}^{k} \frac{\partial f_i}{\partial x_j}(x_0) h_j.$$

On reconnaît un simple produit matrice-vecteur. Bien sûr, le fait que f soit différentiable en x_0 implique que toutes ces dérivées partielles existent en x_0.

— La composée de deux applications différentiables est différentiable. Si $f : U \to F$ est différentiable en $x_0 \in U \subset E$ et $g : V \to G$ est différentiable en $f(x_0) \in V \subset F$, U et V ouverts de leur espace respectif, alors $g \circ f : U \to G$ est différentiable en x_0 et sa différentielle est la composée des différentielles de f et g, $d(g \circ f)_{x_0} = dg_{f(x_0)} \circ df_{x_0}$.

— Avec des choix de bases dans les trois espaces vectoriels, comme la matrice de la composée de deux applications linéaires est le produit de leurs matrices (dans le même ordre), on en déduit pour les matrices jacobiennes $\nabla(g \circ f)(x_0) = \nabla g(f(x_0)) \nabla f(x_0)$.

— En explicitant tout cela avec des dérivées partielles, on obtient la très importante formule de dérivation des fonctions composées de plusieurs variables, qu'il faut absolument savoir appliquer quelles que soient les circonstances, même les plus adverses,

$$\frac{\partial (g \circ f)_l}{\partial x_j}(x_0) = \sum_{i=1}^{m} \frac{\partial g_l}{\partial y_i}(f(x_0)) \frac{\partial f_i}{\partial x_j}(x_0),$$

pour $j = 1, \ldots, k$ et $l = 1, \ldots, n$ où n est la dimension de G. [2] C'est une application immédiate de la formule générale donnant les coefficients d'un produit matriciel en fonction des coefficients des matrices dont on effectue le produit. C'est de l'algèbre linéaire en fait (dont on ne saurait trop rappeler combien elle est fondamentale).

— On aura sûrement besoin de l'inégalité des accroissements finis, et plus généralement de l'inégalité de Taylor-Lagrange, ainsi que de la formule de Taylor avec reste intégral (de préférence), parfois avec reste de Taylor-Lagrange quand c'est possible, ou encore, quand ce n'est pas la peine de trop se fatiguer, avec un reste exprimé sans autre forme de procès et un peu vaguement avec des O (notation de Landau).

Quelques rappels sont donnés à l'annexe B.

1.2.4 Equations différentielles ordinaires

On suppose connue la théorie élémentaire d'existence et d'unicité des solutions d'équations différentielles ordinaires. La plupart des résultats utilisés dans le cours sont rappelés dans l'annexe C, mais ne seront pas détaillés en cours.

De large portions de ce polycopié sont directement empruntées à des supports de cours rédigés par Hervé Le Dret et par Marie Postel (Sorbonne Université, LJLL).

2. En fait, on a seulement besoin de se rappeler du cas $n = 1$, manifestement.

Chapitre 2

Notion de schéma numérique pour les EDO

2.1 Equations différentielles ordinaires : rappels

Dans cette section, nous rappelons la définition d'une équation différentielle ordinaire, la notion de problème de Cauchy et un premier algorithme de résolution numérique (la méthode d'Euler). L'analyse théorique des EDO a été faite en L2 – les résultats principaux sont rappelés Appendice C.

2.1.1 Définition

Dans tout ce qui suit, I est un intervalle ouvert de $\mathbb{R}$, de la forme $I =]0, T[$ le plus souvent, avec $T > 0$, plus généralement parfois de la forme $]t_0, T[$ avec $T > t_0$, mais la généralité supplémentaire apportée par ce t_0 par rapport à $t_0 = 0$ n'est qu'apparente. On note $\bar{I} = [0, T]$ l'intervalle fermé correspondant. Soit $m \geq 1$ un nombre entier. Étant donnée une fonction continue f définie sur $\bar{I} \times \mathbb{R}^m$ et à valeurs dans $\mathbb{R}^m$, donc qui a tout couple (t, y) avec $t \in \bar{I}$ et $y \in \mathbb{R}^m$ associe un vecteur $f(t, y) \in \mathbb{R}^m$, on s'intéresse au problème suivant : trouver les fonctions $y : \bar{I} \mapsto \mathbb{R}^m$ dérivables sur I, qui satisfont

$$\forall t \in I, \quad y'(t) = f\big(t, y(t)\big). \tag{2.1.1}$$

On rappelle qu'une fonction $y \colon \bar{I} \to \mathbb{R}^m$ est dérivable en un point $t \in I$ si et seulement si toutes ses fonctions composantes, qui sont des fonctions définies sur $\bar{I}$ à valeurs réelles $t \mapsto y_i(t)$, $i = 1, \ldots, m$, sont dérivables au point t. Le vecteur dérivé $y'(t)$ a alors pour composantes les dérivées $y_i'(t)$ des composantes de y.

Pour ce qui concerne la notation des dérivées, on utilisera de façon essentiellement interchangeable $y', y'', \ldots, y^{(n)}$, $\frac{dy}{dt}, \frac{d^2y}{dt^2}, \ldots, \frac{d^ny}{dt^n}$. Dans certains contextes, comme celui de la dynamique des systèmes de points matériels, on sera parfois amenés à noter traditionnellement $\dot{y}$ et $\ddot{y}$ les dérivées première et seconde de y par rapport à t. [1]

On dit que le problème (2.1.1) est un *système d'équations différentielles ordinaires* ou *système d'EDO* ou *EDO* [2] tout court, du premier ordre, car seule la dérivée première de

1. Alors prononcées " y point " et " y deux points " ou " y point point ".

2. Abréviation que l'on utilisera plus ou moins systématiquement dans la suite pour éviter d'écrire l'expression système d'équations différentielles ordinaires.

y par rapport à t y apparaît. Dans le cas où $m = 1$, mais pas uniquement, on parle simplement d'équation différentielle ordinaire, ou encore parfois plus rapidement d'équation différentielle. La fonction f est appelée le *second membre de l'EDO* ou encore *fonction second membre de l'EDO*.

L'égalité (2.1.1) est pour chaque t une égalité entre deux vecteurs de $\mathbb{R}^m$. On peut l'écrire composante par composante sous la forme

$$y_i'(t) = f_i\big(t, y_1(t), y_2(t), \dots, y_m(t)\big), \tag{2.1.2}$$

où $f_i : \bar{I} \times \mathbb{R}^m \to \mathbb{R}$ désigne la i-ème composante de la fonction f, soit donc un jeu de m équations scalaires pour $i = 1$ jusqu'à m, à satisfaire par le m-uplet des fonctions scalaires inconnues $y_i : \bar{I} \to \mathbb{R}$.

Une solution de l'EDO est donc une courbe paramétrée dans $\mathbb{R}^m$, $t \mapsto y(t)$, qui se débrouille pour faire en sorte que sa dérivée, ou son vecteur tangent au point t, vaut exactement ce que vaut le second membre f en t et au point $y(t)$. Il faut penser à la fonction second membre comme à un champ de vecteurs dépendant du temps : à chaque instant t, on a en tout point y de $\mathbb{R}^m$ la donnée d'un vecteur $f(t, y)$ de $\mathbb{R}^m$.

Si l'on pense à une interprétation cinématique plutôt que géométrique, où t représente le temps et $y(t)$ la position d'un point mobile dans $\mathbb{R}^m$ à l'instant t, alors $y'(t)$ représente la vitesse instantanée du mobile à l'instant t. Cette vitesse est donc égale à $f(t, y(t))$. C'est d'ailleurs de cette façon que les EDO sont apparues historiquement, en lien avec la mécanique du point matériel.

Prenons l'exemple d'un baton qu'on jette dans une rivière. En première approximation, il est transporté par l'eau à la vitesse du courant. Sa position le long de l'axe de la rivière est donc régie par l'EDO

$$x'(t) = f(t, x(t))$$

où la fonction $f(t, x)$ donne la vitesse du courant à la position x et au temps t. On peut imaginer que cette vitesse dépend de x si le lit de la rivière est irrégulier (la vitesse sera d'autant plus grande que la largeur sera petite), et du temps (la vitesse aumente avec le débit, et varie donc en fonction des précipitations). La figure 2.1 donne un exemple de champ de vecteurs correspondant à une variation sinusidale en x et en t. La longueur des flèches est proportionnelle à la force du courant en un point x et un temps t. Les trois courbes correspondent à trois solutions de l'EDO pour trois positions initiales différentes.

Pour approfondir et visualiser la notion de champ de vecteur, on pourra visionner la vidéo disponible sur le site http://www.chaos-math.org/

2.1.2 Le problème de Cauchy

On sait bien qu'en général, si l'on se donne une EDO raisonnable, celle-ci va avoir une infinité de solutions. Si l'on admet que de nombreux phénomènes naturels ou artificiels sont régis par des équations différentielles, comment la nature choisit-elle la solution qui se réalise parmi cette infinité possible ? En d'autre termes, comment obtient-on de l'unicité ? Cette question est liée à celle du *déterminisme* de la physique classique et on l'exprime à l'aide de la notion de *problème de Cauchy* [3].

3. Augustin Louis, baron Cauchy, 1789–1857.

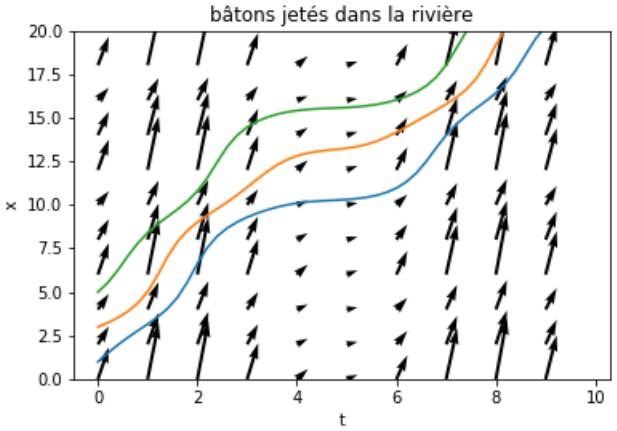


FIGURE 2.1 – Champ de vecteurs du courant dans une rivière $f(x,t) = (2 + 0.9\cos(x))(1 + 0.9\sin(t))$ et positions de trois bâtons lancés à $x_0 = 1$, $x_0 = 3$ et $x_0 = 5$ à l'instant initial (voir le notebook cours_1.ipynb).

Définition 2.1.1 *On appelle* problème de Cauchy*, la conjonction d'une EDO* (2.1.1) *et d'une condition additionnelle, la* donnée initiale*,*

$$\begin{cases} y'(t) = f(t, y(t)), \text{ pour tout } t \in I, \\ y(t_0) = y_0, \end{cases} \tag{2.1.3}$$

où $t_0 \in \bar{I}$ est un instant donné et y_0 un vecteur donné de $\mathbb{R}^m$.

On considérera le plus souvent pour simplifier que $I =]0, T[$ et que $t_0 = 0$, ce qui ne nuit pas à la généralité [4]. La condition initiale est donc

$$y(0) = y_0.$$

Notons que pour le problème à N corps présenté plus haut, la donnée initiale consiste à se donner les N positions initiales des corps célestes, $q_i(0) \in \mathbb{R}^3$, $i = 1, \ldots, N$, et pour chacun d'entre eux, sa vitesse initiale $\dot{q}_i(0) \in \mathbb{R}^3$, $i = 1, \ldots, N$, donc au total $6N$ conditions scalaires.

Ce problème de Cauchy a-t-il une solution, et si oui, est-elle unique ? Pour répondre à cette question fondamentale, il serait évidemment très agréable de pouvoir exhiber la solution à l'aide d'une formule ne faisant intervenir que des fonctions bien connues comme les fonctions polynomiales, les fractions rationnelles, les exponentielles, les logarithmes, les fonctions trigonométriques et trigonométriques hyperboliques directes et inverses. [5] C'est ce que l'on appelle *résoudre analytiquement* ou *intégrer* l'EDO. Malheureusement, c'est rarement possible. Il s'agit bien d'une impossibilité de principe, et non pas d'une impossibilité d'incompétence...

4. Pourquoi d'ailleurs ?

5. Les fonctions que l'on obtient en combinant les éléments de cette liste par les opérations algébriques usuelles et la composition forment ce que l'on appelle les *fonctions élémentaires*. Si l'on est plus savant, on peut aussi en faire intervenir d'autres plus sophistiquées, appelées *fonctions spéciales*, souvent définies d'ailleurs comme solutions de telle ou telle EDO.

Le cas dit " à variables séparées "

En dimension $m = 1$, il existe une classe d'équations différentielles que l'on peut avoir une chance d'intégrer au sens ci-dessus [6]. Il s'agit des EDO dont les fonctions second membre f sont à *variables séparées*, c'est-à-dire sous forme d'un produit d'une fonction de t par une fonction de y,

$$f(t, y) = g(t)h(y),$$

où g et h sont deux fonctions d'une seule variable. Il s'agit manifestement d'une condition extrêmement restrictive. La plupart des EDO scalaires ne sont pas à variables séparées. Dans ce cas particulier, l'équation différentielle se met sous la forme

$$\frac{y'(t)}{h(y(t))} = g(t),$$

(en supposant que l'on n'est pas en train de diviser par 0, [7] on procède ici un peu " à la physicienne " comme on dit dans les cercles mathématiques). Au membre de gauche, on reconnaît la dérivée de la fonction $t \mapsto R(y(t))$ où R désigne une primitive de $1/h$. Au membre de droite, on a une fonction g à intégrer, notons G une de ses primitives sur $[0, T]$. Il vient donc, pour tout $t \in [0, T]$,

$$R(y(t)) - R(y(0)) = \int_0^t \frac{y'(t)}{h(y(t))}\, dt = \int_0^t g(t)\, dt = G(t) - G(0),$$

soit

$$R(y(t)) = R(y_0) + G(t) - G(0).$$

Supposons que la fonction R soit bijective sur l'intervalle auquel appartient $y(t)$ (nous procédons toujours à la physicienne) et y admette donc une fonction réciproque notée R^{-1}. On en déduit alors que

$$y(t) = R^{-1}\big(R(y_0) + G(t) - G(0)\big).$$

On conclut donc, avec un léger manque de rigueur, que s'il y a une solution au problème de Cauchy, alors celle-ci est unique et donnée par la formule ci-dessus.

À ce stade-là, il faut encore vérifier que la formule en question donne bien une solution du problème de Cauchy, car ce n'est absolument pas garanti par ce qui précède, même si l'on fait abstraction des irrégularités qui ont émaillé le parcours. Pour la donnée initiale, c'est assez évident :

$$y(0) = R^{-1}\big(R(y_0) + G(0) - G(0)\big) = R^{-1}\big(R(y_0)\big) = y_0.$$

Pour l'EDO, il faut savoir dériver une fonction réciproque et une fonction composée, ce qui ne devrait pas poser de problème de principe à ce niveau d'études.

$$\begin{aligned} y'(t) &= \frac{d}{dt}\big(R^{-1}\big(R(y_0) + G(t) - G(0)\big)\big) \\ &= \frac{1}{R'\big(R^{-1}\big(R(y_0) + G(t) - G(0)\big)\big)} \times G'(t) \\ &= g(t)h(y(t)) = f(t, y(t)), \end{aligned}$$

6. Bien que cela ne soit pas gagné d'avance.

7. Ce qui est rigoureusement interdit en mathématiques.

vu que $R' = \frac{1}{h}$ et $G' = g$ (et $R(y_0) - G(0)$ est une constante). On a donc trouvé, en croisant un peu (en fait beaucoup) les doigts, l'unique solution du problème de Cauchy dans ce cas simple.

Bien sûr, pour que l'EDO soit résoluble analytiquement, encore faut-il que les deux calculs de primitives soient faisables analytiquement, c'est-à-dire s'expriment avec des fonctions élémentaires. C'est pourquoi l'affaire n'était pas gagnée d'avance. Par exemple, les primitives de la fonction $t \mapsto e^{-t^2}$ ne peuvent pas s'exprimer à l'aide des fonctions élémentaires. On a la même difficulté pour la fonction réciproque d'ailleurs. Mais enfin, même si le calcul explicite n'est pas possible, on a quand même ainsi une petite idée de la solution.

Il existe d'autres familles d'équations différentielles dont on peut calculer explicitement les solutions, par changement d'inconnues, ou bien quand il s'agit d'équations différentielles exactes. On pourra se reporter à ses cours de L1 ou bien au site web de l'université en ligne :

http://uel.unisciel.fr/mathematiques/eq_diff/eq_diff/co/eq_diff.html

La théorie élémentaire d'existence et unicité de solutions d'EDO a été vue en L2. Elle est reprise de façon assez détaillée en appendice de ce cours. L'objectif du cours est plutôt de présenter des méthodes d'approximation numérique de solutions d'EDO, qui donnent lieu à des algorithmes programmables sur ordinateur. Nous commençons par le premier tel algorithme introduit dans l'histoire : la méthode d'Euler (qui a servi à l'origine à justifier l'existence de solutions).

2.1.3 Approche historique par la méthode d'Euler

Dans ce paragraphe, nous allons décrire sans démonstration et sans être très précis, une approche pour l'existence d'une solution du problème de Cauchy. Cette approche annonce également la problématique du cours, celle de l'approximation numérique, mais utilisée ici pour montrer l'existence et l'unicité. C'est une approche *constructive*, puisqu'elle consiste à montrer la convergence d'une suite de solutions approchées connues explicitement et calculables numériquement vers une solution de (2.1.3).

Bien avant l'informatique et le calcul scientifique, il est apparu judicieux d'introduire des procédures de calcul approché pour les solutions d'équations différentielles. [8] Elles reposent sur la notion de discrétisation.

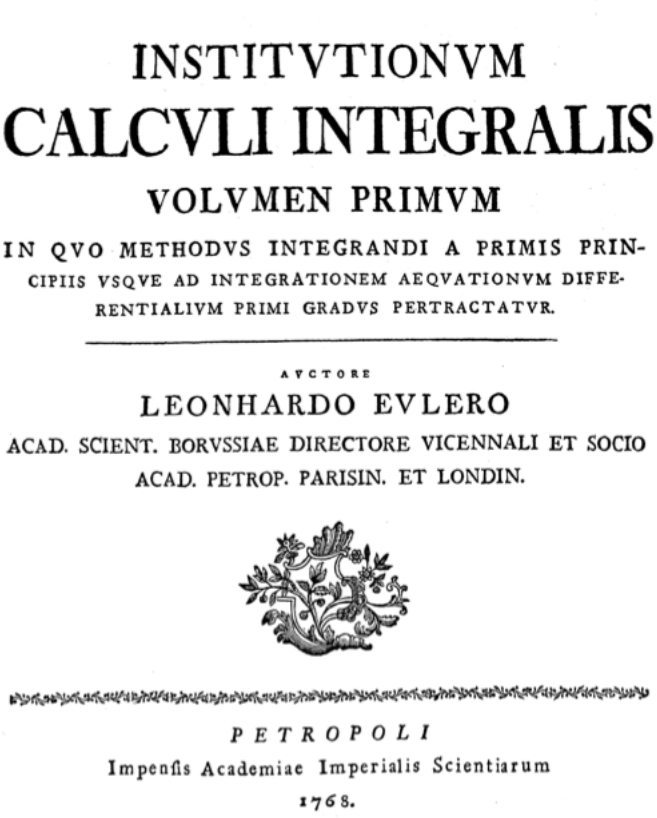
Soit n un entier strictement positif. On définit une subdivision de l'intervalle $[t_0, T]$

$$t_0 < t_1 < \ldots < t_{n-1} < t_n = T.$$

Soit une fonction $y(t)$ dérivable sur $[t_0, T]$. Si les points t_i, $i = 0, \ldots, n$, sont suffisamment rapprochés les uns des autres, on peut raisonnablement approcher la dérivée $y'(t_i)$ aux points t_i par le taux de variation de y, appelé également *quotient aux différences finies*

$$y'(t_i) \approx \frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i}.$$

8. La méthode d'Euler date de 1768 dans *Institutionum Calculi Integralis, Libri Prioris Pars Prima Sectio Secunda, Caput VII, De Integratione Æquationum Differentialum Per Approximationem.*



INSTITVTIONVM
CALCVLI INTEGRALIS
VOLVMEN PRIMVM
IN QVO METHODVS INTEGRANDI A PRIMIS PRINCIPIIS VSQVE AD INTEGRATIONEM AEQVATIONVM DIFFERENTIALIVM PRIMI GRADVS PERTRACTATVR.

AVCTORE
LEONHARDO EVLERO
ACAD. SCIENT. BORVSSIAE DIRECTORE VICENNALI ET SOCIO
ACAD. PETROP. PARISIN. ET LONDIN.

PETROPOLI
Impensis Academiae Imperialis Scientiarum
1768.

FIGURE 2.2 – La source.

En tant que telle, cette approximation un peu vague ne sert pas à grand-chose, puisque l'on ignore les valeurs $y(t_i)$. L'idée est de remplacer ces valeurs par des valeurs y_i, $i = 0, \dots, n$, qui soient calculables par récurrence en partant de y_0, la donnée initiale connue, et dont on espère qu'elles seront proches des valeurs exactes $y(t_i)$ si tout se passe bien. Pour cela, notant que $y'(t_i) = f(t_i, y(t_i))$, on imite l'EDO sous la forme

$$\frac{y_{i+1} - y_i}{t_{i+1} - t_i} = f(t_i, y_i).$$

Cette définition fonctionne clairement, puisqu'elle se réécrit comme

$$\begin{aligned} y_1 &= y_0 + (t_1 - t_0) f(t_0, y_0) \\ &\vdots \\ y_n &= y_{n-1} + (t_n - t_{n-1}) f(t_{n-1}, y_{n-1}). \end{aligned}$$

On l'a déjà dit, l'espoir est que ces valeurs calculables sont en fait des approximations des valeurs exactes, qui ne sont pas en général calculables, en un sens à préciser. La méthode est évidemment définie dans ce but, mais il n'est aucunement évident que ce but soit atteint à ce point de l'analyse...

On note $h_i = t_{i+1} - t_i$ et $h = \max_{i=0,\dots,n-1} h_i$. Le nombre h est appelé le *pas de la discrétisation*. La ligne brisée reliant les points $(t_i, y_i)_{i=0,\dots,n}$ est appelée le *polygone d'Euler*. C'est le graphe de la fonction affine par morceaux y_h définie par

$$y_h(t) = y_i + (t - t_i) f(t_i, y_i), \quad \text{pour } t \in [t_i, t_{i+1}]. \tag{2.1.4}$$

La Figure 2.3 représente ce polygone pour $n = 5$, pour la fonction $y(t) = t(1 - t)$ solution

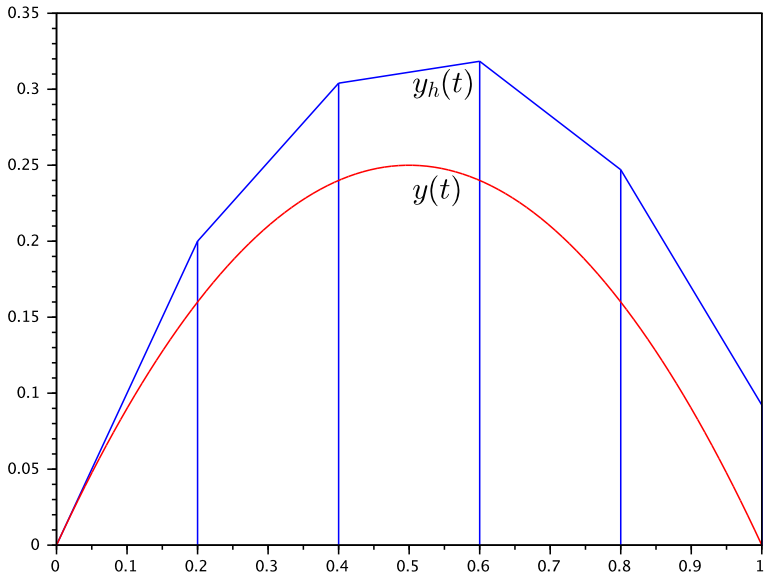


FIGURE 2.3 – Polygone d'Euler pour $y'(t) = 1 - 2t^2 - 2y(t)$, $T = 1$ et $n = 5$.

de $y'(t) = 1 - 2t^2 - 2y(t)$ et pour un pas uniforme. On s'attend à ce que y_h soit une approximation de la solution $y(t)$, convergeant vers $y(t)$ lorsque $n \to \infty$, ce qui implique $h \to 0$.

Exemple 2.1.1 Considérons le problème de Cauchy $y'(t) = ay(t)$ sur $[0, T]$, $y(0) = y_0$, dont la solution est $y(t) = e^{at} y_0$ et choisissons une subdivision uniforme $h_i = t_{i+1} - t_i = h = \frac{T}{n}$ pour simplifier. Il est facile de voir que, pour $T > 0$ fixé, la valeur approchée $y_i = y_h(t_i)$ est donnée par

$$y_h(t_i) = \left(1 + \frac{aT}{n}\right)^i y_0.$$

Si $y_0 = 0$, alors $y_h(t) = 0 = y(t)$ pour tout h et tout va bien. Supposons $y_0 \neq 0$ et posons $z_h(t) = \frac{y_h(t)}{y_0}$. Fixons t dans l'intervalle $[0, T]$. Si l'on pose $i = \left[\frac{t}{h}\right] = \left[n\frac{t}{T}\right]$, où le crochet désigne la partie entière, alors on voit que $t_i \leq t < t_{i+1}$, ce qui implique que $|t - t_i| \leq \frac{T}{n}$. Naturellement, cet indice i dépend de t et de h, même si on ne l'écrit pas explicitement. Par conséquent, en prenant n assez grand pour que $1 + \frac{aT}{n} > 0$, comme $z_h(t) = (1 + a(t - t_i)) z_h(t_i)$,

on a

$$\begin{aligned}\ln(z_h(t)) &= \ln(1 + a(t - t_i)) + \left[n\frac{t}{T}\right]\ln\left(1 + \frac{aT}{n}\right)\\ &= O\left(\frac{1}{n}\right) + \left[n\frac{t}{T}\right]\frac{aT}{n} + O\left(\frac{1}{n}\right)\\ &= at + O\left(\frac{1}{n}\right).\end{aligned}$$

En effet, comme $n\frac{t}{T} - 1 \leq \left[n\frac{t}{T}\right] \leq n\frac{t}{T}$, on voit que $at - \frac{aT}{n} \leq \left[n\frac{t}{T}\right]\frac{aT}{n} \leq at$ pour $a \geq 0$ et un encadrement analogue pour $a < 0$. On en déduit que $y_h(t)$ tend vers $y(t)$ quand $n \to +\infty$. Avec un peu plus de soin, on établit que la convergence est uniforme. ♢

L'exemple ci-dessus n'est pas très éclairant, puisque c'est un cas où l'on sait que la solution existe, on a même une formule explicite pour celle-ci. On peut en fait montrer la convergence du schéma d'Euler, voir Figure 2.4, dans le cas général modulo quelques hypothèses supplémentaires. On ne va pas les donner ici. [9] Sous ces hypothèses, la suite de fonctions affines par morceaux y_h converge uniformément vers une fonction y qui se trouve être dérivable (ce que les y_h ne sont pas) et solution du problème de Cauchy.

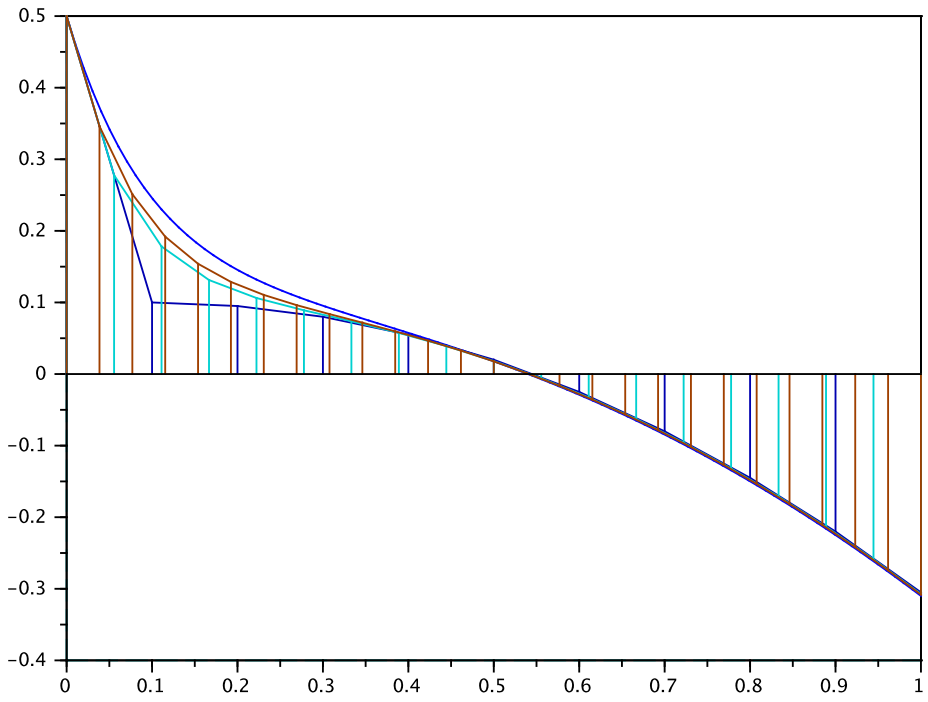


FIGURE 2.4 – Convergence des polygones d'Euler pour le problème de Cauchy $y'(t) = 1 - 5t^2 - 10y(t)$, $y(0) = 0,5$. On a tracé la solution exacte en bleu et trois polygones de couleurs différentes correspondant respectivement à $n = 10$ (bleu foncé), 18 (cyan) et 26 (brun).

9. On aura cette preuve par des voies détournées par la suite : existence de la solution d'abord par un autre biais, puis convergence de la méthode d'Euler, parmi bien d'autres méthodes d'approximation.

Retenons que ce n'est pas une méthode très performante du point de vue numérique, mais que c'est la plus simple de toutes et qu'à ce titre elle reste utile pour appréhender rapidement le comportement des solutions.

Ceci constitue le premier exemple d'approximation numérique d'une EDO, par la méthode d'Euler. Nous allons maintenant généraliser le principe pour définir un grand nombre d'autres *méthodes d'approximation numérique*, que nous appellerons aussi des *schémas numériques*. Nous aborderons ensuite l'étude mathématique de ces schémas avec en ligne de mire leur *convergence* vers la solution du problème de Cauchy de départ. Cette étude portera sur deux notions essentielles, la *stabilité* du schéma et sa *consistance*. On parlera aussi d'*estimation d'erreur* et l'on appliquera tout cela à plusieurs grandes familles de schémas numériques.

Pour mettre en oeuvre (ou juste développer) certains de ces schémas, nous serons confrontés à des problèmes qui ne sont pas directement reliés aux EDO, mais qui font partie de l'analyse numérique au sens large – tels l'analyse numérique matricielle, les algorithmes de résolution de problèmes non linéaires, la théorie de l'approximation ou encore l'intégration numérique de fonctions.

2.2 La notion de schéma numérique

On se donne un problème de Cauchy générique de la forme (2.1.3) que l'on supposera satisfaire de bonnes hypothèses assurant existence, unicité et régularité de la solution. Dans l'hypothèse réaliste où il est impossible d'en donner une solution analytique, si l'on veut disposer d'informations quantitatives sur la solution, il faut donc définir des procédés d'approximation effectivement calculables, à la main dans le passé, sur ordinateur aujourd'hui. Cela implique en particulier que de tels procédés ne fassent intervenir qu'un nombre fini d'inconnues scalaires, alors qu'une fonction fait naturellement intervenir une infinité non dénombrable de valeurs scalaires.

On commence donc par discrétiser, c'est-à-dire remplacer du continu par du discret, l'intervalle $I = [0, T]$ en y plaçant $N + 1$ points $t_n = nh$, $n = 0, \dots, N$, appelés *points de discrétisation*, uniformément espacés de $h = T/N$ (h est appelé *pas de la discrétisation*), pour un entier $N > 0$ donné. En particulier $t_0 = 0$ est l'instant initial et $t_N = T$ l'instant final. Le cas d'une discrétisation à pas variable $h_n = t_{n+1} - t_n$, ajusté de manière adaptative pour optimiser la précision du schéma sera traité ultérieurement au paragraphe 8.2.

L'approximation numérique du problème de Cauchy consiste à construire une suite indexée par $N \in \mathbb{N}^*$ de valeurs $y_0^N, \dots y_N^N$ censées approcher les valeurs exactes de la solution aux points de discrétisation, $y(t_0), \dots, y(t_N)$, en un sens que l'on précisera plus loin. Comme $y(t_0) = y_0$ est connu, on prendra (presque) toujours $y_0^N = y_0$ et, même si cela peut créer de l'ambiguïté, on notera généralement y_n^N par y_n (mais il faut garder à l'esprit la dépendance par rapport à N).

Un *schéma numérique* est la donnée d'une telle construction, nous en verrons de nombreux exemples. Notons que dans le contexte de l'approximation numérique, on parle couramment de " résoudre " l'équation avec un schéma, alors qu'on ne fait en réalité que calculer une approximation nécessairement entachée d'erreur de la solution. Le vocabulaire est correct par contre si l'on parle d'un schéma convergent en tant que procédé abstrait d'approximation. Après tout, mis à part les objets construits en un nombre fini d'étapes à partir des nombres entiers, tous les objets de l'analyse sont définis par des approximations diverses et variées.

Dans la pratique, par contre, on ne peut pas faire tendre N vers l'infini, mais on n'effectue les calculs que pour une ou un petit nombre de valeurs de N, d'où les guillemets entourant le mot résoudre plus haut.

Il y a plusieurs façons de procéder pour construire des schémas numériques.

1. On écrit l'équation (2.1.1) à l'instant t_n (ou à un instant de discrétisation voisin),

$$y'(t_n) = f(t_n, y(t_n)),$$

relation exactement satisfaite. On remplace alors la dérivée $y'(t_n)$ du membre de gauche par un quotient aux différences, obtenu généralement en utilisant des développements de Taylor faisant intervenir les valeurs exactes de la fonction inconnue y à des instants de discrétisation voisins de t_n. Dans un deuxième temps, on remplace les valeurs exactes de la fonction inconnue dans le quotient aux différences et dans le membre de droite par les fameuses approximations, encore potentielles à ce stade. Voici quelques exemples :

(a) L'approximation

$$y'(t_n) \approx \frac{y(t_{n+1}) - y(t_n)}{h} \tag{2.2.1}$$

remplace la dérivée par un quotient aux différences contenant les valeurs de y en deux points de discrétisation. C'est une approximation qui semble raisonnable quand h est petit. Le deuxième temps de la construction consiste à remplacer ces valeurs exactes par des valeurs approchées potentielles dans les deux membres de l'équation

$$\frac{y_{n+1} - y_n}{h} = f(t_n, y_n),$$

ce qui conduit au schéma

$$y_{n+1} = y_n + hf(t_n, y_n), \tag{2.2.2}$$

qui est appelé *schéma d'Euler explicite*, ou schéma d'Euler progressif, ou plus simplement schéma d'Euler.

Attention, une erreur trop fréquente est de confondre $y(t_n)$ et y_n ! Si l'on pouvait confondre les deux, il n'y aurait aucune raison de se casser la tête à définir des schémas numériques...

Le schéma d'Euler (2.2.2) se présente sous la forme d'une récurrence vectorielle, définissant une suite finie de vecteurs $y_0, y_1, \ldots, y_N$ de $\mathbb{R}^m$, dont il faut assurer l'initialisation. On prendra donc pour y_0^N (on note ici l'exposant N implicite parce qu'il est utile de le voir, on l'oubliera ensuite [10]) soit la valeur initiale exacte y_0, si celle-ci est connue, soit une approximation de y_0 dépendant donc de N. À partir de là, la récurrence se déroule sans accroc.

Jusqu'ici, il ne s'agit que d'une recette, certes a priori raisonnable, mais dont on n'a aucune garantie qu'elle fournisse bien ce que l'on espère, à savoir des approximations des valeurs exactes $y(t_n)$.

Cette méthode est dite *explicite* car y_{n+1} est donnée explicitement par une formule connue appliquée à t_n et y_n. Il n'y a pas d'équation supplémentaire à résoudre pour l'obtenir. Plus généralement, si la valeur de y_{n+1} est donnés explicitement en fonction de certaines des valeurs calculées aux instants antérieurs, $y_0, \ldots, y_n$, qui ont déjà été calculées précédemment lors de la mise en œuvre du schéma, on dit que l'on a affaire à un schéma explicite.

10. Voir remarque plus haut sur cet exposant implicite.

(b) L'approximation

$$y'(t_{n+1}) \approx \frac{y(t_{n+1}) - y(t_n)}{h} \tag{2.2.3}$$

conduit de la même façon à partir de $y'(t_{n+1}) = f(t_{n+1}, y(t_{n+1}))$ au schéma *implicite*

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}). \tag{2.2.4}$$

Cette fois-ci, pour calculer y_{n+1}, il faut résoudre une équation, en général non linéaire (sauf si $y \to f(t, y)$ est affine en y), d'où le qualificatif d'implicite. Il n'est pas évident que cette équation ait une solution ni que celle-ci soit unique, on verra plus loin sous quelles conditions ceci est vrai. Quand on y adjoint une condition initiale comme précédemment, on obtient donc une autre récurrence vectorielle. Le schéma (2.2.4) est appelée *schéma d'Euler implicite* ou rétrograde.

(c) L'approximation centrée

$$y'(t_n) \approx \frac{y(t_{n+1}) - y(t_{n-1})}{2h} \tag{2.2.5}$$

conduit au schéma

$$y_{n+1} = y_{n-1} + 2hf(t_n, y_n), \tag{2.2.6}$$

appelé *schéma leapfrog* (saute-mouton). Ce schéma est explicite, mais c'est une récurrence à deux pas. Sa mise en ?uvre demande par conséquent de connaître y_0, disons la donnée de Cauchy, et y_1 qui n'est pas une donnée du problème, et qu'il faut donc se procurer autrement d'une façon ou d'une autre (forcément avec un schéma à un pas, mais convenablement choisi).

(d) On peut aussi utiliser des combinaisons de plusieurs approximations de $y'(t_n)$. Par exemple une combinaison linéaire de (2.2.1), (2.2.3) et (2.2.5) conduit au schéma

$$\alpha(y_{n+1} - y_n) + \beta(y_{n+1} - y_n) + \gamma(y_{n+1} - y_{n-1}) = \alpha hf(t_n, y_n) + \beta hf(t_{n+1}, y_{n+1}) + 2\gamma hf(t_n, y_n),$$

ou toute autre variante raisonnable. Les paramètres α, β et γ sont à choisir au mieux pour que le schéma ait les meilleures propriétés d'approximation possibles.

2. Une deuxième manière de construire un schéma numérique utilise les techniques d'intégration numérique ou de quadrature (voir [4] chapitre 2). En intégrant l'EDO (2.1.1) sur un intervalle $[t_n, t_{n+1}]$, on obtient

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} y'(s)\, ds = \int_{t_n}^{t_{n+1}} f\big(s, y(s)\big)\, ds. \tag{2.2.7}$$

On peut donc calculer $y(t_{n+1})$ connaissant $y(t_n)$, pourvu que l'on sache aussi calculer l'intégrale

$$I_n = \int_{t_n}^{t_{n+1}} f\big(s, y(s)\big)\, ds. \tag{2.2.8}$$

Évidemment, on ne sait pas calculer cette intégrale puisqu'on ne connait pas la solution exacte [11]. Par contre, on peut l'approcher à l'aide d'une des multiples formules de quadrature numérique qui existent déjà dans la nature, voir la Figure 2.9 pour l'interprétation

11. Et même si on la connaissait... mais est-ce que l'on n'est pas en train de tourner un peu en rond ?

géométrique des méthodes les plus élémentaires d'intégration numérique. On procédera ici aussi en deux temps : approximation de l'intégrale par une formule de quadrature faisant intervenir les valeurs exactes en des points de discrétisation (si possible), puis remplacement de toutes les valeurs exactes par des valeurs approchées potentielles. Dans ce qui suit, on n'écrit plus la donnée initiale.

(a) Approchons l'intégrale (2.2.8) par la méthode des rectangles à gauche (avec un seul rectangle).

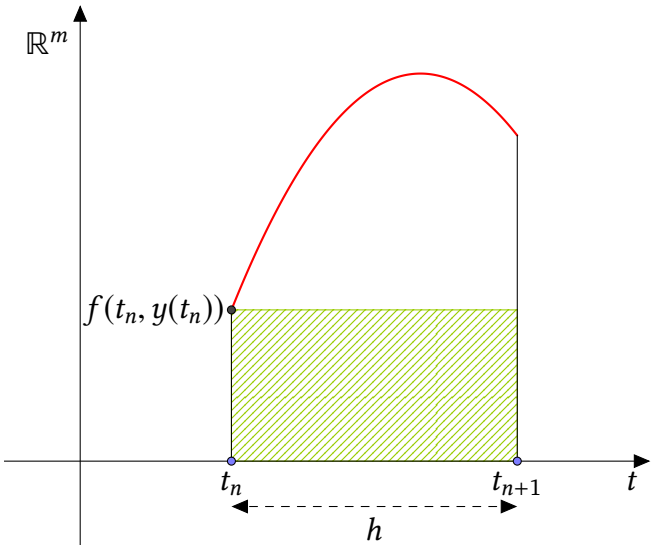


FIGURE 2.5 – À gauche, la fonction $t \mapsto f(t, y(t))$ étant tracée en rouge.

On obtient donc

$$I_n \approx h f\big(t_n, y(t_n)\big).$$

Remplaçant maintenant dans (2.2.7) et dans l'approximation précédente les valeurs exactes $y(t_n)$ par des valeurs que l'on espère approchées y_n, il vient

$$y_{n+1} - y_n = h f(t_n, y_n),$$

soit

$$y_{n+1} = y_n + h f(t_n, y_n),$$

On retrouve ainsi le schéma d'Euler explicite, déception (de courte durée). Dans le cas (on le rappelle pas très intéressant) où $f(t, y) = g(t)$ ne dépend pas de y, le schéma redonne la méthode des rectangles à gauche composée pour approcher une intégrale, voir Figure 2.9. En effet, dans ce cas et avec $y_0 = 0$, on a $y(t) = \int_0^t g(s)\,ds$ et $y_{n+1} = y_n + hg(t_n)$, d'où $y_n = h \sum_{k=0}^{n-1} g(t_k)$.

(b) Le même procédé appliqué avec la méthode des rectangles à droite, toujours avec un seul rectangle, donne

$$y_{n+1} = y_n + h f(t_{n+1}, y_{n+1}).$$

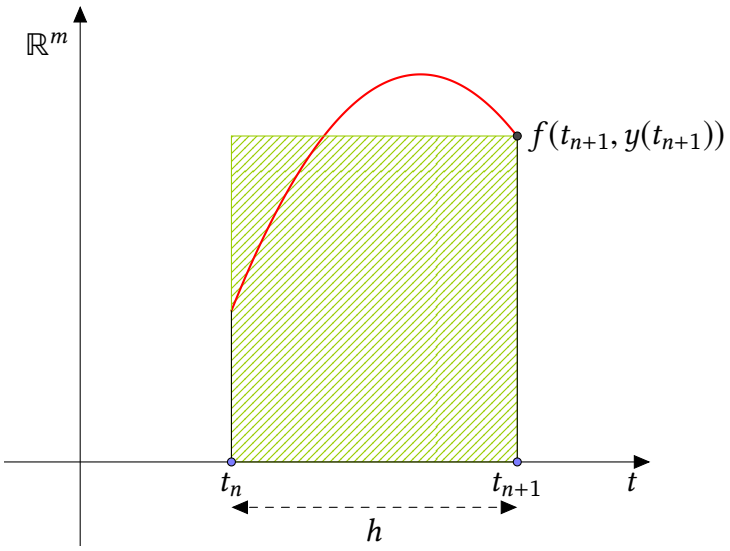


FIGURE 2.6 – À droite.

C'est le schéma d'Euler implicite, toujours rien de nouveau jusque là. Dans le cas où $f(t, y) = g(t)$ ne dépend pas de y, le schéma redonne la méthode des rectangles à droite composée pour approcher une intégrale, voir Figure 2.9, $y_n = h \sum_{k=1}^{n} g(t_k)$.

(c) Si l'on approche (2.2.8) par la méthode du point milieu, [12] on trouve l'approximation

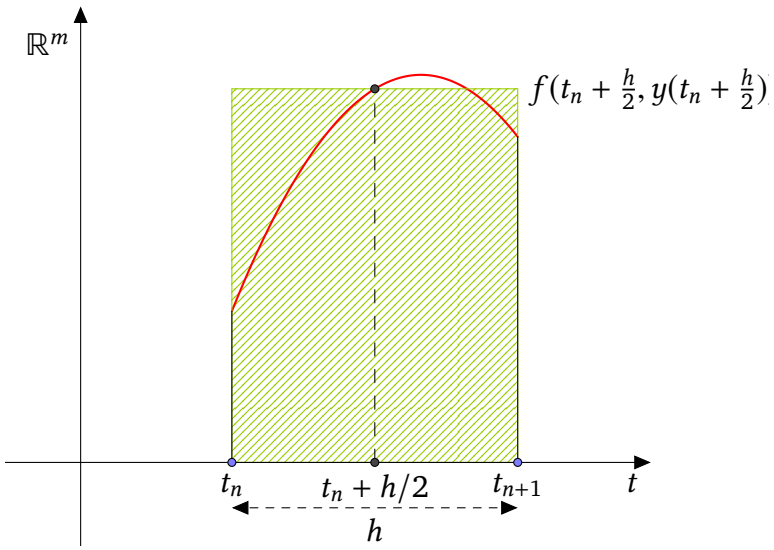


FIGURE 2.7 – Au milieu.

12. En tant que méthode de quadrature, la méthode du point milieu, qui est aussi une méthode des rectangles, est nettement plus précise que les méthodes des rectangles à gauche et à droite quand $h \to 0$. On peut espérer que le schéma numérique que l'on en tire soit plus performant que les schémas d'Euler, ce qui en fait se révélera bien être le cas.

$$I_n \approx hf\left(t_n + \frac{h}{2}, y\left(t_n + \frac{h}{2}\right)\right).$$

Cela ne permet pas de construire directement un schéma numérique suivant la recette habituelle, puisque la valeur $y(t_n + h/2)$ ne correspond pas à un point de discrétisation. Néanmoins, on peut reprendre pour cette valeur intermédiaire l'idée du schéma d'Euler et dire que

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

Remplaçant les valeurs exactes $y(t_n)$ par des valeurs approchées y_n, on obtient ainsi un schéma appelé *schéma d'Euler modifié* (ou schéma du point milieu),

$$y_{n+1} - y_n = hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f(t_n, y_n)\right),$$

soit

$$y_{n+1} = y_n + hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f(t_n, y_n)\right), \tag{2.2.9}$$

qui est un schéma explicite. On voit que les choses commencent à se compliquer un peu, avec des compositions de la fonction f avec elle-même qui ne sont pas exactement intuitives. Dans le cas où $f(t, y) = g(t)$ [...] la méthode du point milieu composée [...] Figure 2.9, $y_n = h\sum_{k=0}^{n-1} g(t_k + \frac{h}{2})$.

(d) En approchant (2.2.8) par la méthode des trapèzes (avec un seul trapèze, cf. Figure 2.8), on obtient le schéma

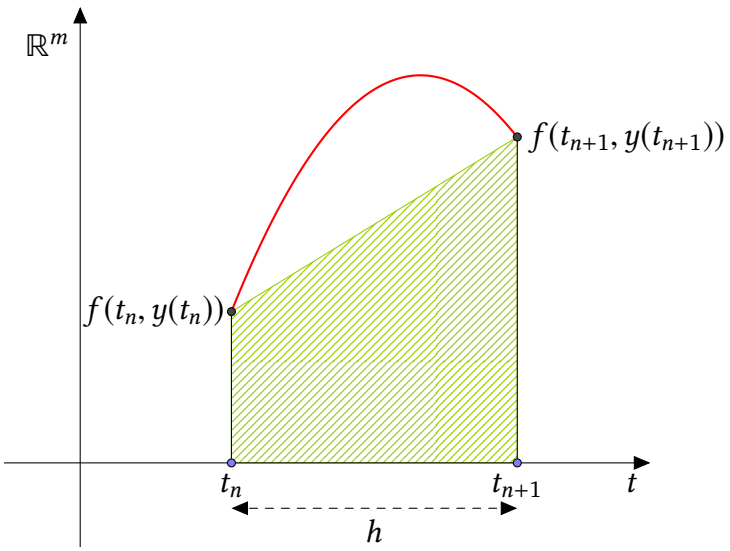


FIGURE 2.8 – Avec un trapèze.

$$y_{n+1} = y_n + \frac{h}{2}\big(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\big), \tag{2.2.10}$$

appelé *schéma de Crank-Nicolson* [13]. C'est un schéma implicite. Dans le cas où $f(t, y) = g(t)$ [...] la méthode des trapèzes composée [...] Figure 2.9, $y_n = h\big(\frac{1}{2}g(t_0) + \sum_{k=1}^{n-1} g(t_k) + \frac{1}{2}g(t_n)\big)$.

À chaque discrétisation de (2.2.8), on peut ainsi associer un schéma numérique pour le problème de Cauchy.

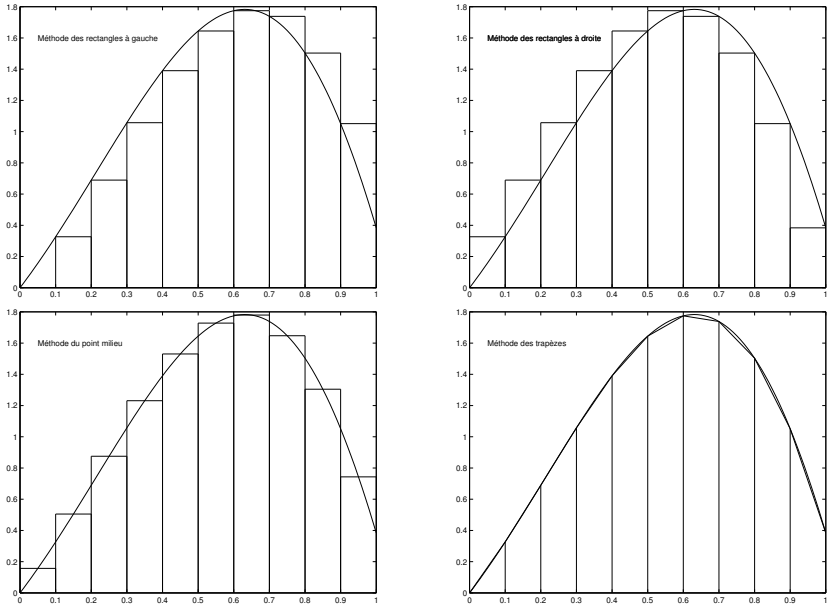


FIGURE 2.9 – Illustration des formules de quadratures utilisées dans le contexte de l'intégration numérique : on approche l'aire située sous la courbe par la somme des aires des rectangles (ou des trapèzes).

3. La dernière catégorie de schémas que l'on abordera dans ce cours est celle des schémas symplectiques, destinés spécifiquement aux systèmes hamiltoniens comme ceux de la mécanique céleste. [14]

Les schémas précédemment introduits, Euler, leapfrog, Crank-Nicolson, sont d'un usage généraliste. On peut les utiliser pour n'importe quelle EDO. Néanmoins, pour certaines familles d'EDO qui possèdent une structure supplémentaire, il peut se faire que des schémas numériques spécialisés soient plus indiqués, en particulier si ces schémas sont adaptés pour tenter de prendre en compte la structure supplémentaire en question. C'est le cas pour les systèmes hamiltoniens. On rappelle que ces systèmes différentiels s'écrivent sous la forme générique

$$\begin{cases} \dot{q} = \dfrac{\partial H}{\partial p}, \\ \dot{p} = -\dfrac{\partial H}{\partial q}, \end{cases} \tag{2.2.11}$$

où le hamiltonien H est une fonction des variables de position $q = (q_i)_{i=1,\dots,m}$ et de moment ou impulsion $p = (p_i)_{i=1,\dots,m}$ et où l'on a écrit $\frac{\partial H}{\partial p}$ à la place de $\nabla_p H$ etc. Ces systèmes ont la propriété de conserver le hamiltonien au cours du temps. Qui plus est, ils conservent également les volumes $2m$-dimensionnels au sens suivant. Il existe une notion de volume k-dimensionnel sur tout $\mathbb{R}^k$, vol_k, qui généralise naturellement la longueur pour $k = 1$, l'aire pour $k = 2$ et le volume pour $k = 3$. Le volume du parallélotope défini par k vecteurs de $\mathbb{R}^k$ est simplement la valeur absolue de leur déterminant. On imagine comment définir le volume

13. John Crank, 1916–2006 ; Phyllis Nicolson, 1917–1968.

14. La géométrie cachée derrière les systèmes hamiltoniens s'appelle la géométrie symplectique.

d'une partie de $\mathbb{R}^k$ en la découpant en tous petits parallélotopes... [15] Pour tout $t \geq 0$, on note

$$\begin{aligned} \varphi_t : \quad \mathbb{R}^{2m} &\longrightarrow \mathbb{R}^{2m} \\ (q_0, p_0) &\longmapsto \varphi_t(q_0, p_0) = (q(t), p(t)), \end{aligned}$$

où (q, p) est la solution de (2.2.11) muni des conditions initiales

$$q(0) = q_0, \quad p(0) = p_0.$$

L'application φ_t s'appelle le *flot* de l'EDO. On admet le théorème de Liouville suivant : pour tout ouvert borné D de $\mathbb{R}^{2m}$ et tout t,

$$\mathrm{vol}_{2m}(D) = \mathrm{vol}_{2m}(\varphi_t(D)).$$

En particulier, pour $m = 1$, on a conservation de l'aire usuelle par le flot dans l'espace des phases $\mathbb{R}^2$. C'est la seule dimension où l'on puisse visualiser cette conservation.

Il est possible et intéressant de construire des schémas numériques de précision arbitrairement grande conservant à la fois les aires et le hamiltonien (en moyenne). L'exemple le plus simple s'obtient en modifiant très légèrement le schéma d'Euler, qui prend ici la forme

$$\begin{cases} q_{n+1} = q_n + h\dfrac{\partial H}{\partial p}(q_n, p_n), \\ p_{n+1} = p_n - h\dfrac{\partial H}{\partial q}(q_n, p_n), \end{cases}$$

en changeant juste un indice dans la deuxième ligne

$$\begin{cases} q_{n+1} = q_n + h\dfrac{\partial H}{\partial p}(q_n, p_n), \\ p_{n+1} = p_n - h\dfrac{\partial H}{\partial q}(q_{n+1}, p_n). \end{cases}$$

Le schéma obtenu s'appelle *schéma d'Euler symplectique*. Remarquons que, contrairement aux apparences, le schéma d'Euler symplectique reste explicite. On effet, on calcule d'abord q_{n+1} en fonction de (q_n, p_n) via la première ligne, puis p_{n+1} via la seconde. On a donc en fait $p_{n+1} = p_n - h\frac{\partial H}{\partial q}\big(q_n + h\frac{\partial H}{\partial p}(q_n, p_n), p_n\big)$.

Insistons bien une nouvelle fois sur le fait qu'aucune de ces constructions aussi tarabiscotée soit-elle ne garantit que les valeurs y_n calculées par un de ces schémas approchent bien les valeurs exactes $y(t_n)$. On a simplement défini ces valeurs de façon a priori raisonnable par rapport au problème de Cauchy considéré. Montrer a posteriori que ces méthodes fonctionnent est ce qu'on appelle effectuer *l'analyse numérique* de ces schémas. Ce que l'on fera d'ailleurs dans la suite.

Par curiosité, regardons ce que donnent les schémas définis plus haut dans le cas d'une équation scalaire linéaire $y'(t) = ay(t)$, c'est-à-dire $f(t, y) = ay$. On sait dans ce cas que la solution du problème de Cauchy n'est autre que $y(t) = e^{at}y_0$. Pour le schéma d'Euler, on obtient

$$y_{n+1} = y_n + hay_n = (1 + ha)y_n, \text{ d'où } y_n = (1 + ha)^n y_0.$$

La discussion de l'exemple 2.1.1 montre que l'on converge bien vers ce qu'on veut. Pour le schéma d'Euler implicite, on trouve

$$y_{n+1} = y_n + hay_{n+1}.$$

C'est bien une équation pour y_{n+1}, il se trouve que c'est l'un des rares cas où elle se résout facilement et explicitement. En effet, on en déduit que

$$y_{n+1} = \frac{y_n}{1 - ha}, \text{ d'où } y_n = \frac{y_0}{(1 - ha)^n},$$

à condition que h soit assez petit pour que $1 - ha \neq 0$, c'est-à-dire pour N assez grand. Pour la méthode d'Euler modifiée, on obtient

$$y_{n+1} = y_n + ha\left(y_n + \frac{h}{2}ay_n\right) = \left(1 + ha + \frac{h^2a^2}{2}\right)y_n, \text{ d'où } y_n = \left(1 + ha + \frac{h^2a^2}{2}\right)^n y_0.$$

15. La théorie de la mesure, c'est quand même un peu plus compliqué que cela.

Enfin pour le schéma de Crank-Nicolson, qui est implicite, il vient

$$y_{n+1} = y_n + \frac{h}{2}(ay_n + ay_{n+1}),\ \text{d'où}\ y_{n+1} = \frac{1+\frac{h}{2}a}{1-\frac{h}{2}a}y_n,\ \text{d'où}\ y_n = \left(\frac{1+\frac{h}{2}a}{1-\frac{h}{2}a}\right)^n y_0,$$

là aussi pour h assez petit. On peut voir en suivant les même lignes que dans l'exemple 2.1.1 que l'on converge bien aussi dans ces trois derniers cas vers ce que l'on veut, voir Figure 2.10. [16]

L'équivalent de ces exemples pour le schéma d'Euler symplectique correspond à l'hamiltonien le plus simple possible $H(q,p) = \frac{1}{2}p^2 + \frac{1}{2}q^2$. On obtient

$$q_{n+1} = q_n + hp_n \text{ et } p_{n+1} = p_n - hq_{n+1} = (1-h^2)p_n - hq_n,$$

c'est-à-dire

$$\begin{pmatrix} q_n \\ p_n \end{pmatrix} = \begin{pmatrix} 1 & h \\ -h & 1-h^2 \end{pmatrix}^n \begin{pmatrix} q_0 \\ p_0 \end{pmatrix}.$$

On voit que la matrice qui apparaît est de déterminant 1 et donc que l'application $\begin{pmatrix} q_0 \\ p_0 \end{pmatrix} \mapsto \begin{pmatrix} q_n \\ p_n \end{pmatrix}$ conserve les aires dans $\mathbb{R}^2$. Notons que dans ce cas, le schéma d'Euler (pas symplectique pour un sou) donne

$$q_{n+1} = q_n + hp_n \text{ et } p_{n+1} = p_n - hq_n,$$

soit

$$\begin{pmatrix} q_n \\ p_n \end{pmatrix} = \begin{pmatrix} 1 & h \\ -h & 1 \end{pmatrix}^n \begin{pmatrix} q_0 \\ p_0 \end{pmatrix},$$

avec une matrice qui n'est pas de déterminant 1.

2.3 Schémas numériques généraux

La forme générale d'un schéma numérique pour approcher les solutions d'un problème de Cauchy est la suivante

$$y_{n+1} = y_n + h\Phi(t_{n+p}, y_{n+p}, t_{n+p-1}, y_{n+p-1}, \dots, t_{n-q}, y_{n-q}, h),$$

où p et q sont des entiers relatifs tels que $p \geq -q$, de sorte que $n+p \geq n-q$ pour tout n, et Φ est une fonction de $([0,T]\times\mathbb{R}^m)^{p+q+1}\times[0,1]$, à valeurs dans $\mathbb{R}^m$. En effet, si l'on part de $n+p$ et que l'on descend jusqu'à $n-q$, cela se fait en $p+q+1$ étapes. Pour rendre l'écriture unique, on suppose que $n+p$ (respectivement $n-q$) est le plus grand (respectivement plus petit) indice qui apparaît effectivement dans le schéma. On se restreint, un peu arbitrairement, à $[0,1]$ pour la variable h, qui est le pas de la discrétisation, car celle-ci va tendre vers 0 dans la suite. N'importe quel autre intervalle compact contenant 0 conviendrait aussi bien, d'ailleurs on sera parfois contraints de prendre un tel autre intervalle $[0,h_0]$.

Cette forme englobe tous les exemples précédents, sauf le schéma leapfrog qui ne rentre pas tout à fait naturellement dans cette case. Par exemple, pour le schéma d'Euler, on a $\Phi(t_n, y_n, h) = f(t_n, y_n)$ soit $p = q = 0$, ou pour le schéma d'Euler implicite $\Phi(t_{n+1}, y_{n+1}, h) = f(t_{n+1}, y_{n+1})$ soit $p = 1$, $q = -1$. Dans la suite, on se limitera à $p \leq 1$.

Si $p \leq 0$, le schéma est dit explicite, sinon il est dit implicite, l'idée étant toujours la même : avec un schéma explicite, la donnée des valeurs y_j correspondant à des instants antérieurs à t_n, donc normalement déjà calculées, donne directement y_{n+1} par une formule

16. Que donne le schéma leapfrog sur cet exemple, d'ailleurs ?

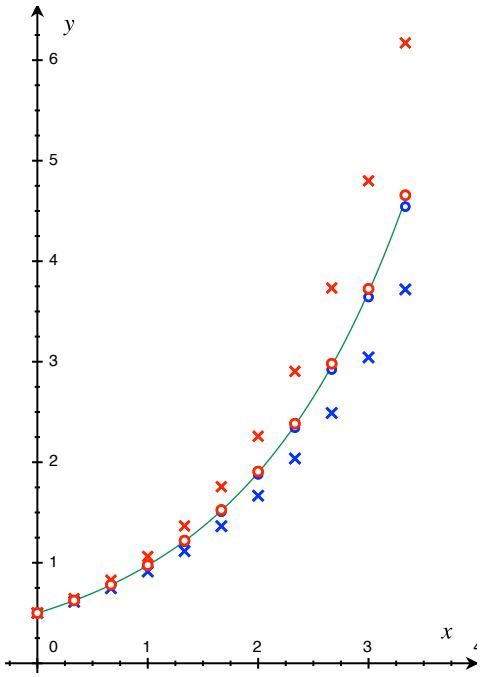


FIGURE 2.10 – × : Euler, × : Euler implicite, ○ : Euler modifiée, ○ : Crank-Nicolson, tous avec le même pas $h = \frac{1}{3}$. En vert, la solution exacte. On en ressort avec l'impression nette que Euler modifiée et Crank-Nicolson donnent de meilleurs résultats que Euler et Euler implicite. Intuition à confirmer plus tard.

connue qu'il suffit alors d'appliquer, alors que dans le cas d'un schéma implicite, on doit résoudre une équation en général non linéaire pour déterminer y_{n+1}. [17] Parmi les schémas déjà vus, les schémas d'Euler implicite et de Crank-Nicolson sont implicites, les autres sont explicites. On verra un peu plus loin que la terminologie explicite-implicite est légèrement différente dans le cas des schémas de Runge-Kutta.

Si $q \geq 1$, on dira que le schéma est à $q+1$ pas, et si $q \leq 0$ qu'il est à un pas. Par exemple, les schémas d'Euler, Euler implicite, point milieu et Crank-Nicolson sont à un pas. Le schéma leapfrog est considéré comme un schéma à deux pas.

Pour démarrer un schéma à un pas, il suffit d'une valeur y_0, dans la mesure du possible prise égale à $y(0)$, à moins que cette valeur ne soit pas exactement connue, auquel cas il faut se contenter d'une approximation $y_0 \approx y(0)$ (ce qui est le cas général dans la vraie vie). [18] La situation est différente pour les schémas à $q+1$ pas avec $q \geq 1$: ces schémas ne peuvent être utilisés que pour calculer les valeurs y_n pour $n \geq q+1$ et cela suppose connues les $q+1$ premières valeurs $y_0, \ldots y_q$. Or seule y_0 est donnée par le problème de Cauchy. Pour calculer les valeurs manquantes, il faut utiliser d'autres schémas, par exemple un schéma à un pas pour calculer y_1, puis un schéma à deux pas pour calculer y_2, etc. et un schéma à q pas pour calculer y_q, ou bien calculer les q valeurs de $n = 1$ à $n = q$ avec un schéma à un

17. On peut légitimement s'interroger sur la raison de s'imposer cette torture supplémentaire dans les schémas implicites... on verra ça plus tard.

18. Il faudrait donc pour être complètement précis dans la notation distinguer entre le y_0 donnée initiale du problème de Cauchy et le y_0, première valeur d'un schéma numérique. Mais c'est trop lourd, alors on ne le fait pas.

pas en partant de y_0...

Par ailleurs, des variantes sophistiquées des schémas construits sur les principes précédents peuvent être introduites, en particulier pour gérer la taille du pas entre deux points de discrétisation. En effet, il est souvent utile de faire varier ce pas en fonction de la régularité de la solution. Dans les zones où la solution varie beaucoup, il faut suivre au mieux ces variations et l'on utilisera un pas de temps plus petit que dans les zones où elle est varie peu et où il n'est pas utile de calculer trop souvent. On sent que les stratégies pour arriver à faire cela de façon algorithmique ne sont pas forcément évidentes.

On est donc face à un choix impressionnant de schémas numériques, qui ont tous leurs avantages et leurs inconvénients :

— simplicité/difficulté de la conception,
— simplicité/difficulté de la mise en œuvre informatique,
— rapidité d'exécution (nombre d'opérations élémentaires),
— précision (par rapport à la solution exacte qu'on ne connaît pas en général...),
— stabilité par rapport à de petites variations ou erreurs sur la condition initiale.

Nous allons dans les chapitres suivants préciser et quantifier l'évaluation de ces différents critères. Nous verrons également sur des exemples qu'en pratique le choix n'est pas toujours simple.

Chapitre 3

Analyse numérique matricielle

3.1 Motivation

On considère une EDO linéaire d'ordre 2 avec condition initiale et finale : $I =]0,1[$,

$$\forall t \in I, y''(t) = y(t), \quad \text{et } y(0) = 0, y(1) = 1.$$

Bien sûr, l'unique solution de ce problème de Cauchy est donnée par $y(t) = \frac{\exp(t)-\exp(-t)}{\exp(1)-\exp(-1)}$. Supposons qu'on veuille approcher cette solution par la méthode d'Euler (explicite). Soit $N \geq 1$, on décompose $[0,1]$ en N intervalles et on pose $t_n = \frac{n}{N}$ pour $n = 0, \dots, N$, $h = \frac{1}{N}$. La méthode d'Euler consiste à approcher la dérivée par un taux d'accroissement. Pour la dérivée seconde, on écrit

$$y''(t_n) \simeq \frac{y'(t_n) - y'(t_{n-1})}{h} \simeq \frac{\frac{y(t_{n+1})-y(t_n)}{h} - \frac{y(t_n)-y(t_{n-1})}{h}}{h} = \frac{y(t_{n+1}) - 2y(t_n) + y(t_{n-1})}{h^2}.$$

Le schéma d'Euler s'écrit alors

$$y_0 = 0, y_N = 1, \quad \forall n = 1, \dots, N-2, \ \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} = y_n,$$

ou encore

$$y_0 = 0, y_N = 1, \quad \forall n = 1, \dots, N-2, \ y_{n+1} - (2+h^2)y_n + y_{n-1} = 0. \tag{3.1.1}$$

Bien que (3.1.1) soit linéaire, la solution ne peut pas se calculer de proche en proche : comme on ne connaît pas y_1, la première relation $y_2 - (2+h^2)y_1 + y_0 = 0$ ne permet pas de calculer y_2. Par ailleurs, la condition finale $y_N = 1$ donne une relation de compatibilité entre y_{N-1} et y_{N-2} sous la forme $1 - (2+h^2)y_{N-1} + y_{N-2} = 0$. Pour y voir plus clair, récrivons (3.1.1) sous forme matricielle. En posant $Y = (y_n)_{n=1,\dots,N-1}$, on obtient

$$\underbrace{\begin{bmatrix} -(2+h^2) & 1 & 0 & \dots & 0 \\ 1 & -(2+h^2) & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & -(2+h^2) & 1 \\ 0 & \dots & 0 & 1 & -(2+h^2) \end{bmatrix}}_{=: A} \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-2} \\ y_{N-1} \end{bmatrix}}_{=: Y} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -1 \end{bmatrix}}_{=: B}, \tag{3.1.2}$$

ou de manière compacte $AY = B$, où A est une matrice carrée symétrique de taille $(N-1)\times(N-1)$ et B est un vecteur de taille $N-1$. La première question est de savoir si ce système admet bien une unique solution. La matrice A est à diagonale strictement dominante, elle est donc inversible par le lemme d'Hadamard.

D'un point de vue pratique, comment calculer la solution Y ? Ceci est l'objectif de l'analyse numérique matricielle. Nous allons présenter deux types de méthodes : les méthodes directes (qui reviennent à inverser la matrice A d'une manière ou d'une autre) et les méthodes itératives (qui reviennent à approcher la solution Y de manière itérative, sans jamais inverser explicitement A).

3.2 Méthodes directes

On appelle méthode directe de résolution d'un système linéaire une méthode qui converge en un nombre fini (et prédéfini) d'itérations. Dans cette section, nous revisitons d'abord la méthode du pivot de Gauss.

3.2.1 Notation et opérations élémentaires

On considère A une matrice carrée réelle de taille N, avec la convention que a_{ij} est l'entrée de la i-ème ligne et de la j-ème colonne de A (de manière générale, les majuscules sont réservées aux matrices ou vecteurs et les minuscules à leurs entrées). On utilise la notation $A_{i:}$ pour la i-ème ligne $A_{i:} = \{a_{ij}\}_{j=1,\dots,N}$ et $A_{:i}$ pour la i-ème colonne $A_{:i} = \{a_{ji}\}_{j=1,\dots,N}$.

On définit deux opérations élémentaires :

— l'échange des lignes i et j, qu'on note de la façon $D \leftarrow (A_{i:} \leftrightarrow A_{j:})$ (D est la matrice obtenue en échangeant les lignes i et j de A) ;

— la combinaison linéaire de lignes de A, qu'on note de la façon $D_{i:} \leftarrow (\lambda A_{i:} + \mu A_{j:})$ (D est la matrice obtenue en remplaçant la ligne i de A par λ fois la ligne i plus μ fois la ligne j, pour $\lambda, \mu \in \mathbb{R}$).

On notera aussi A^T pour la transposée de A : $a^T_{ij} := a_{ji}$ pour tout $1 \le i, j \le N$.

3.2.2 Factorisation $A = LU$ et résolution de $AY = B$

Notre objectif est de construire deux matrices triangulaires L (inférieure, L comme "lower") et U (supérieure, U comme "upper") telles que $A = LU$, en utilisant la méthode du pivot de Gauss. Nous présentons l'algorithme en supposant qu'on ne divise jamais par zéro (i.e. tous les pivots sont non nuls).

La méthode de Gauss consiste, à chaque itération $1 \le k \le N-1$, à éliminer la variable d'indice k des équations $k+1 \dots N$ de la matrice. On note $A^{(1)} = A$. On procède de manière itérative et on suppose les $k-1$ premières itérations établies au sens où on a obtenu une matrice $A^{(k)}$ de la forme

$$A^{(k)} = \begin{bmatrix} U^{(k)} & A^{(k)}_{12} \\ A^{(k)}_{21} & A^{(k)}_{22} \end{bmatrix} \in \mathbb{R}^{k+(N-k),k+(N-k)},$$

avec

$$U^{(k)} = \begin{bmatrix} a_{11}^{(k)} & \dots & a_{1k}^{(k)} \\ & \ddots & \vdots \\ & & a_{kk}^{(k)} \end{bmatrix}, \quad A_{21}^{(k)} = \begin{bmatrix} 0 & \dots & 0 & a_{k+1,k}^{(k)} \\ 0 & \dots & 0 & \vdots \\ 0 & \dots & 0 & a_{N,k}^{(k)} \end{bmatrix}.$$

A l'itération k, on cherche à interpréter les opérations élémentaires menant à l'élimination de la variable d'indice k en termes de multiplication à gauche par des matrices bien choisies. Supposons que $u_{kk} = a_{kk}^{(k)}$ (le pivot) est non nul. Pour annuler les coefficients de la variable k dans les lignes $k+1, \dots, N$ de A, on fait naturellement les combinaisons linéaires suivantes

$$A_{i:}^{(k+1)} \leftarrow A_{i:}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} A_{k:}^{(k)} \text{ pour } k+1 \le i \le N.$$

En termes matriciels, cela revient à multiplier $A^{(k)}$ à gauche par $E^{(k)} = (e_{ij}) \in \mathbb{R}^{N,N}$, avec

$$e_{ii} = 1 \text{ pour } 1 \le i \le N, \quad e_{ik}^{(k)} = -\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \text{ pour } k+1 \le i \le N, \quad e_{ij} = 0 \text{ autrement,}$$

soit

$$E^{(k)} = \begin{bmatrix} 1 & & & & & \\ 0 & \ddots & & & & \\ \vdots & & 1 & & & \\ \vdots & & -\frac{a_{k+1,k}^{(k)}}{a_{kk}^{(k)}} & 1 & & \\ \vdots & & \vdots & & \ddots & \\ 0 & & -\frac{a_{N,k}^{(k)}}{a_{kk}^{(k)}} & & & 1 \end{bmatrix}.$$

Par construction, la nouvelle matrice $A^{(k+1)} = E^{(k)}A^{(k)}$ est telle que $a_{ij}^{(k+1)} = 0$ pour tout $1 \le j \le k+1$ et $1 \le i < j$. Un point important à noter est que la matrice $E^{(k)}$ est triangulaire inférieure car, pour tout $1 \le i < j \le N$ on a $E_{ij}^{(k)} = 0$. De plus, on a diag $(E(k)) = (1, \dots, 1)$.

En supposant tous les pivots non nuls, à savoir, $a_{kk}^{(k)} \neq 0$ pour tout $1 \le k \le N-1$, on peut donc écrire la méthode de Gauss de la manière compacte

$$E^{(N-1)} \dots E^{(1)} A = U \iff A = LU,$$

où nous avons posé $L := (E^{(N-1)} \dots E^{(1)})^{-1}$ et où $U := A^{(N)} \in \mathbb{R}^{N,N}$ est une matrice triangulaire supérieure. La proposition suivante montre que la matrice L est obtenue au cours des itérations sans coût supplémentaire.

Proposition 3.2.1 (Calcul de la matrice L) *Supposons qu'aucun pivot n'est nul. La matrice L est triangulaire inférieure et on a $L = L^{(1)} \dots L^{(N-1)}$, où pour tout $1 \le k \le N-1$, $L^{(k)} = (E^{(k)})^{-1}$ est explicitement donnée par*

$$l_{ii}^{(k)} = 1 \text{ pour } 1 \le i \le N, \quad l_{ik}^{(k)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \text{ pour } k+1 \le i \le N, \quad l_{ij}^{(k)} = 0 \text{ autrement,}$$

soit

$$L^{(k)} = \begin{bmatrix} 1 & & & & \\ 0 & \ddots & & & \\ \vdots & & 1 & & \\ \vdots & & \frac{a^{(k)}_{k+1,k}}{a^{(k)}_{kk}} & 1 & \\ \vdots & & \vdots & & \ddots \\ 0 & & \frac{a^{(k)}_{N,k}}{a^{(k)}_{kk}} & & 1 \end{bmatrix} = 2\mathrm{Id}_N - E^{(k)}.$$

En particulier, pour tout $1 \leq k \leq N-1$*, le produit* $\tilde{L}^{(k)} := L^{(1)} \dots L^{(k)}$ *prend la forme*

$$\tilde{L}^{(k)} = \begin{bmatrix} 1 & & & & & \\ \frac{a^{(1)}_{21}}{a^{(1)}_{11}} & \ddots & & & & \\ \vdots & \ddots & 1 & & & \\ \vdots & & \frac{a^{(k)}_{k+1,k}}{a^{(k)}_{kk}} & 1 & & \\ \vdots & & \vdots & & \ddots & \\ \frac{a^{(1)}_{N1}}{a^{(1)}_{11}} & \dots & \frac{a^{(k)}_{N,k}}{a^{(k)}_{kk}} & & & 1 \end{bmatrix}$$

et on a $\tilde{L}^{(k)} A^{(k+1)} = A$*.*

Démonstration. Comme $A^{(k+1)}_{kk} = A^{(k)}_{kk}$, pour tout $1 \leq k \leq N-1$ la matrice $A^{(k)}$ peut s'obtenir à partir de la matrice $A^{(k+1)}$ en effectuant les combinaisons linéaires

$$A^{(k)}_{i:} \leftarrow A^{(k+1)}_{i:} + \frac{a^{(k)}_{ik}}{a^{(k)}_{kk}} A^{(k+1)}_{k:} \text{ pour } k+1 \leq i \leq N,$$

qu'on peut récrire sous la forme matricielle $A^{(k)} = L^{(k)} A^{(k+1)} = L^{(k)} E^{(k)} A^{(k)}$ ou encore $A^{(k+1)} = E^{(k)} A^{(k)} = E^{(k)} L^{(k)} A^{(k+1)}$. Ceci montre bien que $E^{(k)}$ est inversive et que $(E^{(k)})^{-1} = L^{(k)}$. La matrice L est bien triangulaire inférieure comme produit de matrices triangulaires inférieures et on obtient la formule explicite donnée par récurrence (faites la preuve !). Enfin, on a par définition de l'algorithme que $E^{(k)} \dots E^{(1)} A = A^{(k+1)}$, soit $A = (E^{(1)})^{-1} \dots (E^{(k)})^{-1} A^{(k+1)} = L^{(1)} \dots L^{(k)} A^{(k+1)} = \tilde{L}^{(k)} A^{(k+1)}$, comme annoncé. ◊

Corollaire 3.2.2 (Déterminant de A) *Soit A une matrice admettant une factorisation $A = LU$ comme ci-dessus, alors*

$$\det A = \det U = \prod_{k=1}^{N} u_{kk}.$$

Démonstration. D'une parte $\det A = \det(LU) = \det L \det U$. D'autre part, comme L et U sont triangulaires, $\det L = \prod_{k=1} l_{kk} = 1$ (cf. diagonale de 1) et $\det U = \prod_{k=1}^{N} u_{kk}$. On en déduit le résultat. ◊

Un décompte rapide montre que le nombre de d'opérations élémentaires (i.e multiplications, divisions et additions) à effectuer pour la factorisation LU est de l'ordre de $2N^3/3$ (en négligeant les termes d'ordre inférieur).

Le résultat suivant donne une condition suffisante pour que la méthode de factorisation soit bien définie (c'est-à-dire une condition qui assure a priori qu'aucun pivot n'est nul).

Théorème 3.2.3 *Soit $A = (a_{ij}) \in \mathbb{R}^{N,N}$ une matrice et posons pour tout $1 \leq k \leq N$*

$$\Delta^{(k)} := \begin{bmatrix} a_{11} & \dots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \dots & a_{kk} \end{bmatrix} \in \mathbb{R}^{k,k}.$$

Si pour tout $1 \leq k \leq N$, $\Delta^{(k)}$ est inversible, alors il existe une unique factorisation $A = LU$ avec U triangulaire supérieure inversible et L triangulaire inférieure dont les entrées diagonales sont 1.

Démonstration. On procède par récurrence pour montrer que l'hypothèse implique que le pivot ne peut pas être nul dans la méthode de Gauss. Comme $a_{11} = \Delta^{(1)} \neq 0$, on peut faire la première itération de la factorisation. Supposons qu'on ait fait $1 \leq k < N-1$ itérations et ainsi construit les matrices $\tilde{L}^{(k)}$ et $A^{(k+1)}$ qu'on décompose en quatre blocs de taille $(k+1)\times(k+1)$, $(k+1)\times(N-1-k)$, $(N-1-k)\times(k+1)$ et $(N-1-k)\times(N-1-k)$

$$\tilde{L}^{(k)} = \begin{bmatrix} \tilde{L}_{11}^{(k)} & 0 \\ \tilde{L}_{21}^{(k)} & \mathrm{Id}_{N-1-k} \end{bmatrix}, \quad A^{(k+1)} = \begin{bmatrix} U^{(k+1)} & A_{12}^{(k+1)} \\ A_{21}^{(k+1)} & A_{22}^{(k+1)} \end{bmatrix} \in \mathbb{R}^{k+1+(N-1-k),k+1+(N-1-k)}.$$

On rappelle que $U^{(k+1)}$ est triangulaire supérieure, que $\tilde{L}^{(k)}$ triangulaire inférieure avec une diagonale de 1 et qu'on a

$$\tilde{L}^{(k)}A^{(k+1)} = A = \begin{bmatrix} \Delta^{(k+1)} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

d'où on déduit par un calcul par blocs que $\tilde{L}_{11}^{(k)}U^{(k+1)} = \Delta^{(k+1)}$. Puisque $\tilde{L}_{11}^{(k)}$ est inversible (comme matrice triangulaire avec une diagonale de 1, et donc de déterminant 1), cela implique $U^{(k+1)} = (\tilde{L}_{11}^{(k)})^{-1}\Delta^{(k+1)}$. La matrice triangulaire supérieure $U^{(k+1)}$ est donc inversible comme produit de deux matrices inversibles (en utilisant l'hypothèse sur $\Delta^{(k+1)}$) et son déterminant est non nul. Comme $\det U^{(k+1)} = \prod_{i=1}^{k+1} a_{ii}^{(k+1)} \neq 0$, le pivot $a_{k+1,k+1}^{(k+1)}$ est nécessairement non nul et on peut passer à l'étape $k+1$. Par la récurrence ci-dessus, on peut poursuivre l'algorithme jusqu'à l'étape $k+1 = N-1$, définissant $\tilde{L}^{(N-1)}$ et $A^{(N)}$, auquel cas on a terminé la factorisation $A = LU$ en posant $U = A^{(N)}$ et $L = \tilde{L}^{(N-1)}$.

On conclut par l'unicité. Supposons qu'on ait $A = L_1U_1 = L_2U_2$ pour deux jeux de matrices L_1, L_2 triangulaires inférieures avec des 1 comme entrées diagonales et U_1, U_2 triangulaires supérieures. Comme A, L_1 et L_2 sont inversibles, U_1 et U_2 sont inversibles et on a l'égalité $U_1U_2^{-1} = L_1^{-1}L_2$. De cette égalité on veut conclure que $M := U_1U_2^{-1} = L_1^{-1}L_2 = \mathrm{Id}$. En notant que l'inverse d'une matrice triangulaire supérieure (respectivement inférieure avec une diagonale de 1) est triangulaire supérieure (respectivement inférieure avec une diagonale de 1) et que le produit de deux matrices triangulaires supérieures (respectivement inférieures avec une diagonale de 1) est triangulaire supérieure (respectivement inférieure

avec une diagonale de 1), on déduit que M est triangulaire supérieure car $M = U_1 U_2^{-1}$ et triangulaire inférieure avec une diagonale de 1 car $M = L_1^{-1} L_2$. Ainsi M est nécessairement diagonale et ses entrées diagonales sont 1, soit $M = \mathrm{Id}$. Ceci prouve que $L_1 = L_2$ et $U_1 = U_2$: la décomposition est bien unique. ◊

On pourrait croire que ce théorème n'a qu'une importance théorique. En effet, comment peut-on savoir a priori que toutes les "sous-matrices" de A sont inversibles ? Si on revient à l'exemple (3.1.2) qui a motivé ce chapitre, on se rend rapidement compte que toutes les sous-matrices ont exactement la même forme et sont donc, tout comme A, inversibles. Les matrices rencontrées lors de la discrétisation d'EDO (et aussi d'équations aux dérivées partielles) ont souvent une structure particulière qu'il conviendra d'essayer d'exploiter en pratique.

D'une manière plus générale, on n'implémente pas cet algorithme directement mais on se permet plutôt d'échanger des lignes (voire des colonnes) de manière à trouver le grand pivot possible (en particulier un pivot non nul). Cette méthode, appelée "pivoting" partiel ou total, sera traitée en exercice.

Une fois qu'on a une factorisation de type $A = LU$ pour la matrice $A \in \mathbb{R}^{N,N}$, la résolution du système linéaire $AY = B$ avec $Y, B \in \mathbb{R}^N$ se fait simplement en résolvant à la suite les deux systèmes triangulaires suivants (par substitutions successives)

$$LX = B, \qquad UY = X.$$

En effet, on a alors $LUY = LX = B$ comme voulu. On parle de descente pour la matrice triangulaire inférieure et de remontée pour la matrice triangulaire inférieure. Traitons en détail la remontée pour résoudre $UY = X$. On rappelle que U a la forme générale

$$U = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1N} \\ & u_{22} & \dots & u_{2N} \\ & & \ddots & \vdots \\ & & & u_{NN} \end{bmatrix}$$

et qu'on a $u_{kk} \neq 0$ car $0 \neq \det U = \prod_{k=1}^{N} u_{kk}$. La solution Y de $UY = X$ est donc donnée par

$$y_N = \frac{x_N}{u_{NN}}, \quad \text{pour } i \text{ de 1 à } N-1, \quad y_{N-i} = \frac{1}{u_{N-i,N-i}} \Big(x_{N-i} - \sum_{j=N-i+1}^{N} u_{N-i,j} y_j \Big).$$

On remarque que le nombre d'opérations (additions, multiplications et divisions) est exactement N^2, ce qui est négligeable devant le nombre d'opérations requises pour la factorisation (de l'ordre de $2N^3/2$).

3.2.3 Méthode de Cholesky

Dans ce paragraphe on traite le cas particulier (et souvent rencontré en pratique) des matrices symétriques définies positives. On rappelle qu'une matrice $A \in \mathbb{R}^{N,N}$ symétrique

est définie positive si et seulement si toutes ses valeurs propres sont strictement positives, ou, de manière équivalente si pour tout vecteur $X \in \mathbb{R}^N$ non nul on a

$$X \cdot AX := \sum_{i,j=1}^{N} a_{ij} x_i x_j > 0,$$

où $X \cdot AX = (X, AX)_{\mathbb{R}^N}$ est le produit scalaire de $\mathbb{R}^N$. Reprenons la factorisation $A = LU$. En appelant D la matrice diagonale qui contient la diagonale de U, on peut écrire $A = LDM^T$ avec $M^T := D^{-1}U$, matrice triangulaire supérieure avec une diagonale de 1 (donc M est triangulaire inférieure avec une diagonale de 1). Sous les hypothèses du Théorème 3.2.3, la décomposition $A = LDM^T$ est unique. Si A est symétrique, on obtient $LDM^T = A = A^T = MDL^T$, d'où on déduit $M = L$ par unicité de la décomposition. On a ainsi $A = LDL^T$, d'où on déduit $D = L^{-1}A(L^{-1})^T$ (en utilisant que l'inverse commute avec la transposition, comme le montre le petit calcul $L^T(L^{-1})^T = (L^{-1}L)^T = \mathrm{Id}$).

Montrons que cela implique que D est elle-même symétrique définie positive. Elle est bien symétrique car $D^T = \left(L^{-1}A(L^{-1})^T\right)^T = L^{-1}A^T(L^{-1})^T = L^{-1}A(L^{-1})^T$ (propriété qu'on savait déjà car D est diagonale !). Soit $X \in \mathbb{R}^N$ non nul. Le vecteur $Y = (L^{-1})^T X$ est aussi non nul (car $(L^{-1})^T$ est inversible), on a

$$X \cdot DX = X \cdot L^{-1}A(L^{-1})^T X = (L^{-1})^T X \cdot A(L^{-1})^T X = Y \cdot AY > 0$$

et D est donc bien définie positive. Commes les entrées de D sont ses valeurs propres, ellse sont strictement positives et on peut définir la racine carrée de D par la formule

$$\sqrt{D} := \mathrm{diag}\,(\sqrt{d_{11}}, \ldots, \sqrt{d_{NN}}).$$

En notant $H := \sqrt{D}L^T$ (qui est une matrice triangulaire supérieure par construction), on obtient la décomposition (dite de Cholesky)

$$A = H^T H.$$

Le résultat suivant donne une construction directe de la matrice H.

Théorème 3.2.4 (Factorisation de Cholesky) *Soit $A \in \mathbb{R}^{N,N}$ une matrice symétrique définie positive. Alors il existe une unique matrice H triangulaire supérieure avec diagonale strictement positive telle que $A = H^T H$. Les éléments de H peuvent être calculés à partir des formules suivantes : $h_{11} = \sqrt{a_{11}}$ et, pour tout, $2 \le i \le N$,*

$$h_{ji} = \frac{1}{h_{jj}}\Big(a_{ij} - \sum_{k=1}^{j-1} h_{ki}h_{kj}\Big), \text{ pour } 1 \le j \le i-1, \quad h_{ii} = \Big(a_{ii} - \sum_{k=1}^{i-1} h_{ki}^2\Big)^{\frac{1}{2}}.$$

Démonstration. On commence par l'unicité et suppose qu'on a $A = H_1^T H_1 = H_2^T H_2$ pour deux telles matrices H_1 et H_2. Comme H_1 et H_2 sont inversibles, en multipliant à gauche par $(H_1^{-1})^T$ et à droite par H_2^{-1} on obtient l'égalité $H_1 H_2^{-1} = (H_1^T)^{-1} H_2^T$. Cela implique d'abord que $(H_1^T)^{-1} H_2^T$ est diagonale (elle est à la fois triangulaire supérieure et inférieure). Elle est donc égale à sa transposée et on peut récrire l'égalité sous la forme

$$H_1 H_2^{-1} = (H_2 H_1^{-1})^T = H_2 H_1^{-1} = (H_1 H_2^{-1})^{-1}.$$

Comme seules Id et $-\mathrm{Id}$ sont leur propre inverse, on a $H_1 H_2^{-1} = \pm \mathrm{Id}$, qu'on récrit $H_1 = \pm H_2$. On a donc nécessairement $H_1 = H_2$ car leurs diagonales sont toutes deux à termes positifs. Cela montre l'unicité.

On passe maintenant à l'existence et aux formules explicites. On raisonne par récurrence sur la dimension N. Pour $N = 1$, le résultat est trivial. Supposons-le vrai en dimension $N-1$ et et prouvons-le en dimension N . On considère une matrice A_N symétrique définie positive, qu'on écrit sous la forme

$$A_N = \begin{bmatrix} A_{N-1} & v \\ v^T & a_{NN} \end{bmatrix},$$

où v est un vecteur de taille $N-1$ et $a_{NN} \in \mathbb{R}$ (on a même $a_{NN} > 0$, mais on ne s'en sert pas). Avant d'appliquer l'hypothèse de récurrence, vérifions que la matrice symétrique A_{N-1} est bien définie positive. Pour cela, Il suffit de montrer que pour tout $X \in \mathbb{R}^{N-1}$ non nul on a $X \cdot A_{N-1} X > 0$. Définissons alors $Y \in \mathbb{R}^N$ via $y_i = x_i$ pour $1 \leq i \leq N-1$ et $y_N = 0$. Si X est non nul, Y est non nul, et comme A_N est définie positive, on a bien

$$0 < Y \cdot A_N Y = \sum_{i,j=1}^{N} (A_N)_{ij} y_i y_j = \sum_{i,j=1}^{N-1} (A_N)_{ij} x_i x_j = X \cdot A_{N-1} X.$$

D'après l'hypothèse de récurrence, il existe H_{N-1} triangulaire supérieure telle que $A_{N-1} = H_{N-1}^T H_{N-1}$. On cherche alors $\beta > 0$ et $h \in \mathbb{R}^{N-1}$ tels qu'on ait l'égalité

$$A_N = \begin{bmatrix} A_{N-1} & v \\ v^T & a_{NN} \end{bmatrix} = \begin{bmatrix} H_{N-1}^T & 0 \\ h^T & \beta \end{bmatrix} \begin{bmatrix} H_{N-1} & h \\ 0 & \beta \end{bmatrix},$$

auquel cas $A_N = H_N^T H_N$ avec $H_N := \begin{bmatrix} H_{N-1} & h \\ 0 & \beta \end{bmatrix}$ triangulaire supérieure à diagonale strictement positive. Vérifions que c'est possible. Par bloc, c'est équivalent aux deux égalités

$$H_{N-1}^T h = v, \quad h \cdot h + \beta^2 = a_{NN}.$$

Le première relation définit bien h de manière unique (c'est une descente pour la matrice triangulaire inférieure H_{N-1}^T, qui correspond à la définition de l'énoncé pour h_{ji}). Reste à vérifier que la seconde relation définit bien un réel $\beta > 0$. Par la première relation, on a $h = (H_{N-1}^{-1})^T v$ et donc $h \cdot h = (H_{N-1}^{-1})^T v \cdot (H_{N-1}^{-1})^T v = v \cdot A_{N-1}^{-1} v$. Avec $Y' \in \mathbb{R}^N$ défini par $y'_i = (A_{N-1}^{-1} v)_i$ pour $1 \leq i \leq N-1$ et $y'_N = -1$, on obtient l'égalité

$$\begin{aligned} Y' \cdot A_N Y' &= (A_{N-1}^{-1} v) \cdot A_{N-1} (A_{N-1}^{-1} v) + a_{NN} - 2 v \cdot A_{N-1}^{-1} v \\ &= a_{NN} - h \cdot h \end{aligned}$$

Comme A_N est symétrique définie positive, $Y' \cdot A_N Y' > 0$ et on peut définir $\beta := \sqrt{Y' \cdot A_N Y'} > 0$ qui satisfait bien la relation $h \cdot h + \beta^2 = a_{NN}$ (et correspond à la définition de l'énoncé pour h_{ii}). ◊

On conclut ce paragraphe sur la forme a priori de la matrice H. Un résultat similaire est vrai pour la factorisation LU, mais il est rendu un peu plus subtil quand on s'autorise (ce qu'il faut) le pivotage. On commence par définir la notion d'enveloppe convexe d'une matrice.

Définition 3.2.5 (Enveloppe convexe d'une matrice) *Soit $A \in \mathbb{R}^{N,N}$ une matrice symétrique définie positive. Pour tout $1 \leq i \leq N$, on pose $m_i(A) := i - \min\{j < i \,|\, a_{ij} \neq 0\}$. On appelle enveloppe convexe de A la suite d'indices*

$$\mathscr{C}(A) := \{(i,j) \,|\, 0 < i - j \leq m_i(A)\}.$$

L'enveloppe convexe de A décrit la structure des termes hors-diagonaux non nuls de A, comme on le voit dans l'exemple suivant (où les croix indiquent les indices retenus dans $\mathscr{C}(A)$).

$$A = \begin{pmatrix} 10 & 0 & 0 & 0 & 1 \\ 0 & 5 & 6 & 0 & 0 \\ 0 & 6 & 10 & 9 & 2 \\ 0 & 0 & 9 & 10 & 3 \\ 1 & 0 & 2 & 3 & 1 \end{pmatrix}, \quad \mathscr{C}(A) \rightsquigarrow \begin{pmatrix} & & & & \\ \times & & & & \\ \times & \times & & & \\ \times & & \times & & \\ \times & & \times & \times & \end{pmatrix}.$$

Corollaire 3.2.6 *Soit $A \in \mathbb{R}^{N,N}$ une matrice symétrique définie positive et H l'unique matrice triangulaire supérieure à diagonale strictement positive telle que $A = H^T H$. Alors $\mathscr{C}(H + H^T) = \mathscr{C}(A)$.*

En particulier, le stockage de H ne prend pas plus de place que le stockage de A (si A est stockée via son enveloppe convexe). Ceci est une conséquence directe des formules explicites du Théorème 3.2.4.

3.3 Méthodes itératives

Les méthodes itératives pour résoudre le système linéaire $AY = B$ sont basées sur un principe itératif et ne visent pas à inverser la matrice A, mais plutôt à calculer une suite d'approximations de la solution Y (d'autant plus précise que l'itération est grande – ou en tout cas on l'espère). Les approches que nous proposons ci-dessous sont basées sur une interprétation variationnelle de l'équation $AY = B$ pour laquelle nous avons besoin que A soit symétrique définie positive, ce que nous supposerons par la suite.

3.3.1 Reformulation de $AY = B$ comme problème de minimisation

Rappelons qu'un point y minimise une fonction j sur un ensemble E si et seulement si $y \in E$ et $j(y) \leq j(x)$ pour tout $x \in E$. Le résultat principal de ce paragraphe est le suivant.

Théorème 3.3.1 *Soit $A \in \mathbb{R}^{N,N}$ une matrice symétrique définie positive et $B \in \mathbb{R}^N$ un vecteur. Soit $J : \mathbb{R}^N \to \mathbb{R}$ la forme quadratique définie par $J(X) := \frac{1}{2} X \cdot AX - B \cdot X$. Alors on a l'équivalence pour $Y \in \mathbb{R}^N$*

$$AY = B \iff J(Y) = \inf\{J(X), X \in \mathbb{R}^N\}.$$

Ce théorème affirme que Y est solution du système linéaire $AY = B$ si et seulement si il minimise la fonction J sur $\mathbb{R}^N$.

Démonstration. Soit Y tel que $AY = B$. Un tel Y existe et est unique car A est inversible. Soit $X \in \mathbb{R}^N$. Par la symétrie de A sous la forme $Y \cdot AX = X \cdot A^T Y = X \cdot AY$, on a

$$\begin{aligned} J(X) - J(Y) &= \frac{1}{2} X \cdot AX - B \cdot X - \frac{1}{2} Y \cdot AY + B \cdot Y \\ &= \frac{1}{2} X \cdot AX - AY \cdot X - \frac{1}{2} Y \cdot AY + AY \cdot Y \\ &= \frac{1}{2}\Big(X \cdot AX + Y \cdot AY - X \cdot AY - Y \cdot AX) \; = \; \frac{1}{2}(X - Y) \cdot A(X - Y). \end{aligned}$$

Comme $(X - Y) \cdot A(X - Y) \geq 0$, on a $J(X) \geq J(Y)$ pour tout $X \in \mathbb{R}^N$ et Y est un bien un minimiseur de J. Réciproquement, si $Y' \in \mathbb{R}^N$ est un minimiseur de J, alors $0 \geq J(Y') - J(Y) = \frac{1}{2}(Y' - Y) \cdot A(Y' - Y) \geq 0$, auquel cas $(Y' - Y) \cdot A(Y' - Y) = 0$ et donc $Y' = Y$, ce qui montre que Y' satisfait bien $AY' = B$.

L'argument ci-dessus marche bien car on sait déjà que le système linéaire $AY = B$ a une solution. Pour la culture, montrons comment procéder en supposant plutôt que c'est pour le problème de minimisation qu'on sait qu'il y a une solution (c'est aussi le cas ici par un argument de convexité). Soit donc $Y' \in \mathbb{R}^N$ un vecteur qui minimise J sur $\mathbb{R}^N$. En particuler, pour tous $t > 0$ et $X \in \mathbb{R}^N$, on obtient en utilisant la symétrie de A

$$\begin{aligned} 0 \leq J(Y' + tX) - J(Y') &= \frac{1}{2}(Y' + tX) \cdot A(Y' + tX) - B \cdot (Y' + tX) - \frac{1}{2} Y' \cdot AY' + BY' \\ &= tX \cdot (AY' - B) + \frac{1}{2} t^2 X \cdot AX. \end{aligned}$$

En divisant par $t > 0$, on obtient

$$X \cdot (AY' - B) + \frac{1}{2} tX \cdot AX \geq 0,$$

puis en faisant tendre t vers 0

$$X \cdot (AY' - B) \geq 0,$$

d'où on déduit $AY' - B = 0$ (prendre X puis $-X$ dans l'inégalité, puis conclure que le seul vecteur orthogonal à tout l'espace est le vecteur nul). ◊

Les méthodes itératives que nous présentons ci-dessous visent à minimiser la forme quadratique J de manière approchée plutôt qu'à résoudre le système linéaire $AY = B$ (ce qui, à convergence, revient au même par le Théorème 3.3.1).

3.3.2 Méthode du gradient à pas fixe

Le point de départ est la minimisation de la fonctionnelle J du Théorème 3.3.1. Cette fonctionnelle est de classe C^∞ (c'est une fonction quadratique). En particulier, si on a une approximation Y_n de la solution de $AY = B$ et qu'on la perturbe par αX avec $\alpha > 0$ et $X \in \mathbb{R}^N$ (pour l'instant, le α est un peu redondant) on a (en utilisant la symétrie de A pour écrire $Y_n \cdot AX = X \cdot AY_n$)

$$\begin{aligned} J(Y_n + \alpha X) &= \frac{1}{2} Y_n \cdot AY_n + \frac{1}{2}(Y_n \cdot AX + X \cdot AY_n) + \frac{1}{2}\alpha^2 X \cdot AX - B \cdot Y_n - \alpha B \cdot X \\ &= J(Y_n) + \alpha(AY_n - B) \cdot X + O(\alpha^2 \|X\|^2_{\mathbb{R}^N}). \end{aligned}$$

Si $AY_n = B$, on a trouvé notre minimiseur. Si non, on appelle $r_n := AY_n - B$ le résidu. Dans ce cas, si on veut que $J(Y_n + \alpha X) < J(Y_n)$, il faut que $r_n \cdot X < 0$. On choisit alors $X = -r_n$ et pose $Y_{n+1} = Y_n - \alpha r_n$. Reste à choisir le scalaire α, c'est-à-dire de combien on avance dans la direction opposée au résidu r_n. Pour cela, on reformule la méthode comme

$$Y_{n+1} \;=\; (\mathrm{Id} - \alpha A)Y_n + \alpha B.$$

Sous quelles conditions cette méthode itérative converge ? Appelons $e_n := Y_n - Y$ l'erreur à l'ordre n. La relation de récurrence prend la forme, en utilisant que $AY = B$,

$$e_{n+1} \;=\; (\mathrm{Id} - \alpha A)Y_n + \alpha B - Y \;=\; (\mathrm{Id} - \alpha A)e_n + Y - \alpha AY + \alpha B - Y = (\mathrm{Id} - \alpha A)e_n.$$

Ainsi, on obtient par récurrence (la rédiger !) que pour tout $n \geq 0$

$$e_n \;=\; (\mathrm{Id} - \alpha A)^{n-1} e_0. \tag{3.3.1}$$

Dans le cas $N = 1$, on connaît bien la condition nécessaire et suffisante pour que $e_n \to 0$, il s'agit juste de $|1 - \alpha a| < 1$. Dans le cas matriciel $N > 1$, il y a un résultat du même type.

Lemme 3.3.2 *Soit $D \in \mathbb{R}^{N,N}$ une matrice symétrique et soit $\rho(D) = \max\{|\lambda|, \lambda \in \text{spectre}\,(D)\}$, où le spectre de D est l'ensemble de ses valeurs propres (on rappelle qu'une matrice symétrique est diagonalisable dans $\mathbb{R}$). On a l'équivalence suivante*

$$\text{Pour tout } X \in \mathbb{R}^N,\ \lim_{n\to\infty} D^n X = 0 \iff \rho(D) < 1.$$

Démonstration. Soit $X_1, \ldots, X_N$ une famille libre de vecteurs propres de D, associés aux valeurs propres $\lambda_1, \ldots, \lambda_N$ (comptées avec leur multiplicité). Tout $X \in \mathbb{R}^N$ se décompose sous la forme $X = \sum_{i=1}^N x_i X_i$ et on a par linéarité de D et définition des vecteurs propres

$$DX = \sum_{i=1}^N x_i \lambda_i X_i$$

et ainsi, par récurrence, pour tout n

$$D^n X = \sum_{i=1}^N x_i \lambda_i^n X_i.$$

Comme la famille $X_1, \ldots, X_N$ est libre, $\lim_{n\to\infty} D^n X = 0$ si et seulement si pour tout $1 \leq i \leq N$, $\lim_{n\to\infty} x_i \lambda_i^n = 0$, soit $|\lambda_i| < 1$, ce qui est bien équivalent à $\rho(D) < 1$. ◊

Ce lemme permet de donner une condition nécessaire et suffisante pour la convergence de la méthode du gradient.

Théorème 3.3.3 (Convergence de la méthode du gradient) *Soit $A \in \mathbb{R}^{N,N}$ une matrice symétrique définie positive,* spectre (A) *l'ensemble de ses valeurs propres et $B \in \mathbb{R}^N$ un vecteur. Soit $\alpha > 0$ un réel, $Y_0 \in \mathbb{R}^N$ un point de départ et, pour tout $n \in \mathbb{N}$, soit $Y_{n+1} = (\mathrm{Id} - \alpha A)Y_n + \alpha B$ l'itérée par la méthode du gradient à pas α. La suite Y_n converge vers la solution $Y \in \mathbb{R}^N$ de $AY = B$ pour tout point de départ $Y_0 \in \mathbb{R}^N$ si et seulement si* $\max_{\lambda \in \text{spectre}\,(A)} |1 - \alpha\lambda| < 1$. *Par*

ailleurs, pour le choix optimal $\alpha = \frac{2}{\lambda_M+\lambda_m}$ *on a l'estimation pour la norme euclidienne* $\|e_n\|_2$ *de l'erreur* $e_n = Y_n - Y$

$$\|e_n\|_2 \leq \left(\frac{\lambda_M - \lambda_m}{\lambda_M + \lambda_m}\right)^n \|e_0\|_2, \tag{3.3.2}$$

où $\lambda_m = \min\{\text{spectre}(A)\}$, $\lambda_M = \max\{\text{spectre}(A)\}$, *et* $\|\cdot\|_2$ *désigne la norme euclidienne sur* $\mathbb{R}^N$ *(soit* $\|X\|_2^2 := \sum_{i=1}^N x_i^2$*).*

Démonstration. D'après (3.3.1), avec la notation $e_n = Y_n - Y$, on a $e_n = (\mathrm{Id} - \alpha A)^n e_0$ pour tout n. En appliquant le lemme 3.3.2 à la matrice $D = \mathrm{Id} - \alpha A$, on a la convergence $e_n \to 0$ (et donc $Y_n \to Y$ pour tout choix de Y_0), si et seulement $\rho(\mathrm{Id} - \alpha A) < 1$. Comme Id est diagonale dans toute base, les vecteurs propres X_i de A sont aussi vecteurs propres de $\mathrm{Id} - \alpha A$, cette fois avec les valeurs propres $1 - \alpha\lambda_i$ (où les λ_i sont les valeurs propres de A). La condition $\rho(D) < 1$ est donc équivalente à $\max_{\lambda\in\text{spectre}\,(A)} |1 - \alpha\lambda| < 1$. On peut vérifier que le choix qui rend $\max_{\lambda\in\text{spectre}\,(A)} |1 - \alpha\lambda|$ minimal est $\alpha = \frac{2}{\lambda_M+\lambda_m}$, auquel cas $\max_{\lambda\in\text{spectre}\,(A)} |1 - \alpha\lambda| = \frac{\lambda_M-\lambda_m}{\lambda_M+\lambda_m}$. En effet, pour ce choix de α, $1 - \alpha\lambda = \frac{\lambda_m+\lambda_M-2\lambda}{\lambda_m+\lambda_M}$, est une fonction décroissante en λ, positive pour $\lambda \in [\lambda_m, \frac{\lambda_m+\lambda_M}{2}]$ et négative pour $\lambda \in [\frac{\lambda_m+\lambda_M}{2}, \lambda_M]$, et sa valeur absolue pour $\lambda \in [\lambda_m, \lambda_M]$ est maximale en $\lambda = \lambda_m$ et $\lambda = \lambda_M$ et vaut bien $\frac{\lambda_M-\lambda_m}{\lambda_M+\lambda_m}$.

Pour montrer (3.3.2), il suffit de prendre la norme euclidienne de l'égalité (3.3.1), c'est-à-dire, $\|e_n\|_2 = \|(\mathrm{Id} - \alpha A)^n e_0\|_2$, qui implique pour la norme subordonnée sur les matrices

$$\|e_n\|_2 \leq \|\mathrm{Id} - \alpha A\|_2^n \|e_0\|_2,$$

où $\|\mathrm{Id} - \alpha A\|_2 := \sup_{\|X\|_2=1} \|(\mathrm{Id} - \alpha A)X\|_2$ (les normes du membre de droite sont sur les vecteurs de $\mathbb{R}^N$ – définition A.2.1 de la norme subordonnée dans l'annexe A). En écrivant le vecteur unitaire X sous la forme $X = \sum_{i=1} x_i X_i$ pour une base orthonormée de $\mathbb{R}^N$ de vecteurs propres $\{X_1, \ldots, X_N\}$ de A, on obtient par orthogonalité des X_i, et le fait que $\|X_i\|_2 = 1$ et $\|X\|_2^2 = \sum_{i=1}^N x_i^2 = 1$

$$\begin{aligned}\|(\mathrm{Id} - \alpha A)X\|_2^2 = \|\sum_{i=1}^N (1 - \alpha\lambda_i) x_i X_i\|_2^2 &= \sum_{i=1}^N (1 - \alpha\lambda_i)^2 x_i^2 \\ &\leq \max_{\lambda\in\text{spectre}\,(A)} |1 - \alpha\lambda|^2 \sum_{i=1}^N x_i^2 = \left(\frac{\lambda_M - \lambda_m}{\lambda_M + \lambda_m}\right)^2,\end{aligned}$$

d'où on déduit $\|\mathrm{Id} - \alpha A\|_2 = \frac{\lambda_M-\lambda_m}{\lambda_M+\lambda_m}$ (choisir pour X le vecteur propre associé à λ_m pour voir que ce n'est pas juste une borne, mais bien la norme de la matrice) et donc (3.3.2). ◊

L'estimation (3.3.2) est intéressante et montre que plus les valeurs propres sont éloignées les unes des autres, plus la méthode du gradient est lente à converger. On réécrit traditionnellement le facteur dans (3.3.2) sous la forme

$$\frac{\lambda_M - \lambda_m}{\lambda_M + \lambda_m} = \frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1},$$

où le conditionnement $\mathrm{cond}_2(A)$ de A est le ratio $\mathrm{cond}_2(A) := \lambda_M/\lambda_m$ (la plus grande sur la plus petite valeur propre).

Montrons sur un exemple que le conditionnement d'une matrice donne une mesure du module de continuité de l'inverse. Pour cela, considérons la matrice A d'ordre 4 et les deux vecteurs X_1 et X_2 suivants

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \quad X_1 = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}, \quad X_2 = X_1 + 0.01 \times \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 32.01 \\ 22.99 \\ 33.01 \\ 30.99 \end{pmatrix}.$$

Un calcul direct (ou en utilisant un ordinateur) donne $\lambda_m \simeq 0.0102$ and $\lambda_M \simeq 30.28867$, d'où $\mathrm{cond}_2(A) \simeq 2984.1$, ce qui est plutôt grand. Regardons maintenant les solutions Y_1 et Y_2 aux systèmes $AY_1 = X_1$ et $AY_2 = X_2$:

$$Y_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad Y_2 = \begin{pmatrix} 1.82 \\ -0.36 \\ 1.35 \\ 0.79 \end{pmatrix}.$$

En particulier, on a $\frac{\|X_1 - X_2\|_2}{\|X_1\|_2} \simeq 3.33\ 10^{-4}$, $\frac{\|Y_1 - Y_2\|_2}{\|Y_1\|_2} \simeq 1$, d'où sur cet exemple

$$\frac{\|Y_1 - Y_2\|_2}{\|Y_1\|_2} \simeq \mathrm{cond}_2(A) \frac{\|X_1 - X_2\|_2}{\|X_1\|_2}.$$

3.3.3 Méthode du gradient conjugué

La méthode du gradient conjugué a un point de départ similaire à la méthode de gradient, au détail (fondamental) près qu'on impose en plus une condition d'orthogonalité sur les directions de descentes (on ne prend pas simplement $-r_n$).

On considère toujours $A \in \mathbb{R}^{N,N}$ une matrice symétrique définie positive et on rappelle qu'elle définit un produit scalaire sur $\mathbb{R}^N$, via la forme bilinéaire $X, X' \mapsto X \cdot AX' = \sum_{ij} a_{ij} x_i x'_j$. Soit $B \in \mathbb{R}^N$ un vecteur. On veut résoudre le système linéaire $AY = B$ de manière itérative, on choisit un point de départ $Y_0 \in \mathbb{R}^N$ et on définit le résidu $r_0 = AY_0 - B$. On va choisir des directions de descente w_n (pour $n \geq 0$) obtenues via $w_0 := -r_0$, puis par récurrence via

$$w_n = -r_n - \beta_n w_{n-1}, \tag{3.3.3}$$

où les β_n seront choisis pour assurer des conditions d'orthogonalité. Etant donnée la direction de descente, on met à jour Y_n en posant

$$Y_{n+1} = Y_n + \alpha_{n+1} w_n, \tag{3.3.4}$$

où α_{n+1} sera aussi choisi pour assurer des conditions d'orthogonalité. Avec ces définitions on a l'égalité

$$r_{n+1} = AY_{n+1} - B = AY_n + \alpha_{n+1} A w_n - B = r_n + \alpha_{n+1} A w_n. \tag{3.3.5}$$

Pour fixer les valeurs de α_n et β_n, on impose que pour tout $0 \leq j \leq n-1$

$$w_n \cdot A w_j = 0, \tag{3.3.6}$$
$$r_n \cdot w_j = 0, \tag{3.3.7}$$

c'est-à-dire que la direction de descente w_n à l'ordre n est orthogonale (pour le produit scalaire induit par A) à toutes les directions précédentes (condition sur β_n), et que le reste r_n à l'ordre n est orthogonal (pour le produit scalaire euclidien) à toutes les directions de descente précédentes (condition sur α_n). Si cela est bien possible (on ne l'a pas encore montré), la méthode converge nécessairement en au plus N itérations : en effet, après N itérations, on a N directions de descente w_n orthogonales entre elles (c'est donc une base de $\mathbb{R}^N$) et le reste à l'ordre N, qui est par définition orthogonal aux vecteurs de cette base de $\mathbb{R}^N$, est donc nécessairement nul. Ceci démontre

Théorème 3.3.4 *Soit $A \in \mathbb{R}^{N,N}$ une matrice symétrique définie positive et $B \in \mathbb{R}^N$ un vecteur. Si la méthode de gradient conjugué pour la résolution du système $AY = B$ est bien définie, alors elle converge en au plus N itérations.*

Le lemme suivant montre que la méthode du gradient conjugué est bien définie.

Lemme 3.3.5 *Pour tout $Y_0 \in \mathbb{R}^N$, il existe des paramètres α_n, β_n et une famille de vecteurs w_n pour $n \geq 1$ tels que les conditions* (3.3.6) *et* (3.3.7) *sont satisfaites pour l'itération définie par* (3.3.3) *et* (3.3.4)*, tant que $r_n \neq 0$.*

Démonstration. On procède par récurrence. On rappelle que $w_0 = -r_0$. Pour $n = 1$, on a $Y_1 = Y_0 + \alpha_1 w_0$ où on choisit α_1 de manière à satisfaire (3.3.7) : en utilisant (3.3.5),

$$0 = w_0 \cdot r_1 = w_0 \cdot (r_0 + \alpha_1 A w_0) = w_0 \cdot r_0 + \alpha_1 w_0 \cdot A w_0 \iff \alpha_1 = -\frac{w_0 \cdot r_0}{w_0 \cdot A w_0},$$

qui est bien défini (et non nul) si $w_0 = -r_0 \neq 0$. Si $r_1 \neq 0$, on définit alors la nouvelle direction de descente $w_1 = -r_1 - \beta_1 w_0$ en choisissant β_1 de manière à satisfaire (3.3.6) :

$$0 = w_0 \cdot A w_1 = w_0 \cdot A(-r_1 - \beta_1 w_0) \iff \beta_1 = -\frac{w_0 \cdot A r_1}{w_0 \cdot A w_0},$$

où on note que $w_0 \cdot A w_0 \neq 0$ si $r_0 \neq 0$.

Supposons maintenant qu'on ait construit $\alpha_i \neq 0, \beta_i, w_i \neq 0, r_i \neq 0$ pour tout $i \leq n$ satisfaisant (3.3.6) et (3.3.7). On pose $Y_{n+1} = Y_n + \alpha_{n+1} w_n$ où on choisit α_{n+1} de manière à satisfaire (3.3.7) at l'ordre $n+1$ pour $j = n$: en utilisant (3.3.5),

$$0 = w_n \cdot r_{n+1} = w_n \cdot (r_n + \alpha_{n+1} A w_n) = w_n \cdot r_n + \alpha_{n+1} w_n \cdot A w_n \iff \alpha_{n+1} = -\frac{w_n \cdot r_n}{w_n \cdot A w_n},$$

quantité bien définie puisque $w_n \neq 0$. Montrons maintenant que (3.3.7) at l'ordre $n+1$ est aussi vérifié pour $j \leq n-1$: en utilisant (3.3.5), on a

$$w_j \cdot r_{n+1} = w_j \cdot r_n + \alpha_{n+1} w_j \cdot A w_n$$

qui est bien nul grâce à (3.3.6) et (3.3.7) à l'ordre n. Par ailleurs, par définition de w_n et en utilisant (3.3.7) à l'ordre n, on a (puisqu'on a supposé $r_n \neq 0$)

$$w_n \cdot r_n = -r_n \cdot r_n - \beta_n r_n \cdot w_{n-1} = -r_n \cdot r_n < 0 \tag{3.3.8}$$

et donc $\alpha_{n+1} \neq 0$.

Une fois qu'on a Y_{n+1}, on met à jour la direction de descente $w_{n+1} = -r_{n+1} - \beta_{n+1} w_n$. On choisit β_{n+1} de manière à satisfaire (3.3.6) à l'ordre $n+1$ pour $j = n$:

$$0 = w_{n+1} \cdot A w_n = w_n \cdot A(-r_{n+1} - \beta_{n+1} w_n) \iff \beta_{n+1} = -\frac{w_n \cdot A r_{n+1}}{w_n \cdot A w_n}.$$

Notons que si $r_{n+1} = 0$, i.e. $AY_{n+1} = B$, l'algorithme a convergé et $w_{n+1} = 0$. Si non, par définition de w_{n+1} et en utilisant (3.3.7) à l'ordre $n+1$, on a $w_{n+1} \cdot r_{n+1} = -r_{n+1} \cdot r_{n+1} - \beta_{n+1} r_{n+1} \cdot w_n = -r_{n+1} \cdot r_{n+1} \neq 0$, ce qui montre que $w_{n+1} \neq 0$. Supposons donc $r_{n+1} \neq 0$ et vérifions que ce choix de β_{n+1} implique bien (3.3.6) à l'ordre $n+1$ pour $j \leq n-1$. Par symétrie de A et la définition de w_{n+1}, on a

$$w_{n+1} \cdot A w_j = w_j \cdot A(-r_{n+1} - \beta_{n+1} w_n) = -w_j \cdot A r_{n+1} - \beta_{n+1} w_j \cdot A w_n = -r_{n+1} \cdot A w_j,$$

en utilisant (3.3.6) à l'ordre n qui implique que $w_j \cdot A w_n = 0$. Montrons qu'également le terme $r_{n+1} \cdot A w_j$ est nul. D'une part, la définition $w_i = -r_i - \beta_i w_{i-1}$ et la condition (3.3.7) à l'ordre $i \leq n$ impliquent (par récurrence) que $V_{i+1} := \mathrm{vect}(r_0, \ldots, r_i) = \mathrm{vect}(w_0, \ldots, w_i)$ pour $i \leq n$ et, en utilisant (3.3.5), $A w_j = \frac{1}{\alpha_{j+1}}(r_j - r_{j+1}) \in V_{n+1}$ (qui a bien un sens car $\alpha_{j+1} \neq 0$). D'autre part, par (3.3.7) à l'ordre $n+1$, r_{n+1} est orthogonal à V_{n+1}. Ainsi $A w_j$ et r_{n+1} sont orthogonaux et on déduit bien que $w_{n+1} \cdot A w_j = 0$. ◇

En résumé, on a montré que la méthode du gradient s'écrit de la manière suivante : tant que $r_n \neq 0$,

$$\alpha_{n+1} = -\frac{w_n \cdot r_n}{w_n \cdot A w_n}, \tag{3.3.9}$$

$$\begin{aligned} Y_{n+1} &= Y_n + \alpha_{n+1} w_n, \\ r_{n+1} &= r_n + \alpha_{n+1} A w_n, \end{aligned} \tag{3.3.10}$$

$$\begin{aligned} \beta_{n+1} &= -\frac{w_n \cdot A r_{n+1}}{w_n \cdot A w_n}, \\ w_{n+1} &= r_{n+1} - \beta_{n+1} w_n, \end{aligned}$$

et qu'elle converge en au plus N itérations. Bien que la méthode converge en un nombre fini d'itérations, on la considère plutôt comme une méthode itérative. En effet, elle donne souvent de bonnes approximations en peu d'itérations (dépendant notamment du conditionnement de A). En tous les cas, la convergence est monotone.

Proposition 3.3.6 *Soit $A \in \mathbb{R}^{N,N}$ une matrice symétrique définie positive, $B \in \mathbb{R}^N$ un vecteur et J la fonctionnelle quadratique associée par le théorème 3.3.1. Pour toute itération $n < N$, si $r_n \neq 0$, alors $J(Y_{n+1}) < J(Y_n)$.*

Démonstration. On commence par calculer $J(Y_{n+1}) - J(Y_n)$ et on utilise (3.3.9) et (3.3.10) pour obtenir la dernière ligne

$$\begin{aligned} & J(Y_{n+1}) - J(Y_n) \\ &= \frac{1}{2}(Y_n + \alpha_{n+1} w_n) \cdot A(Y_n + \alpha_{n+1} w_n) - B \cdot (Y_n + \alpha_{n+1} w_n) - \frac{1}{2} Y_n \cdot A Y_n + B \cdot Y_n \\ &= \alpha_{n+1} w_n \cdot (A Y_n - B) + \frac{1}{2} \alpha_{n+1}^2 w_n \cdot A w_n \\ &\overset{(3.3.9)\&(3.3.10)}{=} \alpha_{n+1} w_n \cdot r_n - \frac{1}{2} \alpha_{n+1} w_n \cdot r_n \\ &= \frac{1}{2} \alpha_{n+1} w_n \cdot r_n. \end{aligned}$$

L'inégalité $J(Y_{n+1}) < J(Y_n)$ est maintenant une conséquence de (3.3.8) puisqu'on a supposé $r_n \neq 0$. ◊

3.4 Quelle méthode choisir en pratique ?

— l'aspect mémoire
— l'aspect nombre d'itérations
— l'aspect parallélisme
— le conditionnement
— le nombre de fois que le système est résolu

Chapitre 4

Schémas à un pas : analyse générale

4.1 Schémas explicites

4.1.1 Formulation du schéma

Comme leur nom l'indique, il s'agit de schémas dans lesquels le calcul de y_{n+1} ne dépend que de y_n (et de h et de t_n, cela va sans dire) et qui s'écrivent sous la forme générique

$$y_{n+1} = y_n + hF(t_n, y_n, h), \tag{4.1.1}$$

où F est une fonction définie et continue [1] sur $[0,T] \times \mathbb{R}^m \times [0,1]$ à valeurs dans $\mathbb{R}^m$, que l'on sait écrire explicitement [2].

Par exemple, le schéma d'Euler correspond à $F(t,y,h) = f(t,y)$ et le schéma d'Euler modifié à $F(t,y,h) = f\big(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\big)$. En toute rigueur, dans ce dernier cas F n'est pas définie si $t+\frac{h}{2} > T$, mais l'on peut contourner l'obstacle en supposant disposer d'un prolongement de f au delà de T, qui conserve les bonnes propriétés de f. [3] Le schéma lui-même n'utilise jamais de valeurs de F faisant intervenir un tel prolongement, et c'est heureux car ce dernier est arbitraire. Dans la suite, on ignorera donc cette petite difficulté sans conséquence. La donnée d'un schéma explicite à un pas est donc la donnée d'une telle fonction F.

Comment identifie-t-on la fonction F sur un schéma donné sous la forme (4.1.1) ? Certainement pas en disant que $F(t_n, y_n, h) =$ une fonction de (t_n, y_n, h) ! C'est la même difficulté que la lecture de la fonction second membre f à partir de la donnée d'une EDO dont on a parlé tout au début. En effet, F est une fonction des variables libres et indépendantes les unes des autres $(t,y,h) \in [0,T] \times \mathbb{R}^m \times [0,1]$. Les variables $t_n = nh$, y_n donné par la récurrence et h ne sont ni libres [4], ni indépendantes les unes des autres. En posant $F(t_n, y_n, h) =$ une fonction de (t_n, y_n, h), on définit tout au plus une application de $\mathbb{N}\times\mathbb{N}^*$ dans $\mathbb{R}^m$ (et encore, si le schéma est explicite), pas une application de $[0,T]\times\mathbb{R}^m\times[0,1]$ dans $\mathbb{R}^m$. Comme dans le cas continu où il était important d'oublier la variable (t) dans

1. Donc continue par rapport au triplet (t,y,h), on l'avait notée Φ dans la section précédente.

2. En effet, pour vraiment mériter le qualificatif d'explicite, il faut bien sûr que F soit donnée par une formule explicite, sinon ce n'est pas du jeu.

3. Par exemple, $f(t,y) = f(T,y)$ pour tout $t > T$ et tout $y \in \mathbb{R}^m$ qui est continue et lipschitzienne par rapport à y etc., etc.

4. Même h ne l'est pas, puisqu'il est de la forme T/N avec N entier.

$y(t)$ pour la remplacer par un y générique, il faut ici oublier l'indice n (on rappelle que la notation indicielle n'est qu'une façon de noter une application de $\mathbb{N}$ à valeurs dans un ensemble) et laisser h parcourir tout l'intervalle $[0, 1]$ [5]. Encore une erreur courante facile à corriger (mais le mieux est de comprendre pourquoi il est important de le faire). C'est ce qu'on a fait plus haut pour le schéma d'Euler et pour le schéma d'Euler modifié.

Nous allons nous intéresser au problème suivant : trouver des hypothèses suffisantes sur F pour que le schéma (4.1.1) converge, c'est-à-dire, pour que y_n^N tende vers $y(t_n)$ quand N tend vers $+\infty$ (ou quand h tend vers 0, ce qui revient au même) en un sens précisé à la définition 4.1.1. [6]

Précisons donc cette notion de convergence d'un schéma numérique pour régler cette histoire d'interdépendance non écrite entre n et N.

Définition 4.1.1 *Le schéma* (4.1.1) *est dit convergent si, pour toute donnée initiale* $y(0)$ *du problème continu,*

$$\lim_{\substack{h\to 0\\ y_0^N\to y(0)}} \sup_{0\le n\le N} \|y_n^N - y(t_n)\| = 0. \tag{4.1.2}$$

Dans cette définition, $\|\cdot\|$ désigne une norme quelconque définie sur $\mathbb{R}^m$, le choix particulier n'ayant pas d'importance puisque toutes les normes sur $\mathbb{R}^m$ sont équivalentes. Il faut également se rappeler que h et N sont liés par la relation $h = T/N$, le temps final T étant usuellement considéré comme fixé. Donc dire $h \to 0$ est équivalent à dire $N \to +\infty$. Dans le cas où l'on prend $y_0^N = y(0)$, on peut naturellement se passer de la deuxième condition sous la limite.

Dans la suite, pour abréger la notation, on sous-entendra à nouveau le N en exposant dans y_n. On l'a juste remis ici pour clarifier que c'est bien par rapport à ce N que la limite est prise et que la convergence du schéma a lieu.

La notion de convergence que l'on vient d'introduire est bien entendue valable pour les schémas généraux, et pas seulement les schémas à un pas explicites.

Pour illustrer la convergence d'un schéma numérique, on considère l'exemple suivant.

Exemple 4.1.1 Le problème de Cauchy

$$y'(t) = 6\big(y(t) - \varphi(t)\big) + \varphi'(t), \tag{4.1.3}$$

avec la condition initiale $y(0) = \varphi(0)$, a visiblement pour solution exacte $y(t) = \varphi(t)$ mais celle-ci peut être difficile à calculer numériquement. La Figure 4.1 montre l'approximation obtenue par un schéma d'Euler explicite, dans le cas $\varphi(t) = t^2/(1 + t^2)$. On y a représenté la solution exacte et les solutions approchées obtenues avec différents pas de discrétisation h. On observe que plus h est petit et plus les valeurs calculées sont proches des valeurs exactes aux instants correspondants, de façon uniforme sur l'intervalle. L'approximation obtenue par le schéma d'Euler converge quand le pas h tend vers 0 (en tout cas sur le dessin, ce n'est pas encore prouvé).

La convergence d'un schéma est liée à deux notions indépendantes l'une de l'autre : la stabilité et la consistance que nous définissons maintenant.

5. Ou plus généralement $[0, h_0]$ pour un certain $h_0 > 0$.

6. Il faut en effet être précis car n n'est pas indépendant de N ! À n fixé, le point $t_n = nT/N$ bouge avec N, donc dire y_n^N tend vers $y(t_n)$ ne veut rien dire à strictement parler.

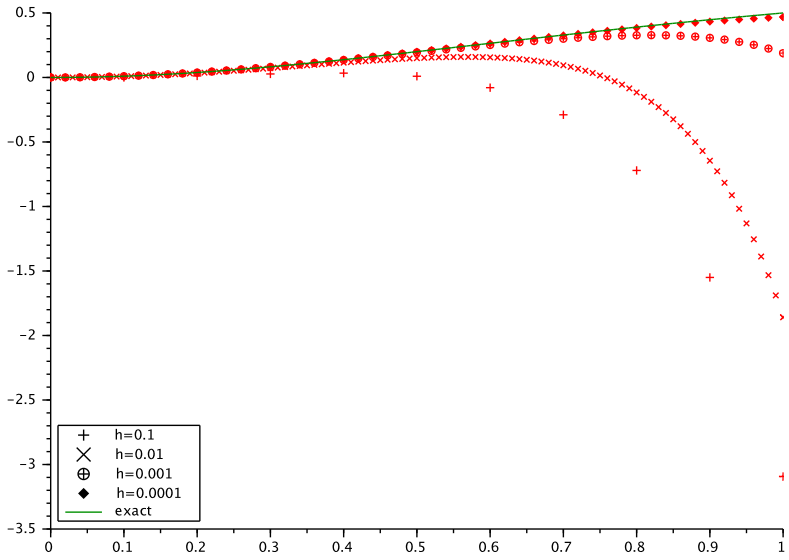


FIGURE 4.1 – Résolution de (4.1.3) avec le schéma d'Euler. On n'a pas tracé tous les points pour les deux plus petites valeurs de h (1 point sur 20 et 1 point sur 200 respectivement), il y en a trop et on ne voit plus rien.

4.1.2 Stabilité

L'idée de stabilité pour un schéma numérique est qu'une perturbation de la donnée initiale et du second membre ne doit pas être amplifiée au delà de tout contrôle par le schéma. Plus précisément,

Définition 4.1.2 *Le schéma* (4.1.1) *est* stable *s'il existe une constante C indépendante de N telle que, pour toute suite de vecteurs $(\eta_n)_{0\le n\le N}$, les suites $(y_n)_{0\le n\le N}$ et $(z_n)_{0\le n\le N}$ de $\mathbb{R}^m$ définies respectivement par*

$$y_0 \in \mathbb{R}^m \text{ et } y_{n+1} = y_n + hF(t_n, y_n, h) \text{ pour } 0 \le n \le N-1$$

et

$$z_0 = y_0 + \eta_0 \text{ et } z_{n+1} = z_n + hF(t_n, z_n, h) + \eta_{n+1} \text{ pour } 0 \le n \le N-1,$$

sont telles que

$$\max_{0\le n\le N} \|z_n - y_n\| \le C \sum_{n=0}^{N} \|\eta_n\|. \tag{4.1.4}$$

La constante C est appelée *constante de stabilité* du schéma. [7] Quand on a affaire à un schéma stable, la perturbation produite sur la solution numérique par les erreurs η_n reste donc de l'ordre du cumul de ces erreurs, modulo cette constante.

La proposition suivante donne une condition suffisante de stabilité d'un schéma numérique à un pas.

7. Il conviendrait de réserver ce nom à la plus petite constante telle que cette inégalité ait lieu, mais cette valeur optimale est en général inaccessible.

Proposition 4.1.3 *S'il existe une constante $M > 0$ telle que pour tous y et z dans $\mathbb{R}^m$, $h \in [0,1]$ et $t \in [0,T]$, on ait*

$$\|F(t,y,h) - F(t,z,h)\| \leq M\|y-z\|, \tag{4.1.5}$$

alors le schéma (4.1.1) *est stable.*

En d'autres termes, la condition suffisante est que la fonction F soit lipschitzienne par rapport à y, uniformément par rapport à t *et par rapport à* h. On note la similarité avec les hypothèses du théorème de Cauchy-Lipschitz global.

Pour démontrer ce résultat on va utiliser un lemm qui est le pendant discret du lemme de Grönwall C.3.6 vu en L2, décliné en deux versions.

Lemme 4.1.4 (Lemme de Grönwall discret, version 1) *Soit $(u_n)_{n\geq 0}$ une suite de réels. On suppose qu'il existe deux réels λ et μ, avec $\lambda \neq 0$, $1+\lambda \geq 0$ et $u_0 + \frac{\mu}{\lambda} \geq 0$, tels que*

$$\forall n \geq 0, \quad u_{n+1} - u_n \leq \lambda u_n + \mu.$$

Alors pour tout $n \geq 1$, on a

$$u_n + \frac{\mu}{\lambda} \leq \left(u_0 + \frac{\mu}{\lambda}\right) e^{\lambda n}. \tag{4.1.6}$$

Démonstration. La suite $v_n = u_n + \mu/\lambda$ vérifie l'inégalité $v_{n+1} \leq (1+\lambda)v_n$. Montrons que $v_n \leq (1+\lambda)^n v_0$ pour tout n. On procède par récurrence. C'est vrai pour $n = 0$, clairement. Supposons que $v_n \leq (1+\lambda)^n v_0$. On en déduit que $v_{n+1} \leq (1+\lambda)v_n \leq (1+\lambda)(1+\lambda)^n v_0 = (1+\lambda)^{n+1} v_0$ car $1+\lambda \geq 0$. Et comme $1+\lambda \leq e^\lambda$ et $v_0 \geq 0$, on a $v_{n+1} \leq e^{\lambda n} v_0$, d'où le résultat annoncé. ◊

Bien sûr, le résultat $u_n \leq (1+\lambda)^n\big(u_0 + \frac{\mu}{\lambda}\big) - \frac{\mu}{\lambda}$ est plus précis et n'a pas besoin de l'hypothèse $u_0 + \frac{\mu}{\lambda} \geq 0$, mais la réécriture avec l'exponentielle donne un parallèle avec la version continue du lemme.

Lemme 4.1.5 (Lemme de Grönwall discret, version 2) *Soient $(u_n)_{n\geq 0}$ et $(\mu_n)_{n\geq 0}$ deux suites de réels avec $u_0 \geq 0$ et $\mu_n \geq 0$. On suppose qu'il existe un réel λ, avec $\lambda \neq 0$ et $1+\lambda \geq 0$, tel que*

$$\forall n \geq 0, \quad u_{n+1} - u_n \leq \lambda u_n + \mu_n.$$

Alors pour tout $n \geq 1$, on a

$$u_n \leq e^{\lambda n} u_0 + \sum_{k=0}^{n-1} e^{\lambda k} \mu_{n-k-1}. \tag{4.1.7}$$

Démonstration. Montrons par récurrence sur n que

$$u_n \leq (1+\lambda)^n u_0 + \sum_{k=0}^{n-1} (1+\lambda)^k \mu_{n-1-k}. \tag{4.1.8}$$

L'inégalité (4.1.8) est vérifiée pour $n = 1$ par hypothèse. De plus

$$\begin{aligned} u_{n+1} &\leq (1+\lambda)u_n + \mu_n \\ &\leq (1+\lambda)^{n+1} u_0 + \sum_{k=0}^{n-1} (1+\lambda)^{k+1} \mu_{n-1-k} + \mu_n \end{aligned}$$

car $1+\lambda \geq 0$,

$$= (1+\lambda)^{n+1}u_0 + \sum_{k=0}^{n}(1+\lambda)^k\mu_{n-k},$$

en changeant l'indice muet de sommation. Le résultat découle alors de la majoration $1+\lambda \leq e^{\lambda}$ et des hypothèses de signe sur u_0 et μ_n. ◊

Notons que si la suite μ_n est constante, la majoration (4.1.7) devient

$$u_n \leq e^{\lambda n}u_0 + \mu\frac{e^{\lambda n}-1}{e^{\lambda}-1}, \tag{4.1.9}$$

majoration plus fine que (4.1.6) quand $\lambda \geq 0$. Par contre, si on ne majore pas par les exponentielles, les deux majorations donnent le même résultat.

Notons également que dans les usages du lemme de Grönwall, l'hypothèse de départ est souvent réécrite sous la forme équivalente $u_{n+1} \leq (1+\lambda)u_n + \mu_n$.

Démonstration de la proposition 4.1.3. Pour deux suites $(y_n)_n$ et $(z_n)_n$ telles que celles définies à la définition 4.1.2, on a pour tout $0 \leq n \leq N-1$,

$$\|z_{n+1} - y_{n+1}\| \leq \|z_n - y_n\| + h\|F(t_n, z_n, h) - F(t_n, y_n, h)\| + \|\eta_{n+1}\|,$$

par l'inégalité triangulaire. D'après l'hypothèse (4.1.5), on en déduit que

$$\|z_{n+1} - y_{n+1}\| \leq (1+hM)\|z_n - y_n\| + \|\eta_{n+1}\|.$$

En appliquant le lemme de Grönwall discret bis 4.1.5 à la suite $u_n = \|z_n - y_n\|$ avec $\lambda = hM$ et $\mu_n = \|\eta_{n+1}\|$, il vient donc

$$\|z_n - y_n\| \leq e^{hMn}\|z_0 - y_0\| + \sum_{k=0}^{n-1} e^{hMk}\|\eta_{n-k}\| = \sum_{k=0}^{n} e^{hMk}\|\eta_{n-k}\|.$$

Or dans la somme, on a $0 \leq k \leq n \leq N$, donc $hk \leq hN = T$ et donc pour tout $n \leq N$,

$$\|z_n - y_n\| \leq e^{MT}\sum_{k=0}^{n}\|\eta_k\| \leq e^{MT}\sum_{k=0}^{N}\|\eta_k\|,$$

ce qui montre que le schéma est stable, (4.1.4), en passant au max sur n au membre de gauche puisque le membre de droite de dépend pas de n, avec une constante de stabilité égale à e^{MT}. ◊

Exemple 4.1.2 Pour le schéma d'Euler, on a $F(t,y,h) - F(t,z,h) = f(t,y) - f(t,z)$. Par conséquent, si la fonction f est globalement lipschitzienne par rapport à y, uniformément par rapport à t, c'est-à-dire satisfait une partie des hypothèses du théorème de Cauchy-Lipschitz global, alors le schéma est stable. ◊

Exemple 4.1.3 Pour le schéma d'Euler modifié, et toujours pour une fonction f globalement

lipschitzienne par rapport à y, uniformément par rapport à t, [8] on a

$$\begin{aligned}\|F(t,y,h)-F(t,z,h)\| &= \left\| f\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right) - f\left(t+\frac{h}{2}, z+\frac{h}{2}f(t,z)\right)\right\| \\ &\leq L\left\| y+\frac{h}{2}f(t,y) - \left(z+\frac{h}{2}f(t,z)\right)\right\|\end{aligned}$$

en utilisant une première fois le caractère lipschitzien de f,

$$\begin{aligned}&= L\left\| y-z+\frac{h}{2}\big(f(t,y)-f(t,z)\big)\right\| \\ &\leq L\left(1+\frac{h}{2}L\right)\|y-z\| \\ &\leq \frac{L(2+L)}{2}\|y-z\|,\end{aligned}$$

en utilisant l'inégalité triangulaire puis une deuxième fois le caractère lipschitzien de f. À la fin, on majore simplement h par 1. Le schéma d'Euler modifié est donc stable. ◊

Notons que dans la stabilité, il n'est fait aucune mention du problème de Cauchy que l'on souhaite approcher. C'est une notion qui en est totalement indépendante. Passons maintenant à la deuxième notion importante, qui elle va prendre en compte le problème de Cauchy.

4.1.3 Consistance

La suite y_n est construite de façon à vérifier l'égalité

$$y_{n+1} - y_n - hF(t_n, y_n, h) = 0.$$

La solution exacte n'a pas de raison de vérifier la même égalité aux instants t_n, mais on espère qu'elle le fait à peu de chose près, l'idée étant que le schéma numérique tente alors bien d'approcher la bonne équation. Pour cela, on souhaite que la quantité $y(t_{n+1})-y(t_n)-hF(t_n, y(t_n), h)$ soit petite en norme. Cette quantité joue un rôle très important dans l'étude des schémas numériques.

Définition 4.1.6 *On appelle* erreur de consistance *(ou erreur de discrétisation locale) du schéma* (4.1.1) *la quantité* $\varepsilon_n \in \mathbb{R}^m$ *définie par*

$$\varepsilon_n = y(t_{n+1}) - y(t_n) - hF(t_n, y(t_n), h),$$

où y est une solution de l'EDO. [9]

L'expression " erreur de discrétisation locale " vient du fait qu'il s'agit de l'erreur commise par le schéma à l'instant t_{n+1} si l'on est parti à l'instant t_n avec la valeur exacte $y(t_n)$, voir figure 4.2. Naturellement, c'est une quantité inconnue, mais on verra que l'on peut l'estimer et elle jouera un rôle intermédiaire important dans l'analyse de convergence d'un schéma.

8. Prolongée par exemple à $[0, T+\frac{1}{2}]\times\mathbb{R}^m$ pour pouvoir bien écrire F.

9. L'erreur de consistance dépend donc du choix de cette solution et de h, même si cela n'apparaît pas dans la notation.

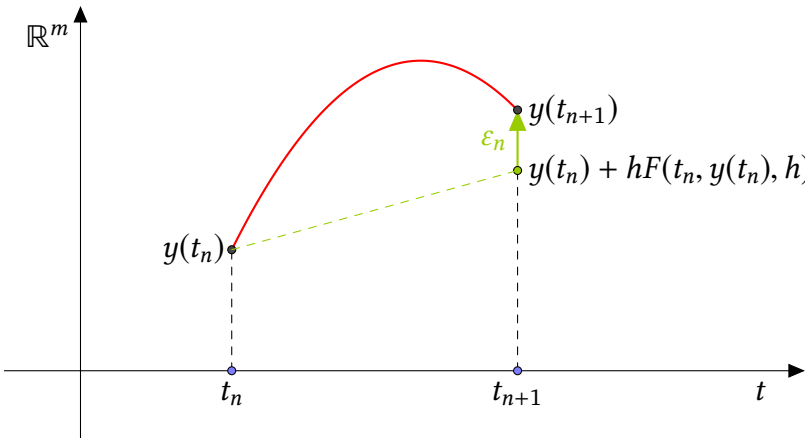


FIGURE 4.2 – Erreur de discrétisation locale, a.k.a. de consistance.

Définition 4.1.7 *Le schéma* (4.1.1) *est dit* consistant *avec l'EDO* (2.1.1) *si pour toute solution y de* (2.1.1)*, on a*

$$\lim_{h\to 0} \sum_{n=0}^{N-1} \|\varepsilon_n\| = 0.$$

L'usage dans ce contexte du mot " consistant " est bien malheureux en français, ce n'est qu'un calque maladroit de l'anglais *consistent*. Il serait plus correct de parler de cohérence avec l'EDO, mais l'usage de consistance semble maintenant tellement enraciné que l'on ne voit plus comment l'éradiquer...

Un schéma est donc ~~cohér~~ consistant si la somme des erreurs de consistance sur tous les instants de discrétisation tend vers 0 avec h, ce pour toute solution y de l'EDO.

La proposition suivante donne une condition nécessaire et suffisante de consistance d'un schéma à un pas.

Proposition 4.1.8 *On suppose l'application F continue sur $[0,T]\times\mathbb{R}^m\times[0,1]$. Le schéma* (4.1.1) *est consistant si et seulement si, pour tout $(t,y)\in I\times\mathbb{R}^m$, on a*

$$F(t,y,0) = f(t,y).$$

Démonstration. On ne traite que la condition suffisante, qui est la partie importante du résultat. On a

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} y'(u)\,du = h\int_0^1 y'(t_n+sh)\,ds = h\int_0^1 f\big(t_n+sh, y(t_n+sh)\big)\,ds,$$

en posant $s = \frac{u-t_n}{h}$. L'erreur de consistance du schéma s'écrit donc

$$\begin{aligned}\varepsilon_n &= h\int_0^1 f\big(t_n+sh, y(t_n+sh)\big)\,ds - hF(t_n, y(t_n), h)\\ &= h\int_0^1 \big(F\big(t_n+sh, y(t_n+sh), 0\big) - F(t_n, y(t_n), h)\big)\,ds,\end{aligned}$$

où l'on a utilisé l'hypothèse $f(\cdot,\cdot) = F(\cdot,\cdot,0)$ et passé la constante dans l'intégrale après avoir mis h en facteur. Il vient alors

$$\|\varepsilon_n\| \leq h \int_0^1 \big\|F\big(t_n + sh, y(t_n + sh), 0\big) - F(t_n, y(t_n), h)\big\|\, ds.$$

Soit $K = \{(t,s,h) \in [0,T]\times[0,1]\times[0,1];\, t+h \leq T\}$. C'est un fermé d'un compact, donc un compact lui-même. Introduisons la fonction $G\colon K \to \mathbb{R}_+$ par

$$G(t,s,h) = \big\|F\big(t+sh, y(t+sh), 0\big) - F(t, y(t), h)\big\|.$$

Cette fonction est continue comme composée de fonctions continues. Elle est donc uniformément continue sur le compact K. Comme par ailleurs $G(t,s,0) = 0$ pour tous t et s, on déduit de cette continuité uniforme que pour tout $\eta > 0$, il existe h_0 tel que pour $h \leq h_0$, $G(t,s,h) = |G(t,s,h) - G(t,s,0)| \leq \eta$ pour tous t et s. [10] Il s'ensuit que pour $h \leq h_0$, on a

$$\|\varepsilon_n\| \leq h\eta \int_0^1 ds = h\eta,$$

d'où

$$\sum_{n=0}^{N-1} \|\varepsilon_n\| \leq hN\eta = T\eta,$$

car il y a N termes dans la somme, ce qui montre la consistance. ◊

Le schéma d'Euler et le schéma d'Euler modifié sont donc consistants. En effet, c'est trivial pour le schéma d'Euler vu que $F(t,y,h) = f(t,y)$ pour tout h donc en particulier pour $h = 0$. Pour celui d'Euler modifié, on a $t + \frac{0}{2} \leq T$, donc $F(t,y,0) = f\big(t + \frac{0}{2}, y + \frac{0}{2} f(t,y)\big) = f(t,y)$.

4.1.4 Convergence

La raison pour laquelle on a fort mystérieusement introduit ces deux notions indépendantes de stabilité et de consistance d'un schéma, est le résultat suivant, fondamental pour la convergence des schémas numériques, parfois connu sous le nom de théorème de Lax [11], quoique probablement pas dû à Lax dans le contexte des EDO.

Théorème 4.1.9 *Si le schéma* (4.1.1) *est stable et consistant alors il est convergent.*

Démonstration. Définissons la suite $(z_n)_n$ par $z_n = y(t_n)$. On a

$$z_{n+1} = z_n + hF(t_n, z_n, h) + \varepsilon_n,$$

par définition de l'erreur de consistance du schéma ε_n. On a donc affaire à une suite perturbée au sens de la définition 4.1.2 de la stabilité, en posant $\eta_{n+1} = \varepsilon_n$ et $\eta_0 = y(0) - y_0$. Comme le schéma est stable, il existe une constante C telle que

$$\max_{0\leq n\leq N} \|z_n - y_n\| \leq C \sum_{n=0}^{N} \|\eta_n\| = C\|\eta_0\| + C \sum_{n=0}^{N-1} \|\varepsilon_n\|.$$

10. Il suffit d'écrire ce qu'est la continuité uniforme de la fonction G. En effet, celle-ci nous dit que pour tout $\eta > 0$, il existe $h_0 > 0$ tel que si $|t_1 - t_2| + |s_1 - s_2| + |h_1 - h_2| \leq h_0$ alors $|G(t_1,s_1,h_1) - G(t_2,s_2,h_2| \leq \eta$. On prend ici $t_1 = t_2 = t$, $s_1 = s_2 = s$, $h_1 = h$ et $h_2 = 0$.

11. Peter David Lax, 1926–

Comme le schéma est consistant, $\sum_{n=0}^{N-1} \|\varepsilon_n\|$ tend vers 0 avec h et le schéma est donc convergent, *cf.* définition 4.1.1. ♢

Nous avons vu que le schéma d'Euler et le schéma d'Euler modifié sont stables (si f est globalement lipschitzienne par rapport à y, uniformément par rapport à t) et consistants (si f est continue par rapport à (t, y)), ils sont donc convergents.

Etant donné un schéma stable, donné a priori par une fonction continue F, si l'on pose $g(t, y) = F(t, y, 0)$, on voit que ce schéma est automatiquement consistant avec l'EDO $y' = g(t, y)$ et converge donc vers les solutions du problème de Cauchy correspondant. On peut donc dire grossièrement qu'un schéma stable converge, mais qu'il ne converge vers ce que l'on veut que s'il est consistant. Sinon, on est train d'approcher une autre EDO. [12]

4.1.5 Ordre d'un schéma, estimation d'erreur

Savoir qu'un schéma numérique est convergent, c'est bien, mais on aimerait savoir aussi à quelle vitesse cette convergence a lieu. En d'autre termes, on souhaite quantifier la qualité de l'approximation. On introduit d'abord une définition qui raffine celle de la consistance.

Définition 4.1.10 *Le schéma* (4.1.1) *est dit d'ordre au moins $p \in \mathbb{N}^*$ si, pour toute solution y de l'EDO* (2.1.1)*, il existe une constante C indépendante de h telle que*

$$\forall N, \quad \sum_{n=0}^{N-1} \|\varepsilon_n\| \leq Ch^p. \tag{4.1.10}$$

Il est d'ordre p s'il n'est en outre pas d'ordre au moins $p + 1$.

La constante C est indépendante de h, par contre, elle va fortement dépendre de la solution y considérée, comme on le verra sur des exemples.

Le résultat qui suit est un simple raffinement du théorème de convergence dans le cas d'un schéma d'ordre p.

Théorème 4.1.11 *On suppose le schéma* (4.1.1) *stable et d'ordre $p \geq 1$ et qu'il existe une constante $C > 0$ telle que $\|y_0 - y(0)\| \leq Ch^p$. Alors il existe $\tilde{C} > 0$ telle qu'on ait l'estimation d'erreur suivante*

$$\max_{0 \leq n \leq N} \|y(t_n) - y_n\| \leq \tilde{C}h^p.$$

Démonstration. Reprendre la démonstration du théorème 4.1.9 en remplaçant l'erreur initiale et les erreurs de consistance par leurs estimations (on rappelle que y_0 et y_n portent un exposant N invisible avec $h = T/N$). La constante $\tilde{C}$ est majorée par $C_1(C_2 + C)$ où C_1 est la constante de stabilité (Definition 2.1.2) et C_2 la constante de la définition (2.1.10)

♢

Remarque 4.1.1 L'intérêt d'une méthode d'ordre p par rapport à une méthode d'ordre $p' < p$ est que sa précision est (asymptotiquement) infiniment meilleure, à pas égal ou à même nombre de points de discrétisation, puisque $h^{p-p'} \to 0$ quand $h \to 0$. Néanmoins, il faut faire attention au fait que l'estimation d'erreur précédente est une estimation dite *a priori*

12. C'est d'ailleurs pourquoi le terme de cohérence serait mieux adapté, mais enfin... tant pis.

qui contient une constante inconnue C. Cette constante fait typiquement intervenir des normes sup de dérivées de la solution y. Elle peut être petite ou grande, on n'en a aucune idée en général, et comme les calculs effectifs se font à $h > 0$, et jamais quand $h \to 0$ (car c'est impossible), la précision supérieure à pas égal n'est pas garantie a priori. On constate en pratique qu'elle l'est le plus souvent.

Un autre facteur à prendre en compte, est que plus l'ordre d'une méthode est élevé, plus le coût de cette méthode (occupation de la mémoire de l'ordinateur, nombre d'opérations pour exécuter l'algorithme donc durée du calcul) est élevé. Il faut donc faire le bilan de cette complexité accrue par rapport au gain de précision, qui permet de prendre moins de points de discrétisation qu'une méthode d'ordre moins élevé, à précision donnée.

Par exemple, pour obtenir une précision de $10^{-s}, s > 0$ avec un schéma d'ordre p, il faut prendre un pas de temps h_p tel que (on suppose $C = 1$) $h_p \leq 10^{-s/p}$. Pour obtenir la même précision avec un schéma d'ordre $p + 1$ (toujours avec $C = 1$, même si ce n'est pas le même C), il faut prendre un pas de temps $h_{p+1} \leq 10^{-s/(p+1)} \simeq h_p 10^{s/p(p+1)}$. Pour $s = 6$ et $p = 2$, le nouveau pas de temps est dix fois plus grand que le précédent, il y a donc dix fois moins de valeurs $y_1, \ldots, y_N$ à calculer, mais le calcul de chaque y_n est plus coûteux que dans la méthode d'ordre p. S'il est moins que dix fois plus coûteux (et que les constantes sont vraiment 1...), on y gagne.

Enfin il est clair que si l'on utilise un schéma d'ordre p, il faut utiliser une approximation de la donnée initiale qui soit du même ordre, sous peine de perdre toute la précision attendue et donc de calculer beaucoup pour rien (gaspillant donc de précieuses ressources). ◊

On a la condition suffisante suivante qui est celle utilisée dans la pratique.

Proposition 4.1.12 *Si pour toute solution y, il existe une constante C' indépendante de h telle que que l'erreur de consistance du schéma vérifie*

$$\forall n \leq N, \quad \|\varepsilon_n\| \leq C' h^{p+1},$$

alors le schéma est d'ordre au moins p. Si l'on a de plus $\|\varepsilon_n\| \geq C'' h^{p+1}$ pour tout $n \leq N$, avec $C'' > 0$ indépendante de h pour au moins une solution y de l'EDO, alors le schéma est d'ordre p.

Démonstration. En effet, dans ce cas

$$\sum_{n=0}^{N-1} \|\varepsilon_n\| \leq C' \sum_{n=0}^{N-1} h^{p+1} = C' N h^{p+1} = C' T h^p,$$

puisque $Nh = T$.

Supposons maintenant que $\|\varepsilon_n\| \geq C'' h^{p+1}$ pour une solution y de l'EDO. Par le même calcul que précédemment, il vient

$$\sum_{n=0}^{N-1} \|\varepsilon_n\| \geq C'' \sum_{n=0}^{N-1} h^{p+1} = C''' h^p.$$

avec $C''' = C'' T > 0$. Or il n'existe aucune constante C'''' indépendante de h telle que $C''' h^p \leq C'''' h^{p+1}$, ce qui se voit immédiatement en faisant tendre h vers 0. ◊

Par exemple pour le schéma d'Euler, l'erreur de consistance est

$$\varepsilon_n = y(t_{n+1}) - y(t_n) - hf\big(t_n, y(t_n)\big) = y(t_{n+1}) - y(t_n) - hy'(t_n).$$

Supposons que la solution y considérée est de classe C^2. [13] On a donc, par l'inégalité de Taylor-Lagrange [14]

$$\|\varepsilon_n\| \leq \frac{\max_{[0,T]} \|y''(t)\|}{2} h^2.$$

Ce schéma est donc d'ordre au moins un avec $C = \frac{T \max_{[0,T]} \|y''(t)\|}{2}$ dès que les solutions de l'EDO sont de classe C^2. Par ailleurs, il est bien évident qu'il n'est pas d'ordre deux. Comme annoncé plus tôt, on remarque que la constante C dépend bien sûr de la solution considérée y par l'intermédiaire de sa dérivée seconde.

Montrons également que le schéma d'Euler modifié est d'ordre deux. Pour ce schéma $F(t, y, h) = f\big(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\big)$ et l'erreur de consistance prend la forme

$$\begin{aligned}\varepsilon_n &= y(t_{n+1}) - y(t_n) - hf\left(t_n + \frac{h}{2}, y(t_n) + \frac{h}{2} f(t_n, y(t_n))\right) \\ &= y(t_{n+1}) - y(t_n) - hf\left(t_n + \frac{h}{2}, y(t_n) + \frac{h}{2} y'(t_n)\right).\end{aligned}$$

Notons que dans ce type de calcul, il est de notre intérêt de simplifier au maximum les expressions en utilisant l'EDO chaque fois que c'est possible. Ici par exemple, on a remplacé $f(t_n, y(t_n))$ par $y'(t_n)$ à l'intérieur du premier terme en f.

Pour simplifier, on se place dans le cas $m = 1$. Nous allons utiliser des développements de Taylor avec des restes exprimés en O pour raccourcir l'écriture. Cachés dans ces O sont des constantes qui ont toute l'uniformité voulue, car il s'agit en fait de restes de Taylor-Lagrange en dimension $m = 1$ ou bien plus généralement de restes intégraux en dimension quelconque. Il faudrait en toute rigueur établir ici cette uniformité, mais on va avoir confiance qu'elle a bien lieu, histoire de ne pas trop y passer de temps.

On écrit donc d'une part le développement de Taylor par rapport à t de la fonction y à l'ordre 2 en $t = t_n$, supposant y de classe C^3, ce qui donne

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

d'où

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

On écrit d'autre part le développement de Taylor à l'ordre 1 du terme en f, [15] cette fois-ci par rapport à h en $h = 0$, en supposant f de classe C^2, ce qui donne

$$f\left(t_n + \frac{h}{2}, y(t_n) + \frac{h}{2} y'(t_n)\right) = f(t_n, y(t_n)) + \frac{h}{2}\left(\frac{\partial f}{\partial t}(t_n, y(t_n)) + \frac{\partial f}{\partial y}(t_n, y(t_n)) y'(t_n)\right) + O(h^2).$$

13. On reviendra plus loin sur ces questions de régularité de la solution, voir proposition 4.1.13.
14. Joseph Louis, comte de Lagrange (en italien Giuseppe Lodovico de Lagrangia), 1736–1813.
15. En effet, inutile d'aller plus loin, puisque ce terme va être multiplié par h dans l'erreur de consistance.

Or on a (voir aussi la proposition 4.1.13)

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

par dérivation des fonctions composées, d'où, en utilisant également l'EDO pour le premier terme,

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

Tous les termes des divers développements de Taylor jusqu'à l'ordre 2 se simplifient visiblement dans l'erreur de consistance et l'on obtient

$$\varepsilon_n = O(h^3),$$

avec un O uniforme par rapport à n, [16] c'est-à-dire que le schéma d'Euler modifié est d'ordre (au moins) 2 quand les solutions sont C^3. Le résultat reste bien sûr valable dans le cas vectoriel $m > 1$. Pour s'assurer que le schéma est bien d'ordre 2 et pas d'un ordre encore plus grand (ce qui serait quand même un peu trop miraculeux), il faut pousser les développements de Taylor un cran plus loin et vérifier que le terme en h^3 dans l'erreur de consistance ne s'annule pas en général. ◊

On reprendra ce type de calculs en toute généralité dans la proposition 4.1.14. À retenir qu'il s'agit toujours d'utiliser des développements de Taylor. Le lecteur ou la lectrice montrera à titre d'exercice que le schéma d'Euler implicite est d'ordre un et que le schéma leapfrog est d'ordre deux. [17]

Remarque 4.1.2 Notons que dans les exemples précédents, l'ordre est obtenu sous une hypothèse de régularité de la solution. En d'autres termes, si l'on utilise la méthode d'Euler modifiée sur une EDO dont les solutions ne sont pas de classe C^3, et il y en a, il ne faut pas espérer voir un gain d'estimation d'erreur par rapport à la méthode d'Euler tout court. Sauf coup de chance. ◊

Remarque 4.1.3 Dans un cas où l'on connaît la solution exacte, on trace l'erreur en fonction du pas de discrétisation en échelle logarithmique. La pente de la droite obtenue nous donne l'ordre p de la méthode. La Figure 4.3 montre les courbes d'erreur calculée et théorique pour les schémas d'Euler implicite, Euler modifié et leapfrog. Pour évaluer graphiquement l'ordre des schémas on a également tracé les allures théoriques $O(h)$ et $O(h^2)$.

Pour parler d'ordre d'un schéma numérique, il faut pouvoir assurer la régularité des solutions de l'EDO. C'est l'objet de la proposition suivante.

Proposition 4.1.13 *On suppose que f est de classe C^p. La solution du problème de Cauchy est alors de classe C^{p+1} avec pour tout $0 \leq k \leq p$,*

$$\forall t \in [0, T], \quad y^{(k+1)}(t) = f^k(t, y(t)),$$

16. Mais encore une fois, il faudrait démontrer cette uniformité, même si c'est à peu près évident.
17. Faire preuve d'un peu d'imagination pour adapter les définitions à ces deux cas.

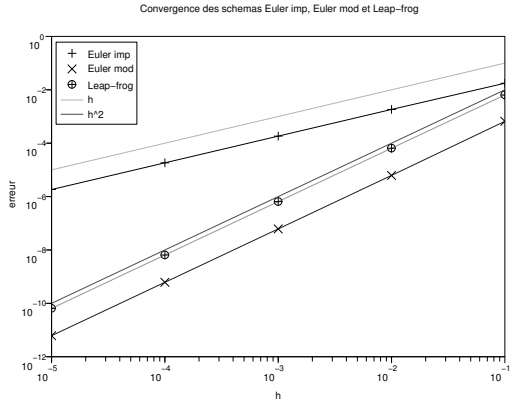


FIGURE 4.3 – Convergence des schémas d'Euler implicite, Euler modifié et leapfrog sur l'EDO $y'(t) = -y(t)$. Erreur estimée numériquement et allures théorique en $O(h)$ et $O(h^2)$.

où la suite $f^k : [0,T] \times \mathbb{R}^m \to \mathbb{R}^m$ *est définie par récurrence par*

$$\forall (t,y) \in [0,T] \times \mathbb{R}^m, \quad \begin{cases} f^0(t,y) = f(t,y), \\ f^{k+1}(t,y) = \dfrac{\partial f^k}{\partial t}(t,y) + \displaystyle\sum_{j=1}^{m} \frac{\partial f^k}{\partial y_j}(t,y) f_j(t,y), \ \textit{pour } 0 \le k \le p-1. \end{cases} \tag{4.1.11}$$

Démonstration. Montrons que toute solution du problème de Cauchy est de classe C^{p+1} avec $y^{(k+1)}(t) = f^k(t, y(t))$, pour $k = 0, \dots, p$, avec f^k de classe C^{p-k}, par récurrence sur k.

Cette relation est trivialement vraie pour $k = 0$ (c'est l'EDO elle-même) et implique que y est de classe C^1 puisque f est continue et y également, par composition de fonctions continues, *cf.* proposition C.3.3. De plus, on a $f^0 = f$ de classe C^{p-0}.

Supposons donc maintenant y de classe C^{k+1} pour un certain $k \le p-1$ avec $y^{(k+1)}(t) = f^k(t, y(t))$ et f^k de classe C^{p-k}. Comme $p - k \ge 1$, la fonction f^k est donc de classe C^1. On a déjà noté que y est de classe C^1. Il s'ensuit que $y^{(k+1)}$ est de classe C^1 par composition des fonctions de classe C^1, ce qui signifie que y est de classe C^{k+2}. Calculant alors sa dérivée, on a pour chaque composante y_i, $i = 1, \dots, m$,

$$\begin{aligned} y_i^{(k+2)}(t) = \frac{dy_i^{(k+1)}}{dt}(t) &= \frac{d}{dt}\big(f_i^k(t, y(t))\big) \\ &= \frac{\partial f_i^k}{\partial t}(t, y(t)) + \sum_{j=1}^{m} \frac{\partial f_i^k}{\partial y_j}(t, y(t)) \frac{dy_j}{dt}(t) \\ &= \frac{\partial f_i^k}{\partial t}(t, y(t)) + \sum_{j=1}^{m} \frac{\partial f_i^k}{\partial y_j}(t, y(t)) f_j(t, y(t)), \end{aligned}$$

par dérivation des fonctions composées à plusieurs variables et le fait que y est solution de l'EDO. Ceci établit la récurrence donnant les fonctions f^k. De plus, on a clairement que f^{k+1}, qui est définie par la formule $f^{k+1}(t,y) = \frac{\partial f^k}{\partial t}(t,y) + \sum_{j=1}^m \frac{\partial f^k}{\partial y_j}(t,y) f_j(t,y)$ pour tout $(t,y) \in [0,T] \times \mathbb{R}^m$, est de classe C^{p-k-1}, puisqu'on a pris des dérivées partielles premières de f^k qui est de classe C^{p-k}, multiplié certaines d'entre elles par des fonctions de classe C^p et additionné le tout.

La récurrence s'arrête pour $k = p - 1$ et donne bien y de classe C^{p-1+2} avec les formules attendues. ◊

Il faut noter que si f n'est pas de classe C^p, alors y n'a en général aucune raison d'être de classe C^{p+1}. Par exemple, si f est simplement continue par rapport à (t, y) et lipschitzienne par rapport à y, sans être de classe C^1, on n'aura pas y de classe C^2, mais seulement de classe C^1. D'où la remarque précédente sur le manque a priori d'intérêt d'utiliser dans ce cas une méthode d'ordre élevé, qui va exiger plus de régularité de la part des solutions pour que cet ordre élevé se manifeste dans l'estimation d'erreur.

Finalement, on voit que la méthode d'Euler est d'ordre 1 dès que f est de classe C^1 et que la méthode d'Euler modifiée est d'ordre 2 dès que f est de classe C^2. Par ailleurs, chacune de ces deux méthodes reste convergente, mais sans le bénéfice de l'ordre, quand f est seulement continue et globalement lipschitzienne par rapport à y uniformément par rapport à t.

On donne maintenant une condition nécessaire et suffisante pour qu'un schéma à un pas soit d'ordre $p \geq 1$. On rappelle que la consistance du schéma est assurée par $F(t, y, 0) = f(t, y)$ pour tous t,y.

Proposition 4.1.14 *On suppose que F est de classe C^p. Le schéma* (4.1.1) *est d'ordre au moins p si et seulement si pour tout $k = 0, \ldots, p-1$, on a*

$$\frac{\partial^k F}{\partial h^k}(t, y, 0) = \frac{1}{k+1} f^k(t, y), \tag{4.1.12}$$

pour tous t, y.

Démonstration. On ne traite que la condition suffisante, qui est la partie importante du résultat. Notons pour commencer que la condition (4.1.12) pour $k = 0$ donne $F(t, y, 0) = f(t, y)$, c'est-à-dire d'abord la consistance du schéma, et de plus que la fonction f est aussi de classe C^p. On peut donc appliquer la proposition 4.1.13. La condition (4.1.12) implique donc que

$$\frac{\partial^k F}{\partial h^k}(t, y(t), 0) = \frac{1}{k+1} y^{(k+1)}(t), \tag{4.1.13}$$

pour toute solution y du problème de Cauchy et tout t dans $[0, T]$.

Reprenons maintenant l'erreur de consistance

$$\varepsilon_n = y(t_{n+1}) - y(t_n) - hF(t_n, y(t_n), h).$$

On écrit le développement de Taylor avec reste intégral à l'ordre p de la fonction y en t_n, ce qui donne

$$y(t_{n+1}) = y(t_n) + \sum_{k=1}^{p} \frac{h^k}{k!} y^{(k)}(t_n) + \frac{h^{p+1}}{p!} \int_0^1 (1-s)^p y^{(p+1)}(t_n + sh)\, ds.$$

On écrit aussi le développement de Taylor avec reste intégral à l'ordre $p-1$ de la fonction F, mais attention par rapport à h en 0, ce qui donne

$$F(t_n, y(t_n), h) = \sum_{m=0}^{p-1} \frac{h^m}{m!} \frac{\partial^m F}{\partial h^k}(t_n, y(t_n), 0) + \frac{h^p}{(p-1)!} \int_0^1 (1-u)^{p-1} \frac{\partial^p F}{\partial h^p}(t_n, y(t_n), uh)\, du.$$

Reportant dans l'erreur de consistance en tenant compte de (4.1.13), il vient

$$\varepsilon_n = \frac{h^{p+1}}{p!}\int_0^1 (1-s)^p y^{(p+1)}(t_n+sh)\,ds - \frac{h^{p+1}}{(p-1)!}\int_0^1 (1-u)^{p-1}\frac{\partial^p F}{\partial h^p}(t_n, y(t_n), uh)\,du$$

d'où

$$\begin{aligned}\|\varepsilon_n\| &\leq h^{p+1}\|\frac{1}{p!}\int_0^1 (1-s)^p y^{(p+1)}(t_n+sh)\,ds - \frac{1}{(p-1)!}\int_0^1 (1-u)^{p-1}\frac{\partial^p F}{\partial h^p}(t_n, y(t_n), uh)\,du\| \\ &\leq \frac{h^{p+1}}{p!}\left(\int_0^1 \|(1-s)^p y^{(p+1)}(t_n+sh)\|, ds + \int_0^1 \|p(1-u)^{p-1}\frac{\partial^p F}{\partial h^p}(t_n, y(t_n), uh)\|\,du\right)\end{aligned}$$

$$\|\varepsilon_n\| = h^{p+1}R_n,$$

où le reste R_n exprimé avec les intégrales ci-dessus est tel que $\|R_n\| \leq C$ avec C indépendante de h et de n, vu que F est de classe C^p et y de classe C^{p+1} et que leurs arguments restent dans des compacts. ♢

Cette condition n'est pas extraordinairement utile en pratique, il est le plus souvent plus indiqué d'utiliser directement des développements de Taylor pour estimer l'erreur de consistance comme on l'a fait précédemment pour les schémas d'Euler et d'Euler modifié.

Pour conclure ce paragraphe, reprenons de nouveau notre exemple favori [18] et utilisons-le pour tester les trois schémas de la Figure 4.3, plus le schéma symplectique. La Figure 4.4 montre les variations de l'inclinaison en fonction du temps pour trois discrétisations différentes et chacun des algorithmes. La Figure 4.5 montre les trajectoires dans l'espace des phases et la Figure 4.6 montre les variations du hamiltonien en fonction du temps. rappelons que ce dernier est constant dans le mouvement réel.

Le quatrième schéma illustré dans ces figures est le schéma d'Euler symplectique. On remarque qu'il possède la propriété intéressante de conserver en moyenne le hamiltonien.

4.2 Schémas implicites

On a déjà introduit quelques schémas implicites (Euler rétrograde, Crank-Nicolson). Avant de se demander s'ils convergent ou non et à quoi ils peuvent bien servir, il faut d'abord s'interroger sur leur caractère bien défini. Une fois ces schémas étudiés, il faut enfin se demander comment les mettre en œuvre en pratique. En effet, ni l'une ni l'autre de ces questions n'est évidente a priori.

Considérons donc un schéma implicite générique à un pas de la forme

$$y_0 = y(0), \quad y_{n+1} = y_n + h\Phi(t_{n+1}, y_{n+1}, t_n, y_n, h), \tag{4.2.1}$$

supposé résoudre le problème de Cauchy pour l'EDO $y'(t) = f(t, y(t))$. À chaque pas de temps, il s'agit de résoudre l'équation en général non linéaire

$$z = \varphi(z), \text{ avec } \varphi(z) = y_n + h\Phi(t_{n+1}, z, t_n, y_n, h). \tag{4.2.2}$$

18. Celui du pendule bien sûr !

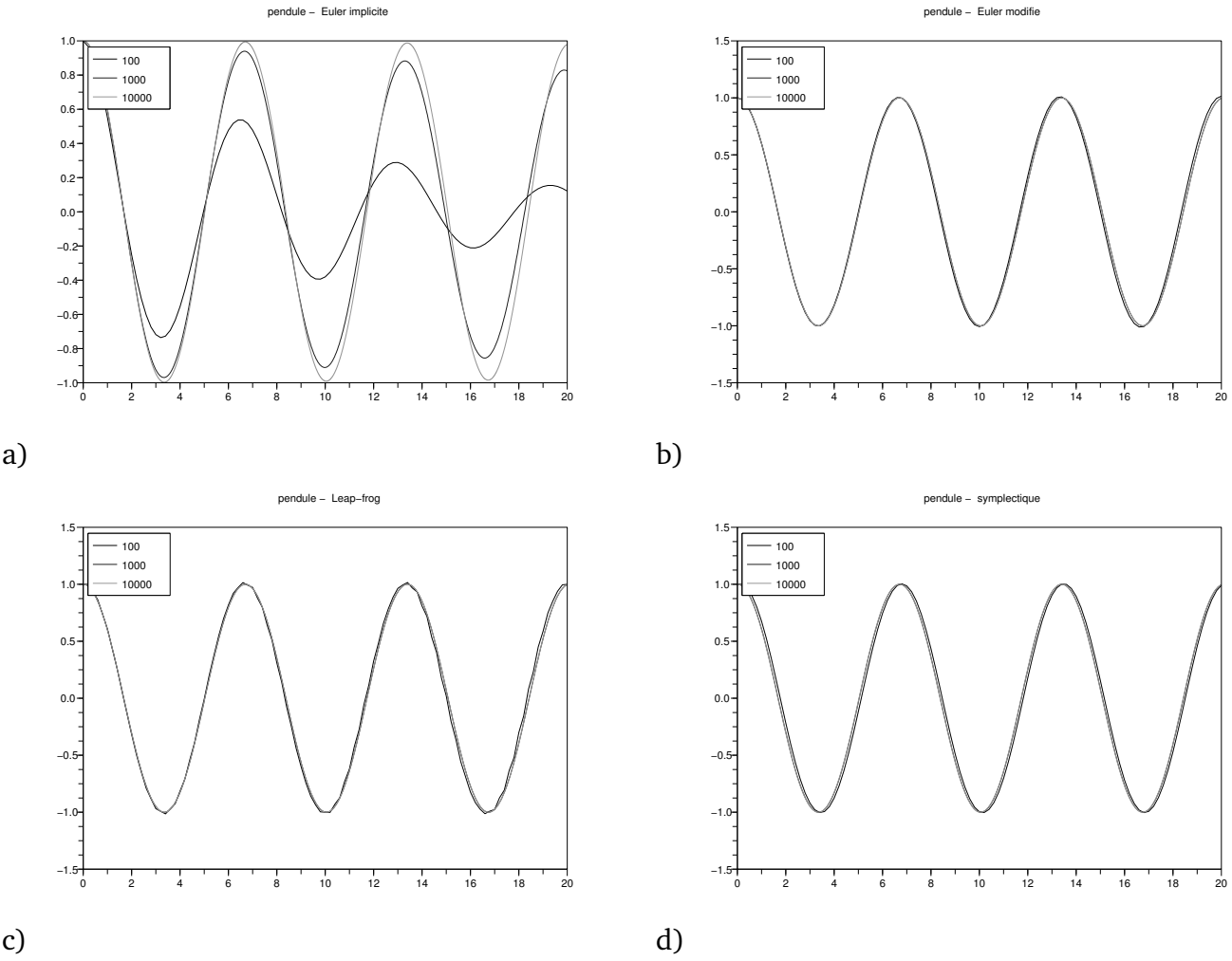


FIGURE 4.4 – Variations de l'inclinaison du pendule en fonction du temps pour trois discrétisations différentes et les schémas a) d'Euler implicite, b) Euler modifié, c) leapfrog d) symplectique.

Un tel z est appelé un *point fixe* de φ, qui est une application de $\mathbb{R}^m$ dans $\mathbb{R}^m$. Bien sûr, en général il n'y a aucune raison pour qu'une telle équation ait au moins une solution d'une part, ou n'en ait qu'au plus une d'autre part. Néanmoins, l'existence et l'unicité d'un point fixe sont assurées si la fonction φ est strictement contractante.

Définition 4.2.1 *Soit (E, d) un espace métrique et φ une application de E dans lui-même. On dit que φ est* strictement contractante *s'il existe $k \in [0, 1[$ tel que pour tout $(x_1, x_2) \in E^2$,*

$$d(\varphi(x_1), \varphi(x_2)) \leq k d(x_1, x_2).$$

En effet, on a le théorème de point fixe de Picard ou de Banach suivant :

Théorème 4.2.2 *Soit (E, d) un espace métrique complet et φ une application strictement contractante de E dans lui-même. Alors φ admet un point fixe unique x^*. De plus, pour tout $x_0 \in E$, la suite récurrente $(x_p)_{p\in\mathbb{N}}$ définie par $x_{p+1} = \varphi(x_p)$ pour tout $p \geq 0$ converge vers le point fixe x^*.*

Démonstration. Montrons d'abord l'unicité. Soient x^* et $\tilde{x}^*$ des points fixes de φ. On a donc $\varphi(x^*) = x^*$ et $\varphi(\tilde{x}^*) = \tilde{x}^*$. Comme φ est strictement contractante, on en déduit que

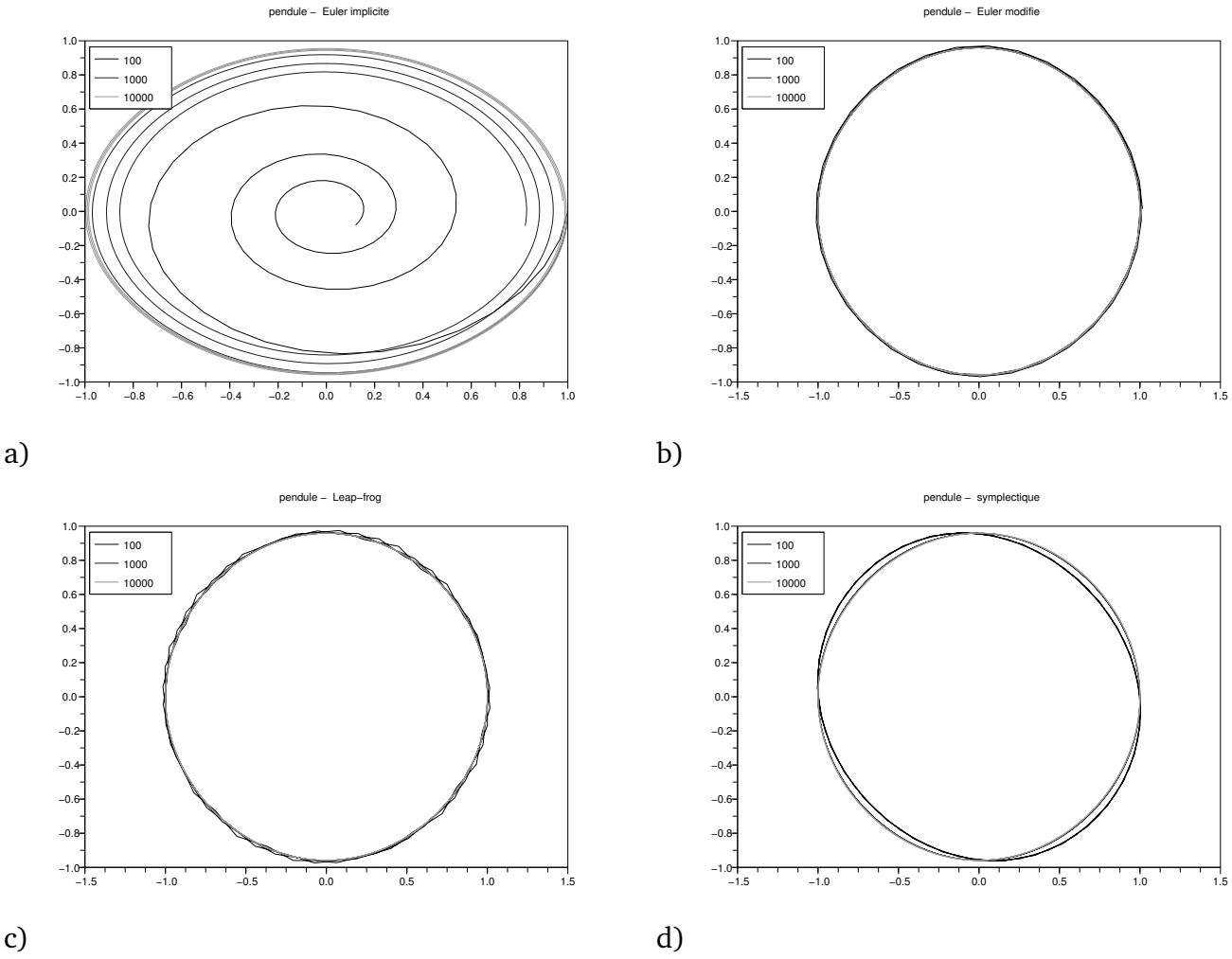


FIGURE 4.5 – Trajectoires du pendule dans l'espace des phases pour trois discrétisations différentes et les schémas a) d'Euler implicite, b) Euler modifié, c) leapfrog, d) symplectique.

$d(x^*, \tilde{x}^*) \leq k d(x^*, \tilde{x}^*)$, soit encore $(1-k)d(x^*, \tilde{x}^*) \leq 0$. Comme $k < 1$, il s'ensuit que $1 - k > 0$, donc nécessairement $d(x^*, \tilde{x}^*) = 0$, c'est-à-dire $x^* = \tilde{x}^*$. [19]

Montrons ensuite l'existence du point fixe. Soit $x_0 \in E$ un point quelconque et (x_p) la suite itérée associée. On a alors

$$d(x_{p+1}, x_p) = d(\varphi(x_p), \varphi(x_{p-1})) \leq k d(x_p, x_{p-1}),$$

d'où par une récurrence immédiate (mais à faire quand même en exercice) $d(x_{p+1}, x_p) \leq k^p d(x_1, x_0)$ pour tout p. Pour tout entier $q > p$, il vient par l'inégalité triangulaire

$$d(x_q, x_p) \leq \sum_{n=p}^{q-1} d(x_{n+1}, x_n) \leq \left(\sum_{n=p}^{q-1} k^n \right) d(x_1, x_0).$$

Or

$$\sum_{n=p}^{q-1} k^n \leq \sum_{n=p}^{\infty} k^n = \frac{k^p}{1-k},$$

19. La complétude de E ne joue aucun rôle pour l'unicité.

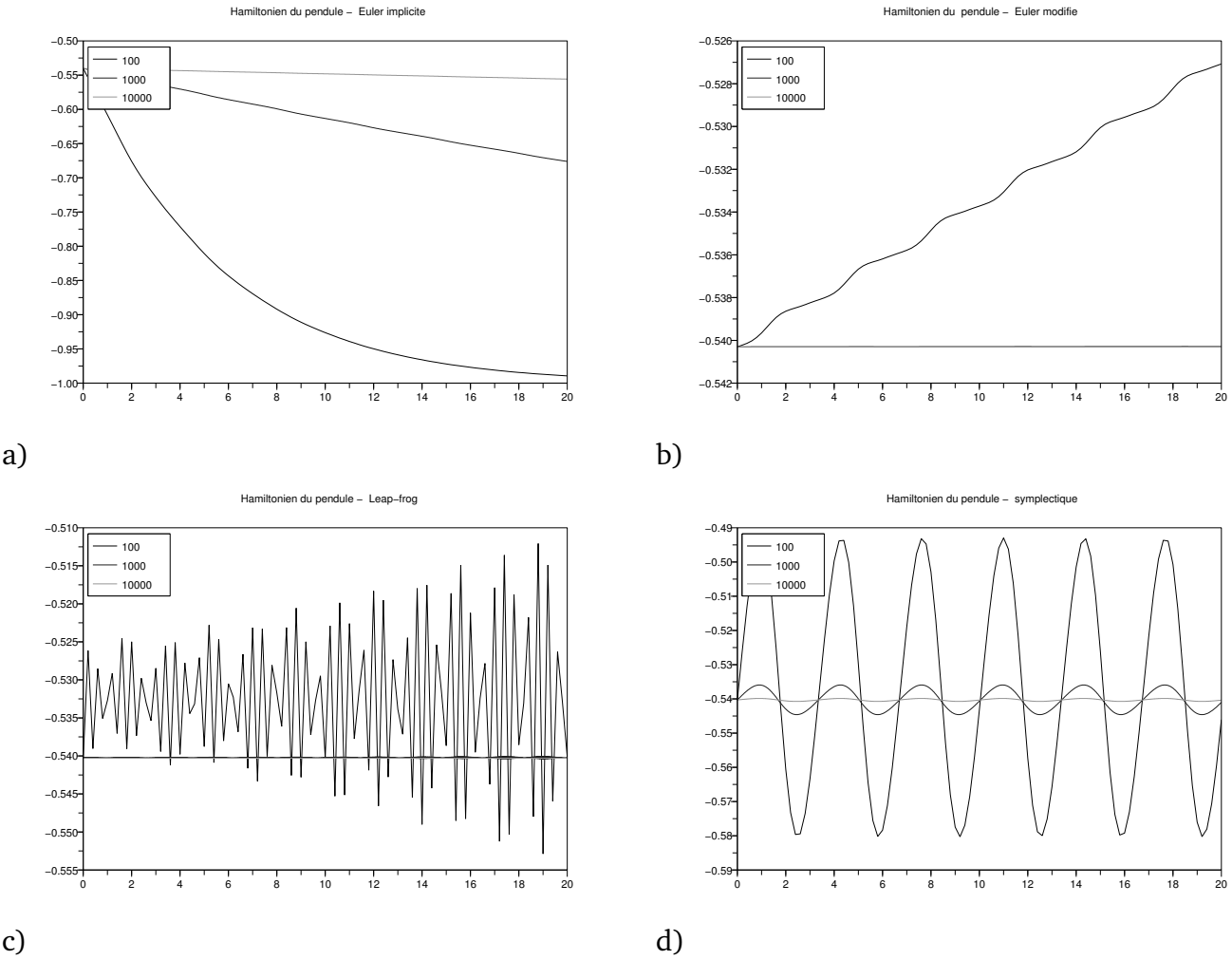


FIGURE 4.6 – Variations du hamiltonien du pendule en fonction du temps pour trois discrétisations différentes et les schémas a) d'Euler implicite, b) Euler modifié, c) leapfrog d) symplectique.

puisque $0 \leq k < 1$. On a donc finalement $d(x_p, x_q) \leq k^p \frac{d(x_1,x_0)}{1-k}$ avec $0 \leq k < 1$, ce qui montre que la suite (x_p) est de Cauchy. Comme E est complet pour la distance d, la suite (x_p) converge vers une limite x^*. Comme φ est contractante, elle est continue, et en passant à la limite dans l'égalité $x_{p+1} = \varphi(x_p)$ quand $p \longrightarrow +\infty$, on obtient $x^* = \varphi(x^*)$. ◊

Revenons au schéma implicite général à un pas, [20] lequel est défini par la donnée d'une fonction $\Phi\colon [0,T] \times \mathbb{R}^m \times [0,T] \times \mathbb{R}^m \times [0,1] \longrightarrow \mathbb{R}^m$. Notons $\Phi(s,z,t,y,h)$ l'image d'un élément générique par cette fonction et φ l'application $z \mapsto y + h\Phi(s,z,t,y,h)$ pour s,t,y,h fixés. Cette application dépend de (s,t,y,h) mais on ne l'écrit pas explicitement.

Proposition 4.2.3 *Si Φ est globalement lipschitzienne par rapport à z, uniformément par rapport à (s,t,y), alors il existe $h_0 > 0$ indépendant de (s,t,y) tel que l'application φ soit strictement contractante pour tout $h \leq h_0$.*

Démonstration. Soit M la constante de Lipschitz uniforme de Φ par rapport à z. On a alors,

20. Le cas des méthodes de Runge-Kutta implicites que l'on verra plus loin étant légèrement différent.

pour tous $z_1, z_2 \in \mathbb{R}^m$,

$$\|\varphi(z_1) - \varphi(z_2)\| = h\|\Phi(s, z_1, t, y, h) - \Phi(s, z_2, t, y, h)\| \leq hM\|z_1 - z_2\|.$$

Par conséquent, si l'on choisit $h_0 > 0$ tel que $h_0 < 1/M$, alors $hM < 1$ pour tout $h \leq h_0$. ◊

Comme $\mathbb{R}^m$ est complet pour la distance induite par n'importe quelle norme, on en déduit le

Corollaire 4.2.4 *Le schéma implicite* (4.2.1) *est bien défini pour tout* $h \leq h_0$.

Pour un schéma implicite, on ne prendra donc pas la variable h dans l'intervalle un peu arbitraire $[0, 1]$ comme précédemment, mais dans $[0, h_0]$ ce qui ne change essentiellement rien.

Dans le cas du schéma d'Euler implicite, où $\Phi(s, z, t, y, h) = f(s, z)$, on voit que la condition est satisfaite dès que que la fonction f est globalement lipschitzienne par rapport à y, uniformément par rapport à t, c'est-à-dire une partie des hypothèses du théorème de Cauchy-Lipschitz global. Par conséquent, le schéma d'Euler implicite est bien défini pour $h < 1/L$. Bien sûr, si on ne connaît pas L, il est difficile de deviner ce que " h suffisamment petit " signifie quantitativement parlant. Intéressons-nous maintenant à la convergence du schéma implicite vers la solution exacte quand le pas de discrétisation h tend vers 0. On pourrait à juste titre considérer qu'il n'y a rien à faire, puisque le théorème de point fixe fournit pour h assez petit une fonction implicite $z = \Theta(s, t, y, h)$ et l'on peut donc écrire

$$y_{n+1} = y_n + h\Phi(t_n + h, \Theta(t_n + h, t_n, y_n, h), t_n, y_n, h),$$

d'où un schéma sous la forme (4.1.1) avec

$$F(t, y, h) = \Phi(t + h, \Theta(t + h, t, y, h), t, y, h). \tag{4.2.3}$$

Bien sûr, cette fonction F est férocement non explicite, mais l'analyse théorique de la convergence n'a que faire du caractère explicite ou pas de la fonction F, c'est une affaire de consistance et de stabilité.

On va quand même reprendre cette étude théorique à partir de la fonction Φ qui est elle explicite. [21] On se placera toujours dans l'hypothèse que h est suffisamment petit pour que le schéma soit bien défini, c'est-à-dire $hM < 1$ où M désigne la constante de Lipschitz de Φ par rapport à z, uniforme par rapport aux autres variables.

Définition 4.2.5 *Le schéma* (4.2.1) *est* stable *s'il existe une constante C indépendante de N telle que, pour toute suite de vecteurs $(\eta_n)_{0\leq n\leq N}$, les suites $(y_n)_{0\leq n\leq N}$ et $(z_n)_{0\leq n\leq N}$ de $\mathbb{R}^m$ satisfaisant respectivement*

$$y_0 \in \mathbb{R}^m \text{ et } y_{n+1} = y_n + h\Phi(t_{n+1}, y_{n+1}, t_n, y_n, h) \text{ pour } 0 \leq n \leq N - 1$$

et

$$z_0 = y_0 + \eta_0 \text{ et } z_{n+1} = z_n + h\Phi(t_{n+1}, z_{n+1}, t_n, z_n, h) + \eta_{n+1} \text{ pour } 0 \leq n \leq N - 1,$$

21. Il n'est pas utile d'en rajouter au delà de toute mesure dans l'implicite.

sont telles que

$$\max_{0\le n\le N} \|z_n - y_n\| \le C \sum_{n=0}^{N} \|\eta_n\|. \tag{4.2.4}$$

L'erreur de consistance du schéma (4.2.1) *est la quantité*

$$\varepsilon_n = y(t_{n+1}) - y(t_n) - h\Phi(t_{n+1}, y(t_{n+1}), t_n, y(t_n), h),$$

où y est une solution de l'EDO. Le schéma (4.2.1) *est consistant si pour toute solution y de* (2.1.1)*, on a*

$$\lim_{h\to 0} \sum_{n=0}^{N-1} \|\varepsilon_n\| = 0.$$

Une fois ces définitions posées, on a par exactement la même démonstration que dans le cas explicite

Théorème 4.2.6 *Un schéma implicite* (4.2.1) *stable et consistant est convergent.*

En termes de conditions suffisantes de stabilité et de convergence, on a des résultats également très semblables.

Proposition 4.2.7 *Soit Φ une fonction globalement lipschitzienne de constante M par rapport à y et z, uniformément par rapport à (s,t,h) et soit $0 < h_0 < 1/M$. Alors pour tout $h \le h_0$, le schéma* (4.2.1) *est stable. Si Φ est continue par rapport à l'ensemble de ses arguments et que $\Phi(t,y,t,y,0) = f(t,y)$ pour tous (t,y), alors le schéma est consistant.*

Démonstration. On remarque d'abord que la constante de Lipschitz M par rapport à y et z vaut aussi a fortiori pour z seul. Le schéma est donc bien défini pour tout $h \le h_0$.

Montrons sa stabilité. Soient deux suites y_n et z_n vérifiant les hypothèses ci-dessus. On a

$$\begin{aligned}\|z_{n+1} - y_{n+1}\| &= \|z_n - y_n + h\left(\Phi(t_{n+1}, z_{n+1}, t_n, z_n, h) - \Phi(t_{n+1}, y_{n+1}, t_n, y_n, h)\right) + \eta_{n+1}\| \\ &\le \|z_n - y_n\| + h\|\Phi(t_{n+1}, z_{n+1}, t_n, z_n, h) - \Phi(t_{n+1}, y_{n+1}, t_n, y_n, h)\| + \|\eta_{n+1}\| \\ &\le \|z_n - y_n\| + hM\|z_{n+1} - y_{n+1}\| + hM\|z_n - y_n\| + \|\eta_{n+1}\|.\end{aligned}$$

P ar conséquent,

$$(1 - hM)\|z_{n+1} - y_{n+1}\| \le (1 + hM)\|z_n - y_n\| + \|\eta_{n+1}\|,$$

et pour $h \le h_0$, on a $1 - hM > 0$, donc en divisant par $1 - hM$,

$$\|z_{n+1} - y_{n+1}\| \le \frac{1 + hM}{1 - hM}\|z_n - y_n\| + \frac{1}{1 - hM}\|\eta_{n+1}\|.$$

On applique alors le lemme de Grönwall discret version 2 4.1.5, avec $u_n = \|z_n - y_n\|$, $\lambda = \frac{1+hM}{1-hM} - 1 = \frac{2hM}{1-hM}$, et $\mu_n = \frac{1}{1-hM}\|\eta_{n+1}\|$ pour en déduire que

$$\|z_n - y_n\| \le \frac{e^{\frac{2MT}{1-hM}}}{1 - hM} \sum_{k=0}^{N} \|\eta_k\|,$$

voir la démonstration de la Proposition 4.1.3 pour le détail de la preuve. La constante qui apparaît n'est pas tout à fait indépendante de h, mais sa dépendance n'est pas bien méchante dans la mesure où pour tout $h \leq h_0$, on a $\frac{e^{\frac{2MT}{1-hM}}}{1-hM} \leq \frac{e^{\frac{2MT}{1-h_0M}}}{1-h_0M}$, et la méthode est stable pour $h \leq h_0$.

On laisse la démonstration de la condition suffisante de consistance en exercice, c'est essentiellement la même que dans le cas explicite. ◊

Dans le cas du schéma d'Euler implicite, $\Phi(s,z,t,y,h) = f(s,z)$, on a donc sous les hypothèses du théorème de Cauchy-Lipschitz global que $M = L$ et le schéma est bien défini et stable pour $h \leq h_0 < 1/L$, et il est consistant. Il est donc convergent.

Les questions d'ordre de méthode dans le cas implicite se formulent exactement de la même façon que dans le cas explicite. Regardons ce que cela donne pour la méthode d'Euler implicite. On suppose donc la fonction f de classe C^1 de telle sorte que toute solution y soit C^2. On calcule l'erreur de consistance

$$\varepsilon_n = y(t_{n+1}) - y(t_n) - hf\big(t_{n+1}, y(t_{n+1})\big) = -\big(y(t_n) - y(t_{n+1}) - (-h)y'(t_{n+1})\big),$$

d'où comme $t_n = t_{n+1} - h$, par l'inégalité de Taylor-Lagrange

$$\|\varepsilon_n\| \leq \frac{h^2}{2} \max_{[0,T]} \|y''\|,$$

et la méthode d'Euler implicite est donc d'ordre 1. On a par conséquent une estimation d'erreur en $O(h)$. On pourra à titre d'exercice traiter le cas du schéma de Crank-Nicolson qui se trouve être d'ordre 2.

Quelques mots maintenant sur les aspects numériques de mise en œuvre des méthodes implicites. Pour une fonction Φ générale, on ne dispose pas de formule explicite donnant y_{n+1}. Il existe bien une fonction implicite, mais on n'a pas d'algorithme pour l'implémenter. Il faut donc en pratique approcher y_{n+1} de manière itérative. La dernière partie du théorème de point fixe nous suggère de construire à chaque pas de temps la suite récurrente

$$y_{n+1}^0 = y_n, \quad y_{n+1}^{p+1} = y_n + h\Phi(t_{n+1}, y_{n+1}^p, t_n, y_n, h), \text{ pour } p = 0, 1, \ldots$$

dont on sait qu'elle converge vers y_{n+1} quand $p \longrightarrow +\infty$ quand h est suffisamment petit. Comme on ne peut pas calculer une infinité de fois, on convient de s'arrêter d'itérer quand y_{n+1}^p a *numériquement convergé* vers sa limite y_{n+1}. Par cela on entend que l'on compare la norme de la différence entre deux itérés successifs avec une tolérance α petite et fixée au préalable, et que l'on s'arrête si l'on est descendu en dessous de cette tolérance, *i.e.*, la première fois que $\|y_{n+1}^{p+1} - y_{n+1}^p\| \leq \alpha$, on s'arrête et on adopte l'approximation y_{n+1}^{p+1} comme valeur de y_{n+1}. En effet, la suite $\|y_{n+1}^{p+1} - y_{n+1}^p\|$ est strictement décroissante tendant vers 0 dès que $h < 1/M$, donc on a la garantie que le processus s'arrête en un nombre fini d'opérations. [22]

En fait, c'est un peu plus compliqué que cela, puisqu'on ne dispose pas de la valeur exacte de y_n, mais d'une approximation calculée de façon analogue à l'étape précédente et ainsi de suite depuis le début. Naturellement tout cela constitue une source supplémentaire d'erreur, mais peu importe, il suffit de la contrôler et la stabilité du schéma fera le reste.

22. De plus, $\|y_{n+1} - y_{n+1}^{p+1}\| \leq \frac{\alpha}{1-hM}$.

Si l'on prend un peu de recul et que l'on regarde ce que l'on a fait, on s'aperçoit que l'on n'a pas implémenté ainsi exactement le schéma implicite, c'est d'ailleurs impossible pour un Φ général, mais que l'on a en fait implémenté un schéma explicite [23] correspondant à des itérations de la fonction Φ en nombre non fixé à l'avance, mais dépendant du déroulement de l'algorithme. On pourrait écrire la fonction explicite correspondante, mais elle est encore plus tordue que (4.2.3). De toutes façons, tout ceci n'est pas bien grave, même un schéma explicite ne peut pas être implémenté exactement pour la simple raison qu'un ordinateur qui calcule en virgule flottante commet donc systématiquement des erreurs d'arrondi et des erreurs d'évaluation de fonctions. On pourrait analyser plus finement toutes ces erreurs, voir le paragraphe 8.1, mais *in fine*, c'est la stabilité du schéma qui permet d'avoir (une certaine) confiance dans le résultat qui sort de la machine.

L'objet du chapitre suivant est précisément d'introduire une méthode performante de résolution numériques d'équations non linéaires qui permet d'approcher y_{n+1}.

23. C'est normal, on ne peut faire de calcul numérique autre qu'explicite.

Chapitre 5

Méthode de Newton

Le théorème du point fixe nous donne l'existence et l'unicité de la solution de $\varphi(z) = z$ quand φ est strictement contractante, et en même temps, une méthode itérative pour approcher la solution. Nous allons voir dans ce paragraphe un algorithme nettement plus efficace pour résoudre approximativement les équations non linéaires mises sous la forme générique $g(z) = 0$. Pour trouver la solution du schéma implicite, on résoudra donc à chaque pas de temps l'équation $g(z) = 0$ après avoir posé $g(z) = \varphi(z) - z$. Il s'agit de la *méthode de Newton*.

5.1 Présentation en dimension 1

Soit c une racine de g, avec g au moins de classe C^1 sur l'intervalle $[a, b]$. On suppose qu'on connaît une valeur approchée x_0 de la racine, d'une façon ou d'une autre. L' idée de la méthode de Newton est de remplacer la courbe représentative de g par sa tangente en x_0, d'équation

$$Y = g'(x_0)(X - x_0) + g(x_0),$$

et de considérer l'intersection de cette tangente avec l'axe des abscisses $Y = 0$, ce qui donne le point suivant

$$x_1 = x_0 - \frac{g(x_0)}{g'(x_0)}.$$

En général, si x_0 est choisi pas trop loin de la racine c de l'équation, on sent bien que x_1 en est une bien meilleure approximation que x_0, *cf.* Figure 5.1. On recommence alors l'opération, ce qui conduit à la suite récurrente

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

Si cette suite est bien définie, *i.e.*, si on n'a pas divisé par zéro en cours de route et n'est pas sorti de l'intervalle de définition de g, et si elle converge vers une valeur c, alors on a $c = c - \frac{g(c)}{g'(c)}$, soit $g(c) = 0$. Avec les notations précédentes, on a en fait écrit une itération de point fixe pour la fonction $h(x) = x - \frac{g(x)}{g'(x)}$. Comme $h'(x) = 1 - \frac{g'(x)^2 - g(x)g''(x)}{g'(x)^2} = \frac{g(x)g''(x)}{g'(x)^2}$, la racine est bien un point fixe super attractif de h si g' ne s'y annule pas.

Analysons plus précisément la méthode de Newton.

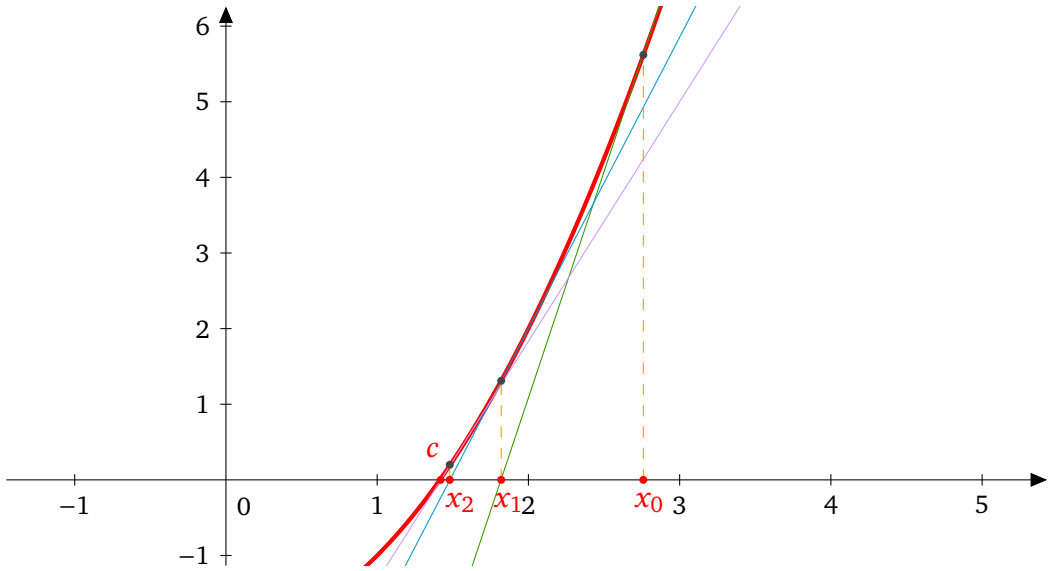


FIGURE 5.1 – La méthode de Newton (zoomer à l'écran pour y voir de plus près).

Théorème 5.1.1 *On suppose que g est de classe C^2 sur l'intervalle $I = [c-r, c+r]$, pour un certain $r > 0$ et que g' ne s'annule pas sur I. Soit*

$$M = \max_{x\in I} |g''(x)|, m = \min_{x\in I} |g'(x)| \text{ et } \alpha = \min\left(r, \frac{2m}{M}\right).$$

Alors pour tout point initial $x_0 \in]c-\alpha, c+\alpha[$, la suite de Newton x_n est bien définie pour tout n et converge vers c quand $n \longrightarrow +\infty$, avec l'estimation

$$|x_n - c| \leq \frac{1}{K}(K|x_0 - c|)^{2^n}, \tag{5.1.1}$$

où $K = \frac{M}{2m}$.

Démonstration. La fonction g' ne s'annulant pas sur I, on peut y définir une fonction h par $h(x) = x - \frac{g(x)}{g'(x)}$. Par définition de la méthode de Newton, si x_n est bien défini, on a $x_{n+1} = h(x_n)$. La suite de Newton est donc celle des itérées de x_0 par h et il suffit par conséquent de montrer que h admet un intervalle invariant pour montrer que la suite est bien définie.

Pour tout $x \in I$, la formule de Taylor-Lagrange nous dit que

$$0 = g(c) = g(x) + (c-x)g'(x) + \frac{(c-x)^2}{2}g''(\xi),$$

pour un certain ξ situé entre c et x. Divisant par $g'(x)$, qui est non nul sur I, on en déduit que

$$\left(x - \frac{g(x)}{g'(x)}\right) - c = \frac{(c-x)^2}{2}\frac{g''(\xi)}{g'(x)}.$$

On voit donc que

$$|h(x) - c| \leq \frac{M}{2m}|x-c|^2.$$

On pose alors $\alpha = \min\left(r, \frac{2m}{M}\right)$. Si $x \in]c-\alpha, c+\alpha[\subset I$, on a donc $|x-c| < \alpha$, d'où

$$|h(x) - c| \leq \frac{M}{2m}\alpha^2 \leq \alpha,$$

puisque $\alpha \leq \frac{2m}{M}$. On en déduit que $h(x) \in]c-\alpha, c+\alpha[\subset I$. On a trouvé un intervalle invariant par h et la suite de Newton x_n est donc bien définie si x_0 est pris dans cet intervalle invariant $]c-\alpha, c+\alpha[$.

De plus, on obtient pour tout n

$$|x_{n+1} - c| \leq K|x_n - c|^2,$$

avec $K = \frac{M}{2m}$, d'où l'estimation (5.1.1) par la Proposition 5.1.2 ci-dessous avec $m_0 = 0$ car $K|x_0 - \bar{x}| < 1$. ◊

Proposition 5.1.2 *Soit x_n une suite réelle qui converge vers $x \in \mathbb{R}$. On dit que la convergence est d'ordre (au moins) $\alpha \geq 1$ quand il existe $\lambda > 0$, avec $\lambda < 1$ quand $\alpha = 1$, et n_0 tels que pour tout $n \geq n_0$ $|x_{n+1} - x| \leq \lambda |x_n - x|^\alpha$. Quand $\alpha = 1$, on dit que la convergence est* linéaire. *Quand $\alpha = 2$, on dit qu'elle est* quadratique. *Enfin, quand $\frac{|x_{n+1}-x|}{|x_n-x|} \to 0$ quand $n \to +\infty$ (en supposant $x_n \neq x$), on dit que la convergence est* surlinéaire.

Si la convergence est linéaire, alors on a pour n assez grand $e_n \leq C\lambda^n$, pour une certaine constante C (on rappelle que dans ce cas $\lambda < 1$). Si la convergence est quadratique, alors on a pour n assez grand

$$e_n \leq C\mu^{2^n}, \tag{5.1.2}$$

pour une autre constante C et pour un certain $\mu < 1$.

Démonstration. On considère $n \geq n_0$. Traitons d'abord la convergence linéaire. On démarre un raisonnement par récurrence en remarquant que $e_{n_0} \leq e_{n_0}$. On suppose donc en guise d'hypothèse de récurrence que $e_n \leq e_{n_0}\lambda^{n-n_0}$. On en déduit que

$$e_{n+1} \leq \lambda e_n \leq e_{n_0}\lambda^{n+1-n_0}.$$

On a ainsi établi l'estimation voulue, avec $C = e_{n_0}\lambda^{-n_0}$.

Pour la convergence quadratique, on a par hypothèse que $e_n \to 0$ quand $n \to +\infty$. Il existe donc n_1 tel que pour tout $n \geq n_1$, $\lambda e_n < 1$. On pose alors $m_0 = \max(n_0, n_1)$ et on part du même scoop, mais en m_0. L'hypothèse de récurrence pour $n \geq m_0$ est ici $e_n \leq \frac{1}{\lambda}(\lambda e_{m_0})^{2^{n-m_0}}$, qui est bien satisfaite pour $n = m_0$. On en déduit que

$$e_{n+1} \leq \lambda(e_n)^2 \leq \lambda\left(\frac{1}{\lambda}(\lambda e_{m_0})^{2^{n-m_0}}\right)^2 = \frac{1}{\lambda}(\lambda e_{m_0})^{2\times 2^{n-m_0}} = \frac{1}{\lambda}(\lambda e_{m_0})^{2^{n+1-m_0}}.$$

On a ainsi établi (5.1.2), avec $\mu = (\lambda e_{m_0})^{2^{-m_0}}$ et $C = \lambda^{-1}$. ◊

Non seulement la méthode de Newton converge sous les hypothèses précédentes, mais elle converge quadratiquement. Par contre, la méthode de Newton demande plus de calculs que les précédentes méthodes, en particulier celui de f', qui n'est pas forcément si simple que ça quand f est elle-même compliquée, et sa convergence n'est pas assurée pour toute valeur initiale de l'itération. Il vaut mieux être suffisamment près de la racine au départ, c'est-à-dire

l'avoir déjà assez bien localisée, par exemple par dichotomie. En effet, le Théorème 5.1.1 donne bien un intervalle de convergence, mais celui-ci est centré sur la racine c inconnue. Si son existence est intéressante, l'utilité pratique de cet intervalle est donc discutable.

On voit par ailleurs que la constante K est d'autant plus grande que la dérivée seconde est grande et que la dérivée première est petite. Ce sont des situations défavorables pour la convergence de la méthode de Newton, à la fois en termes d'intervalle et en termes de vitesse de convergence. Il vaut mieux que la dérivée seconde soit petite (moralement, f très proche de son application affine tangente) avec une forte pente.

Gardons quand même à l'esprit qu'en pratique, quand elle marche, la méthode de Newton est terriblement efficace.

Prenons maintenant l'exemple de la fonction $g(x) = x^2 - 2$. On a $g'(x) = 2x$ donc la méthode de Newton consiste à effectuer l'itération

$$x_{n+1} = x_n - \frac{x_n^2 - 2}{2x_n} = \frac{x_n}{2} + \frac{1}{x_n}.$$

Quand on la démarre à $x_0 = 1$, c'est exactement la méthode de Héron pour approcher $\sqrt{2}$. Celle-ci est donc effectivement quadratique, de nombreux siècles avant Newton. Bien sûr, l'idée de Héron n'était pas du tout celle de Newton, mais se basait sur des considérations géométriques d'aires de rectangles, voir Figure 5.2.

Plus généralement, pour tout $A > 0$, on peut considérer la fonction $g(x) = x^2 - A$, qui fournit des itérations de Héron

$$x_{n+1} = x_n - \frac{x_n^2 - A}{2x_n} = \frac{x_n}{2} + \frac{A}{2x_n} = \frac{1}{2}\Big(x_n + \frac{A}{x_n}\Big),$$

lesquelles convergent quadratiquement vers $\sqrt{A}$.

On passe à la méthode de Newton en dimension supérieure. On va même commencer par se placer dans une généralité telle qu'il n'y a même pas a priori de notion de dimension.

5.2 La méthode de Newton dans $\mathbb{R}^n$

On va s'intéresser ici à trouver les racines ou les zéros d'une application $g\colon \mathbb{R}^n \to \mathbb{R}^n$, ou d'une partie de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}^n$, $g(x) = 0$. Cela a un sens, puisque $\mathbb{R}^n$ contient bien un 0. On a déjà vu le cas $n = 1$.

Première remarque, pourquoi $g\colon \mathbb{R}^n \to \mathbb{R}^n$ et pas plus généralement $g\colon \mathbb{R}^n \to \mathbb{R}^m$ avec m pas nécessairement égal à n ? La raison en est qu'il n'est raisonnable d'espérer trouver des zéros isolés que dans le cas $n = m$. Génériquement, si $m < n$ on va trouver soit l'ensemble vide, soit des zéros qui forment une « grosse » partie de $\mathbb{R}^n$. Par exemple, si $m = 1$ et $n = 2$, on va typiquement tomber sur une courbe, si $m = 1$ et $n = 3$ sur une surface, etc. Par contre, si $m > n$, on va génériquement trouver l'ensemble vide.

Pour se convaincre que ceci est plausible, il suffit de regarder le cas où g est affine ou linéaire. On est ici en train de discuter de la résolution d'un système linéaire de m équations à n inconnues. L'ensemble des zéros est soit un singleton, soit un espace affine de dimension supérieure à 1, soit l'ensemble vide, selon le rang, le noyau et l'image de l'application linéaire associée, et le second membre du système linéaire. La situation est assez semblable dans le

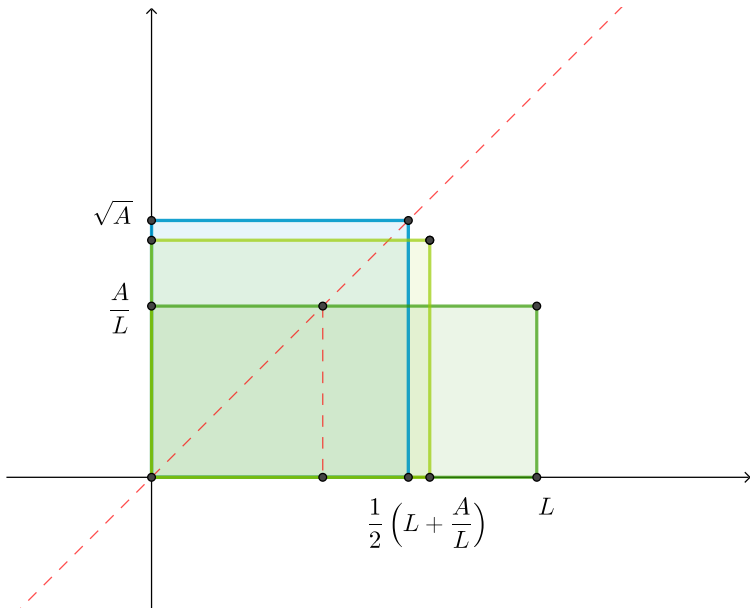


FIGURE 5.2 – La méthode de Héron du point de vue géométrique de Héron : le rectangle vert pâle a la même aire A que le rectangle vert foncé, mais est nettement plus carré, et ainsi de suite, et ainsi de suite...

cas général non linéaire, avec plus de variété [1] naturellement, et il est impossible de conclure de façon générale et universelle comme avec les systèmes linéaires.

Cette section suppose connus les bases du calcul différentiel, reprises en annexe B.

5.2.1 $g(x) = 0$ du point de vue théorique

Étant donné un ouvert U de $\mathbb{R}^n$ et g une application de classe C^1 de U à valeurs dans $\mathbb{R}^n$, on cherche à trouver les $x \in U$ tels que $g(x) = 0$.

Tout d'abord, est-ce un but raisonnable ? On sait bien que oui si $n = 1$, mais en général, ce n'est pas si clair. En effet, dès $n = 2$, les dessins qui sont convaincants pour $n = 1$ doivent être faits en dimension $2 \times 2 = 4$, car le graphe de g est alors un sous-ensemble de $\mathbb{R}^4$. Du coup, on n'y voit plus grand-chose.

Regardons à nouveau et plus précisément ce qui se passe quand g est une fonction affine et $U = \mathbb{R}^n$, c'est-à-dire que $g(x) = Ax - b$ où A est une matrice $n \times n$ et $b \in \mathbb{R}^n$. Les zéros de g sont donc exactement les solutions de l'équation $Ax = b$. On reconnaît un système linéaire de n équations à n inconnues. La discussion de ce système se fait classiquement selon que A est inversible ou non.

Si A est inversible, alors pour tout $b \in \mathbb{R}^n$, il y a une solution unique, donc isolée, qui est donnée par $x = A^{-1}b$.

Si A n'est pas inversible, alors son noyau n'est pas réduit au vecteur nul. Si b appartient à

1. Private joke liée à la géométrie différentielle.

l'image de A, alors l'ensemble des solutions est de la forme $x_p + \ker A$, où x_p est une solution particulière. C'est un sous-espace affine de dimension supérieure à 1, en particulier les solutions ne sont pas isolées. Si par contre b n'appartient pas à l'image de A, alors l'ensemble des solutions est vide.

Quand on prend une matrice carrée « au hasard » (sans donner un sens précis à cette expression), celle-ci va presque sûrement être inversible et on sera dans la première situation. Dit autrement, il faut vraiment soit jouer de malchance, soit s'y prendre de façon délibérée pour avoir une matrice A $n \times n$ non inversible. [2]

On peut penser que la situation générique avec une fonction g non nécessairement affine sera analogue, c'est-à-dire que sauf malchance, si on a un zéro de g, alors ce zéro va être isolé. Dans toute la suite, on va se donner une norme $\|\cdot\|$ sur $\mathbb{R}^n$ fixée une fois pour toutes. Pour fixer les idées, on va prendre $\|x\| = \max_i |x_i|$, mais cela peut aussi être n'importe quelle autre norme, c'est sans importance. [3] Pour cette norme, les boules sont des (hyper)cubes,  ce qui est aussi sans importance. Elles sont bien convexes.

Proposition 5.2.1 *Soit U un ouvert de $\mathbb{R}^n$ et $g\colon U \to \mathbb{R}^n$. On se donne $x \in U$ tel que $g(x) = 0$, on suppose que g y est différentiable et que $\nabla g(x)$ est inversible. Alors x est un zéro isolé de g.*

Démonstration. Définissons $h\colon U \to \mathbb{R}^n$ par $h(y) = (\nabla g(x))^{-1} g(y)$. On a donc $h(x) = 0$ et $\nabla h(x) = I$, la matrice identité, et par définition de la différentiabilité en x, on peut écrire

$$h(y) = h(x) + I(y-x) + \|y-x\|\varepsilon(y-x) = y - x + \|y-x\|\varepsilon(y-x),$$

pour tout $y \in U$. Toujours par définition de la différentiabilité, il existe une boule ouverte $B(x,r)$ telle que $\|\varepsilon(y-x)\| < \frac{1}{2}$ pour tout $y \in B(x,r)$. Par l'inégalité triangulaire, il s'ensuit que

$$\|h(y)\| > \frac{1}{2}\|y-x\|$$

ce qui montre que le seul endroit où h s'annule dans cette boule est au point x, et il en va de même de $g = (\nabla g(x))h$ puisque le noyau de $\nabla g(x)$ est réduit au vecteur nul. ◊

Remarque 5.2.1 En fait, on a beaucoup plus fort quand g est de classe C^1, en faisant appel au *théorème d'inversion locale*, un résultat de calcul différentiel qui nous dit que g est un C^1-difféomorphisme local, en particulier est localement injective, ce qui implique immédiatement que x est un zéro isolé. Le théorème d'inversion locale est d'ailleurs un autre exemple d'application du théorème de point fixe de Banach.

Bien sûr, si $\nabla g(x)$ n'est pas inversible, on ne peut rien dire. ◊

5.2.2 La méthode de Newton(-Raphson)

On est maintenant rassuré qu'il n'est pas a priori idiot de chercher des zéros isolés d'une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$. En dimension n quelconque, la méthode de Newton prend parfois le nom de méthode de Newton-Raphson. Le principe est le même qu'en dimension 1.

2. Par contre, il arrive couramment que l'on tombe sur une matrice inversible, donc bien gentille en théorie, mais qui est telle que le calcul effectif en pratique de la solution est extrêmement difficile, voire impossible avec une précision raisonnable.

3. Encore que ce choix conduise à des calculs parfois plus simples que d'autres choix.

Nous avons toujours un ouvert U de $\mathbb{R}^n$ et g une application de classe C^2 de U à valeurs dans $\mathbb{R}^n$. La classe C^2 est ici entendue [4] au sens où toutes les dérivées partielles secondes de toutes les composantes de g sont bien définies et continues sur U. On cherche à déterminer les $x \in U$ tels que $g(x) = 0$.

On suppose disposer d'une valeur approchée $x_0 \in U$ d'une solution x. Comme dans la méthode de Newton scalaire, l'idée est d'approcher g par sa partie linéaire au voisinage de x_0. En effet, par différentiabilité de g en x_0, on a

$$0 = g(x) = g(x_0) + \nabla g(x_0)(x - x_0) + \|x - x_0\|\varepsilon(x - x_0),$$

où la fonction ε dépend de x_0. Mais comme on ne sait rien sur cette dernière, mis à part le fait que $\varepsilon(h) \to 0$ quand $h \to 0$, on ne peut pas en faire grand-chose directement.

On se contente donc de chercher à résoudre l'équation d'inconnue x_1,

$$g(x_0) + \nabla g(x_0)(x_1 - x_0) = 0, \tag{5.2.1}$$

dans laquelle on a simplement enlevé le reste, et où ne subsiste que la partie linéaire de g, c'est-à-dire l'équivalent de l'équation de la tangente au graphe en dimension 1 (la tangente au graphe est remplacée par un sous-espace affine de $\mathbb{R}^{2n}$ de dimension n). On espère alors, comme en dimension $n = 1$, que x_1 va être une meilleure approximation de x que x_0.

L'équation (5.2.1) en l'inconnue x_1 est en fait un système linéaire de n équations à n inconnues, $\nabla g(x_0)x_1 = \nabla g(x_0)x_0 - g(x_0)$. Si la matrice $\nabla g(x_0)$ est inversible, alors on a une solution unique x_1 simplement donnée par

$$x_1 = x_0 - (\nabla g(x_0))^{-1} g(x_0),$$

et si jamais $x_1 \in U$ et $\nabla g(x_1)$ est encore inversible, alors on peut itérer le processus pour calculer x_2, et ainsi de suite. La méthode de Newton-Raphson consiste donc à construire si possible la suite

$$x_{k+1} = x_k - (\nabla g(x_k))^{-1} g(x_k). \tag{5.2.2}$$

Il faut ajouter « si possible », car rien ne garantit a priori que si $x_k \in U$ et $\nabla g(x_k)$ est inversible, alors $x_{k+1} \in U$ et $\nabla g(x_{k+1})$ est inversible. Il est même facile de construire des contre-exemples.

Remarquons qu'en dimension $n = 1$, multiplier $g(x_k) \in \mathbb{R}^n$ à gauche par l'inverse de la matrice $\nabla g(x_k)$ consiste exactement à diviser par $g'(x_k)$.

Commençons par quelques éléments utiles sur l'espace vectoriel des matrices $M_n(\mathbb{R})$, que l'on munit d'une norme (c'est un espace de dimension n^2) et sur le sous-ensemble des matrices inversibles $GL_n(\mathbb{R})$. On rappelle qu'un ouvert de $M_n(\mathbb{R})$ est une réunion de boules ouvertes, et que ceci ne dépend pas du choix de la norme, puisqu'il s'agit encore d'un espace vectoriel sur $\mathbb{R}$ de dimension finie.

On pourrait prendre n'importe quelle norme sur $M_n(\mathbb{R})$, par exemple $|||A||| = \max_{i,j} |a_{ij}|$, mais il est plus agréable pour travailler d'en choisir une qui soit adaptée à la norme que l'on a déjà adoptée sur $\mathbb{R}^n$. Pour cela, on utilise la notion de *norme matricielle subordonnée*. On renvoie à l'annexe A pour la définition A.2.1 et les propriétés de cette norme.

On rappelle aussi le résultat classique suivant.

4. Pour éviter de devoir introduire les différentielles d'ordre supérieur à 1, par exemple d'ordre 2. Si les différentielles d'ordre 1 ressortent de l'algèbre linéaire, celles d'ordre supérieur à 1 ressortent de l'algèbre multilinéaire. Ce n'est pas fondamentalement beaucoup plus dur, mais c'est définitivement plus lourd. Ici, on peut s'en tirer avec juste des dérivées partielles et cette notion de classe C^2 à base de dérivées partielles secondes coincide avec celle que l'on aurait si l'on avait la vraie différentielle seconde à notre disposition.

Lemme 5.2.2 *L'ensemble des matrices inversibles $GL_n(\mathbb{R})$ est un ouvert de $M_n(\mathbb{R})$.*

Nous donnons une preuve (constructive) de ce résultat à l'annexe A

Proposition 5.2.3 *Si g est de classe C^2 sur Uet $\nabla g(x)$ est inversible, alors il existe une boule centrée en x telle que pour toute donnée initiale dans cette boule, la suite des itérations de Newton est bien définie et converge quadratiquement vers x :*

$$\|x_{k+1} - x\| \leq K\|x_k - x\|^2,$$

pour un certain K.

Démonstration. Comme $\nabla g(x)$ est inversible, il existe une boule $B(\nabla g(x), r)$ dans l'espace des matrices qui ne contient que des matrices inversibles par le Lemme 5.2.2. Comme g est de classe C^1, il existe une boule $B(x, s)$ dans U cette fois telle que si $y \in B(x, s)$ alors $\nabla g(y)$ appartient à $B(\nabla g(x), r)$, et est donc inversible. On commence par se restreindre à la boule $B(x, s)$.

Dans cette boule, on définit une fonction $h\colon B(x, s) \to \mathbb{R}^n$ en posant

$$h(y) = y - (\nabla g(y))^{-1} g(y).$$

Par définition de la méthode de Newton, si x_k est bien défini, on a $x_{k+1} = h(x_k)$. La suite de Newton est donc celle des itérées de x_0 par h et il suffit par conséquent de montrer que h admet une boule invariante pour montrer que cette suite est bien définie pour tout k.

Pour tout $y \in B(x, s)$, on pose $z\colon [0, 1] \to \mathbb{R}^n$, $z(t) = g(y + t(x - y))$ de telle sorte que $z(0) = g(y)$ et $z(1) = g(x) = 0$. Par dérivation des fonctions composées, z est de classe C^2, avec

$$z'(t) = \nabla g(y + t(x - y))(x - y) \text{ et } z''(t) = \nabla^2 g(y + t(x - y))\big((x - y), (x - y)\big).$$

Attention, dans le contexte présent, ∇g est la matrice jacobienne dont les composantes sont indexées par deux indices, et $\nabla^2 g$ est une bête à trois indices [5] dont les composantes sont $\left(\nabla^2 g\right)_{ijk} = \frac{\partial^2 g_i}{\partial x_j \partial x_k}$. Quand on écrit tout en composantes, ceci donne

$$(\nabla g(a)b)_i = \sum_{j=1}^{n} \frac{\partial g_i}{\partial x_j}(a) b_j,$$

et

$$(\nabla^2 g(a)(b, c))_i = \sum_{j,k=1}^{n} \frac{\partial^2 g_i}{\partial x_j \partial x_k}(a) b_j c_k.$$

Posons

$$M = \sup_{y \in B(x,s)} \left(\max_i \sum_{j,k=1}^{n} \left| \frac{\partial^2 g_i}{\partial x_j \partial x_k}(y) \right| \right),$$

il vient alors

$$\sup_{y \in B(x,s)} \|\nabla^2 g(y)(b, c)\| \leq M\|b\|\|c\|.$$

5. On appelle ça un *tenseur*.

Revenant à la fonction z, l'inégalité de Taylor-Lagrange nous dit que

$$\|z(1) - z(0) - z'(0)\| \leq \frac{1}{2} \sup_{t\in[0,1]} \|z''(t)\|.$$

Quand on réexprime ceci en termes de g, on obtient

$$\begin{aligned}\|g(y) + \nabla g(y)(x-y)\| &\leq \frac{1}{2} \sup_{t\in[0,1]} \left\|\nabla^2 g(y + t(x-y))\big((x-y),(x-y)\big)\right\| \\ &\leq \frac{1}{2} M\|x-y\|^2,\end{aligned}$$

puisque le segment qui joint x à y est entièrement dans la boule. Comme

$$h(y) = y - (\nabla g(y))^{-1} g(y) = (\nabla g(y))^{-1}(\nabla g(y) y - g(y)),$$

on voit que

$$h(y) - x = -(\nabla g(y))^{-1}(\nabla g(y)(x-y) + g(y)),$$

d'où

$$\begin{aligned}\|h(y) - x\| &\leq |||(\nabla g(y))^{-1}||| \, \|\nabla g(y)(x-y) + g(y)\| \\ &\leq \frac{1}{2} M |||(\nabla g(y))^{-1}||| \, \|x-y\|^2 \\ &\leq K\|x-y\|^2,\end{aligned}$$

où l'on a posé $K = \frac{1}{2} M \sup_{y\in B(x,s)} |||(\nabla g(y))^{-1}|||$.

On pose alors $\alpha = \min\big(s, \frac{1}{K}\big)$. Si $y \in B(x,\alpha) \subset B(x,s)$, on a donc $\|y-x\| < \alpha$, d'où

$$\|h(y) - x\| \leq K\alpha^2 \leq \alpha,$$

puisque $\alpha \leq \frac{1}{K}$. On en déduit que $h(y) \in B(x,\alpha)$, c'est-à-dire que la boule $B(x,\alpha)$ est invariante par h, avec de plus le fait que ∇g est inversible en tout point de cette boule. La suite de Newton x_k est donc bien définie si x_0 est pris dans la boule $B(x,\alpha)$.

De plus, il vient immédiatement que, comme $x_{k+1} = h(x_k)$ pour tout k,

$$\|x_{k+1} - x\| \leq K\|x_k - x\|^2,$$

d'où la convergence quadratique de la méthode de Newton. ◊

Remarque 5.2.2 Il s'agit essentiellement de la même démonstration qu'en dimension 1. Les mêmes remarques concernant l'éventuelle divergence de la méthode si x_0 est trop loin de x s'appliquent. ◊

5.2.3 Quelques illustrations en dimension 2

Le cas particulier $n = 2$ est particulièrement intéressant quand on restreint l'attention aux fonctions holomorphes, c'est-à-dire quand on identifie $\mathbb{R}^2$ et $\mathbb{C}$, et que l'on considère les fonctions d'un ouvert de $\mathbb{C}$ à valeurs dans $\mathbb{C}$ qui sont dérivables au sens de $\mathbb{C}$. Cette condition est beaucoup plus restrictive qu'être juste différentiable de $\mathbb{R}^2$ dans $\mathbb{R}^2$. De façon très étonnante, sans aucune autre hypothèse que leur dérivabilité, même pas l'hypothèse de classe C^1, les fonctions holomorphes sont en fait automatiquement de classe C^∞ et même analytiques, c'est-à-dire localement égales à leur série de Taylor. De plus, leurs zéros sont forcément isolés. C'est par exemple le cas des fonctions polynomiales (ou de l'exponentielle complexe, mais celle-ci n'a pas beaucoup de zéros). La méthode de Newton va évidemment s'appliquer pour en approcher les zéros, puisqu'il s'agit d'un cas particulier de fonction de $\mathbb{R}^2$ dans $\mathbb{R}^2$.

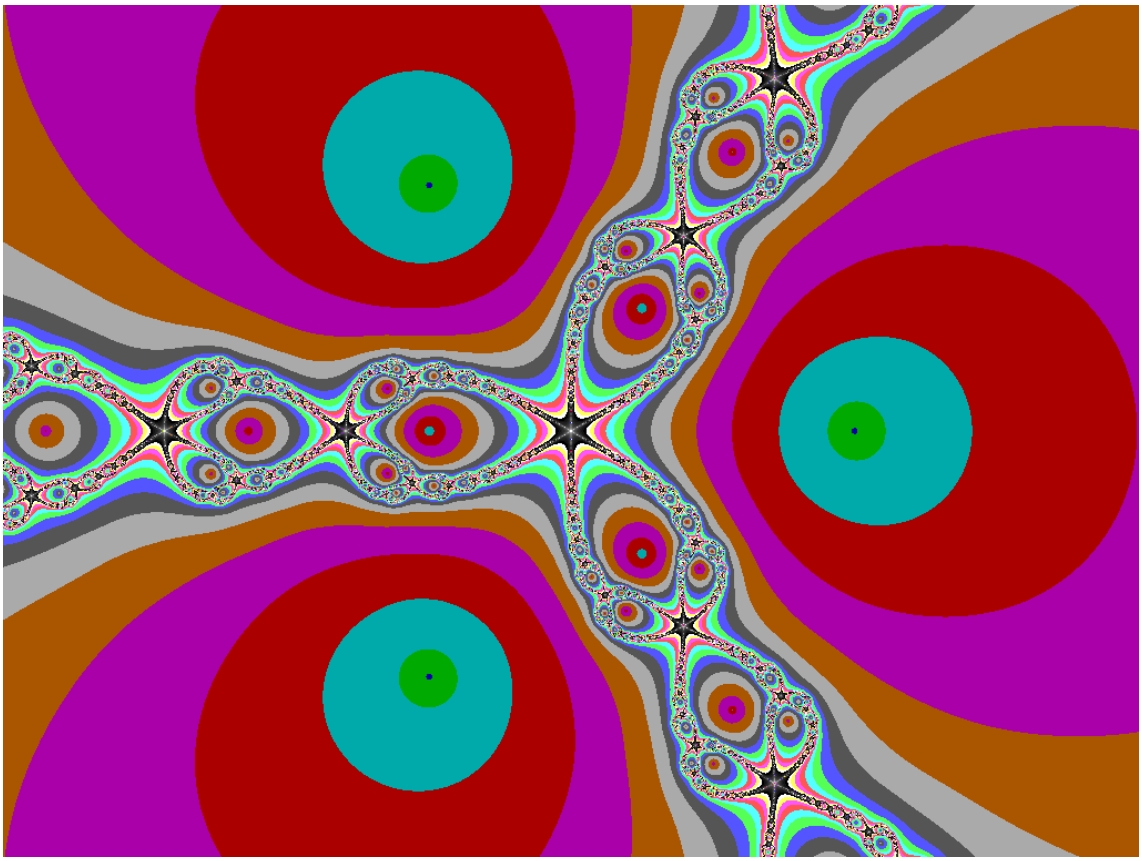
Comme l'action de la différentielle au point z, considérée comme application linéaire de $\mathbb{R}^2$ dans $\mathbb{R}^2$, sur un vecteur est une similitude, qui se traduit du point de vue complexe par la multiplication complexe par le nombre complexe $f'(z)$, l'inverse de la différentielle correspond à la division complexe par $g'(z)$ quand celui-ci n'est pas nul. On se retrouve donc avec la même formulation que dans le cas réel, $z_{k+1} = z_k - g(z_k)/g'(z_k)$, mais dans le plan complexe.

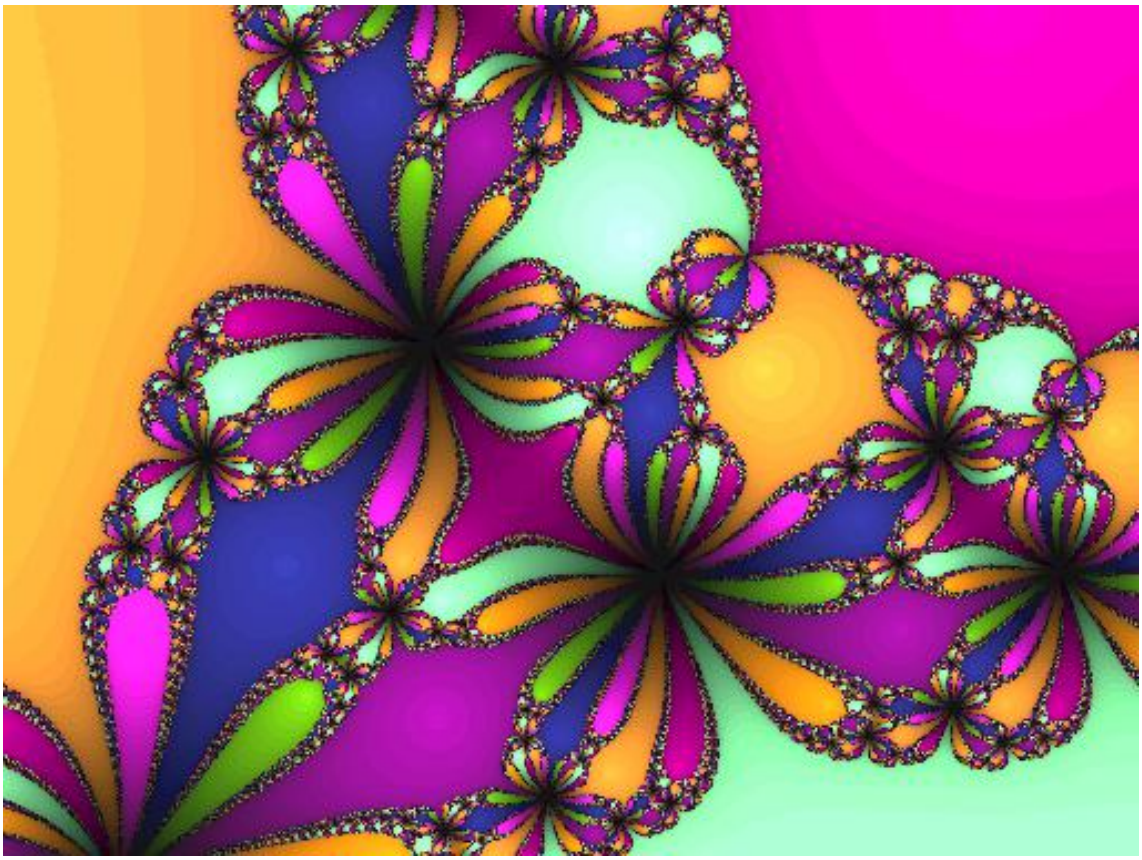
Par la théorie générale, on sait que la méthode de Newton va converger vers une racine quand le point initial z_0 est suffisamment proche de la racine en question. C'est un résultat local. Qu'en est-il plus globalement ? La suite peut converger ou pas et une question qui s'est posée un certain temps, dans le cas des polynômes, est quelle est l'influence du choix de z_0. En particulier, si la suite de Newton converge partant de z_0, vers quelle racine du polynôme converge-t-elle ?

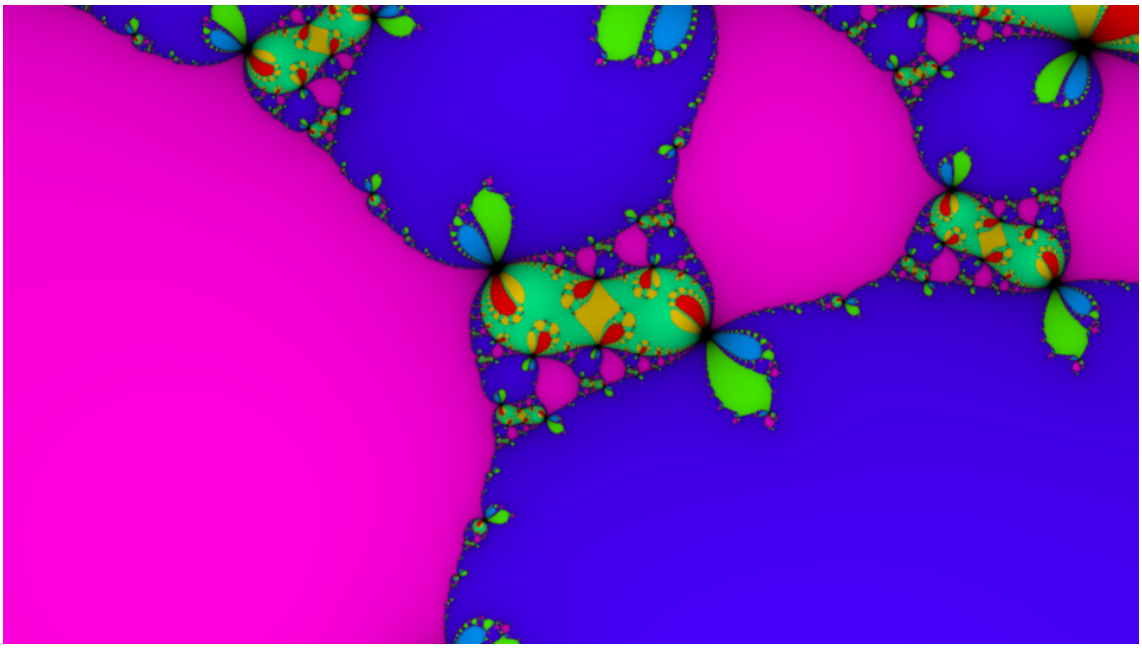
Cayley a montré au 19ème siècle que pour des polynômes de degré deux, cette suite converge vers la racine la plus proche du point de départ. Le plan complexe est donc coupé en deux par une droite, la médiatrice des deux racines (on suppose les racines distinctes), en dehors de laquelle on a convergence vers la racine incluse dans le même demi-plan ouvert. On vérifie par un calcul direct que la médiatrice est invariante par l'itération de Newton. Si le point initial est situé sur cette médiatrice, il s'ensuit que la suite de Newton ne converge pas.

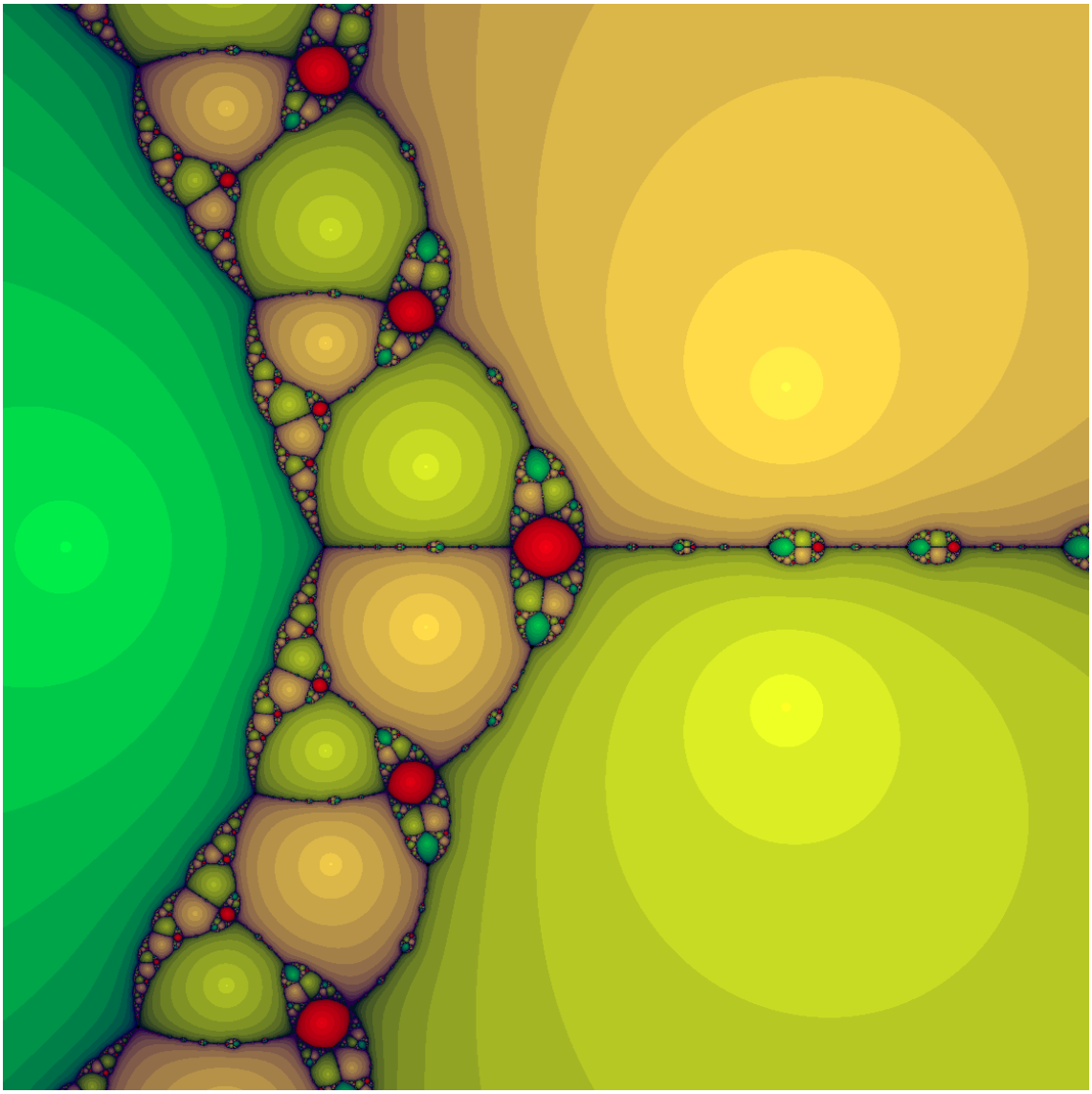
Les choses se complexifient (si l'on ose dire) considérablement à partir du troisième degré. Les bassins d'attraction de chaque racine, c'est-à-dire les régions du plan qui correspondent à une valeur initiale dont la suite des itérées converge vers cette racine, possèdent une structure compliquée. Leur frontière commune a une structure de type fractal. Par exemple, dans le cas d'un polynôme à trois racines distinctes, il peut se faire que chaque point de la frontière soit adhérent en même temps à trois bassins d'attraction différents, ce qui est un peu dur à visualiser. On en effet tendance à penser qu'une frontière ne sépare en général que deux pays, sauf que là, c'est trois pays qui se côtoient en chacun de ses points.

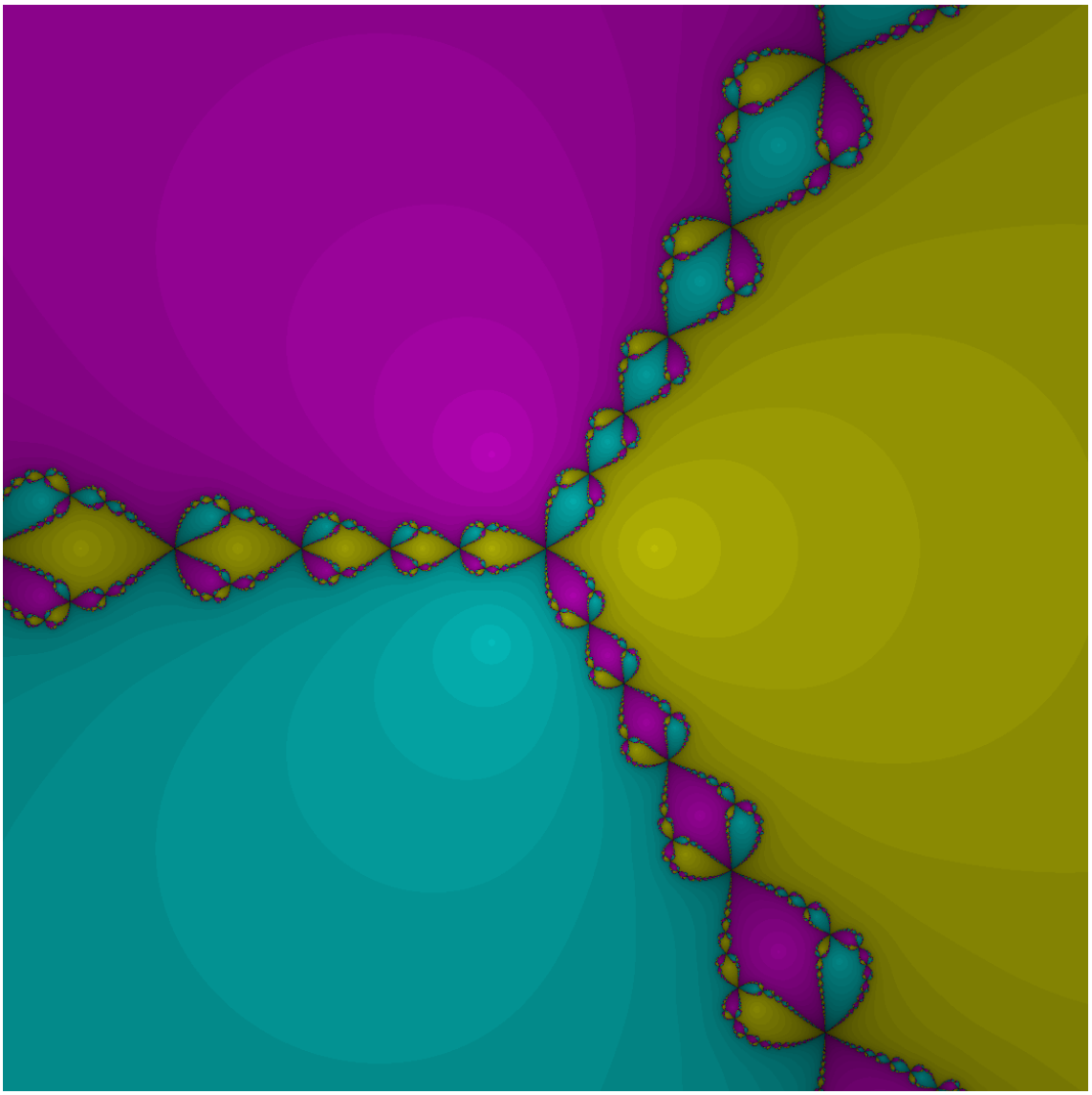
On trouve sur le web de nombreuses images de « fractales de Newton » qui illustrent ces propriétés de convergence ou de divergence d'ailleurs. On y représente souvent (mais pas toujours, comme dans le premier exemple qui suit) les bassins d'attraction d'une certaine couleur, et l'on peut identifier vers quelle racine on converge sur la base de cette couleur.











5.3 Pour en savoir plus : le théorème de Kantorovich

Un théorème évidemment dû à Kantorovich, à propos de la méthode de Newton. En fait, il y a plusieurs versions du théorème de Kantorovich, dont une assez optimale, mais horriblement technique. On va se contenter d'une version plus digeste (mais pas optimale).

Le point principal des théorèmes de type Kantorovich est non seulement de montrer la convergence de la méthode de Newton, mais par la même occasion, de donner des *conditions suffisantes d'existence de racine* que l'on a une petite chance de pouvoir vérifier en pratique ! Ce n'est pas rien. Énonçons le résultat qui va nous occuper dans la suite.

Théorème 5.3.1 (Kantorovich) *Soit U un ouvert de $\mathbb{R}^n$, $x_0 \in U$ et $g\colon U \to \mathbb{R}^n$ de classe C^1. On suppose que $\nabla g(x_0) \in GL_n(\mathbb{R})$ et qu'il existe un nombre $r > 0$ tel que $\overline{B(x_0, r)} \subset U$ et*

$$\left\| \left(\nabla g(x_0)\right)^{-1} g(x_0) \right\| \leq \frac{r}{2}, \tag{5.3.1}$$

$$\forall y, z \in B(x_0, r), \left\| \left(\nabla g(x_0)\right)^{-1} \left(\nabla g(y) - \nabla g(z)\right) \right\| \leq \frac{1}{r} \|y - z\|. \tag{5.3.2}$$

Alors pour tout $y \in B(x_0, r)$, on a $\nabla g(y) \in GL_n(\mathbb{R})$, les itérations de Newton x_k partant de x_0 sont bien définies pour tout k, restent dans $B(x_0, r)$ et convergent vers $x \in \overline{B(x_0, r)}$, qui est de plus l'unique racine de f dans $\overline{B(x_0, r)}$.

On a enfin l'estimation d'erreur

$$\|x_k - x\| \leq \frac{r}{2^k}.$$

Démonstration. On va se simplifier la vie en posant le changement de variable suivant. Si $u \in B(0,1)$ alors $x = x_0 + ru \in B(x_0, r)$ et réciproquement $u = \frac{x-x_0}{r} \in B(0,1)$. On introduit donc $h\colon B(0,1) \to \mathbb{R}^n$ par

$$h(u) = \frac{1}{r}\big(\nabla g(x_0)\big)^{-1} g(x_0 + ru),$$

ou de façon équivalente

$$g(x) = r\nabla g(x_0) h\Big(\frac{x - x_0}{r}\Big).$$

En ce qui concerne les matrices jacobiennes, on obtient

$$\nabla h(u) = \big(\nabla g(x_0)\big)^{-1} \nabla g(x_0 + ru) \text{ et } \nabla g(x) = \nabla g(x_0) \nabla h\Big(\frac{x - x_0}{r}\Big),$$

par dérivation des fonctions composées. En particulier, $h(0) = \frac{1}{r}\big(\nabla g(x_0)\big)^{-1} g(x_0)$ et $\nabla h(0) = I$.

Quand on les réécrit en termes du changement de variables u et h, les hypothèses (5.3.1) et (5.3.2) deviennent plus agréables

$$\|h(0)\| \leq \frac{1}{2}, \tag{5.3.3}$$

$$\forall u, v \in B(0,1), |||\nabla h(u) - \nabla h(v)||| \leq \|u - v\|. \tag{5.3.4}$$

On remarque de plus que les itérations de Newton de g et celles de h se correspondent via le changement de variable. En effet, si $x_k = x_0 + ru_k$ est l'itération de Newton de f, alors

$$\begin{aligned} u_{k+1} = \frac{x_{k+1} - x_0}{r} &= \frac{x_k - x_0}{r} - \frac{\nabla g(x_k)^{-1} g(x_k)}{r} \\ &= u_k - \frac{\nabla h(u_k)^{-1} \nabla g(x_0)^{-1} g(x_k)}{r} = u_k - \nabla h(u_k)^{-1} h(u_k). \end{aligned}$$

Enfin les racines de g et de h dans leurs boules respectives se correspondent manifestement par le changement de variable. On va donc travailler sur h satisfaisant (5.3.3) et (5.3.4) dans la boule unité, et sur ses itérations de Newton, ça sera beaucoup plus confortable.

On prend d'abord $v = 0$ dans la deuxième inégalité (5.3.4). Il vient donc

$$\forall u \in B(0,1), |||\nabla h(u) - I||| \leq \|u\| < 1.$$

On en déduit que $\nabla h(u)$ est inversible partout dans la boule ouverte, avec l'estimation

$$|||\nabla h(u)^{-1}||| \leq \frac{1}{1 - \|u\|} \tag{5.3.5}$$

qui découle immédiatement de la preuve du Lemme 5.2.2. De plus, on a

$$h(u) - h(v) = \int_0^1 \nabla h(v + t(u - v))(u - v)\, dt,$$

si bien que

$$\begin{aligned} \big\|h(u) - h(v) - \nabla h(v)(u - v)\big\| &= \Big\| \int_0^1 \big(\nabla h(v + t(u - v)) - \nabla h(v)\big)(u - v)\, dt \Big\| \\ &\leq \int_0^1 \big\| \big(\nabla h(v + t(u - v)) - \nabla h(v)\big)(u - v) \big\|\, dt \\ &\leq \int_0^1 |||\nabla h(v + t(u - v)) - \nabla h(v)|||\, \|u - v\|\, dt \\ &\leq \|u - v\|^2 \int_0^1 t\, dt = \frac{\|u - v\|^2}{2}, \end{aligned} \tag{5.3.6}$$

toujours par l'hypothèse (5.3.4).

Regardons ce que l'on peut dire de la première itération de Newton avec $u_0 = 0$, $u_1 = 0 - Ih(0) = -h(0)$. Bien sûr,

$$\|u_1\| = \|h(0)\| \leq \frac{1}{2}.$$

Par (5.3.5), on a aussi

$$|||\nabla h(u_1)^{-1}||| \leq \frac{1}{1 - \|u_1\|} \leq 2.$$

Enfin, comme $h(u_1) = h(u_1) + u_1 - u_1 + u_0 = h(u_1) - h(u_0) - \nabla h(u_0)(u_1 - u_0)$, on déduit de (5.3.6) que

$$\|h(u_1)\| \leq \frac{\|u_1\|^2}{2} \leq \frac{1}{8} = \frac{1}{2^3}.$$

Ceci incite fortement à poser l'hypothèse de récurrence suivante pour $k \geq 1$,

$$\|u_k - u_{k-1}\| \leq \frac{1}{2^k}, \|u_k\| \leq 1 - \frac{1}{2^k}, |||\nabla h(u_k)^{-1}||| \leq 2^k \text{ et } \|h(u_k)\| \leq \frac{1}{2^{2k+1}}. \tag{5.3.7}$$

Comme $u_k \in B(0,1)$ par la deuxième inégalité, $u_{k+1} = u_k - \nabla h(u_k)^{-1} h(u_k)$ est bien défini. De plus, $h(u_{k+1}) = h(u_{k+1}) - h(u_k) - \nabla h(u_k)(u_{k+1} - u_k)$. Il vient donc

$$\|u_{k+1} - u_k\| = \|\nabla h(u_k)^{-1} h(u_k)\| \leq |||\nabla h(u_k)^{-1}||| \, \|h(u_k)\| \leq \frac{2^k}{2^{2k+1}} = \frac{1}{2^{k+1}}.$$

Par l'inégalité triangulaire, il s'ensuit que

$$\|u_{k+1}\| \leq \|u_{k+1} - u_k\| + \|u_k\| \leq \frac{1}{2^{k+1}} + 1 - \frac{1}{2^k} = 1 - \frac{1}{2^{k+1}}.$$

L'estimation (5.3.5) donne alors

$$|||\nabla h(u_{k+1})^{-1}||| \leq \frac{1}{1 - \|u_{k+1}\|} \leq 2^{k+1}.$$

Enfin,

$$\|h(u_{k+1})\| = \|h(u_{k+1}) - h(u_k) - \nabla h(u_k)(u_{k+1} - u_k)\| \leq \frac{\|u_{k+1} - u_k\|^2}{2} \leq \frac{1}{2^{2k+3}}.$$

On a ainsi montré que l'itération de Newton est bien définie pour tout k avec les estimations (5.3.7).

L'estimation $\|u_k - u_{k-1}\| \leq \frac{1}{2^k}$ montre comme on l'a vu déjà bien souvent que la suite u_k est de Cauchy, elle converge donc vers un u dans l'espace complet $\overline{B(0,1)}$, avec la vitesse de convergence linéaire donnée par l'estimation d'erreur du théorème. Comme $h(u_k) \to 0$ par la dernière estimation de (5.3.7), on voit que $h(u) = 0$ par continuité.

Il reste à voir qu'il n'y a pas d'autre racine que u dans la boule fermée. Soit v une autre telle racine. On a manifestement $\|u_0 - v\| = \|v\| \leq 1 = \frac{1}{2^0}$. Faisons l'hypothèse de récurrence que $\|u_k - v\| \leq \frac{1}{2^k}$. Il vient

$$\begin{aligned} u_{k+1} - v = u_k - v - \nabla h(u_k)^{-1} h(u_k) &= u_k - v - \nabla h(u_k)^{-1}(h(u_k) - h(v)) \\ &= \nabla h(u_k)^{-1}\big(\nabla h(u_k)(u_k - v) - h(u_k) + h(v)\big), \end{aligned}$$

d'où, en utilisant à nouveau (5.3.6),

$$\|u_{k+1} - v\| \leq |||\nabla h(u_k)^{-1}||| \, \big\|\nabla h(u_k)(u_k - v) - h(u_k) + h(v)\big\| \leq 2^k \frac{\|u_k - v\|^2}{2} \leq \frac{1}{2^{k+1}}.$$

On a ainsi montré que $u_k \to v$, ce qui implique évidemment que $v = u$. ◊

Remarque 5.3.1 i) La grosse différence avec la proposition 5.2.3 est que l'on ne suppose pas l'existence d'une racine, avec convergence dans une boule que l'on ne connaît pas, mais que l'on montre, si un certain nombre

de conditions sont satisfaites, l'existence d'une telle racine avec convergence dans une boule que l'on connaît. Bien sûr, encore faut-il trouver des x_0 et r qui marchent.

ii) Regardons ce que dit le théorème de Kantorovich en dimension 1. Le changement de variable s'écrit $h(u) = g(x_0 + ru)/(rg'(x_0))$, avec h définie sur $[-1, 1]$ et $h'(0) = 1$. Les hypothèses du théorème s'écrivent

$$|h(0)| \leq \frac{1}{2},$$
$$|h'(u) - h'(v)| \leq |u - v|.$$

Il est assez clair en intégrant les inégalités $1 - u \leq h'(u) \leq 1 + u$ que l'on a alors une seule racine dans $[-1, 1]$, qui est négative si $h(0) \geq 0$ et positive si $h(0) \leq 0$. Dans le cas où h est deux fois dérivable, la condition de Lipschitz sur h' est équivalente au fait que $|h''(u)| \leq 1$ pour tout u, une borne sur la courbure du graphe.

iii) Le théorème est vrai pour n'importe quelle norme sur $\mathbb{R}^n$ en utilisant sa norme matricielle subordonnée. Il reste aussi vrai avec la même démonstration dans un espace de Banach quelconque, y compris de dimension infinie, en utilisant la norme d'application linéaire, qui se définit exactement comme une norme matricielle subordonnée.

iv) Le théorème ne fournit qu'une convergence linéaire, et non pas quadratique, pour la méthode de Newton. La raison est que l'on ne fait pas l'hypothèse que g est de classe C^2, mais seulement C^1 avec la matrice jacobienne lipschitzienne (on dit que g est de classe $C^{1,1}$). De plus, la racine trouvée x est dans la boule fermée, et si elle est sur la sphère, rien n'interdit que $\nabla g(x)$ soit non inversible, voir Figure 5.3. On ne peut donc pas espérer plus qu'une convergence linéaire sous ces seules hypothèses. Néanmoins, si g est de classe C^2 et si la racine est dans la boule ouverte, alors on finit tôt ou tard par tomber dans la boule inconnue de la proposition 5.2.3, et la convergence est effectivement quadratique.

v) La version horrible du théorème de Kantorovich donne des conditions vérifiables en pratique et qui assurent la convergence quadratique. En fait, cette version ne suppose pas g de classe C^2 non plus, mais seulement $C^{1,1}$, avec tout un tas d'autres conditions. ◊

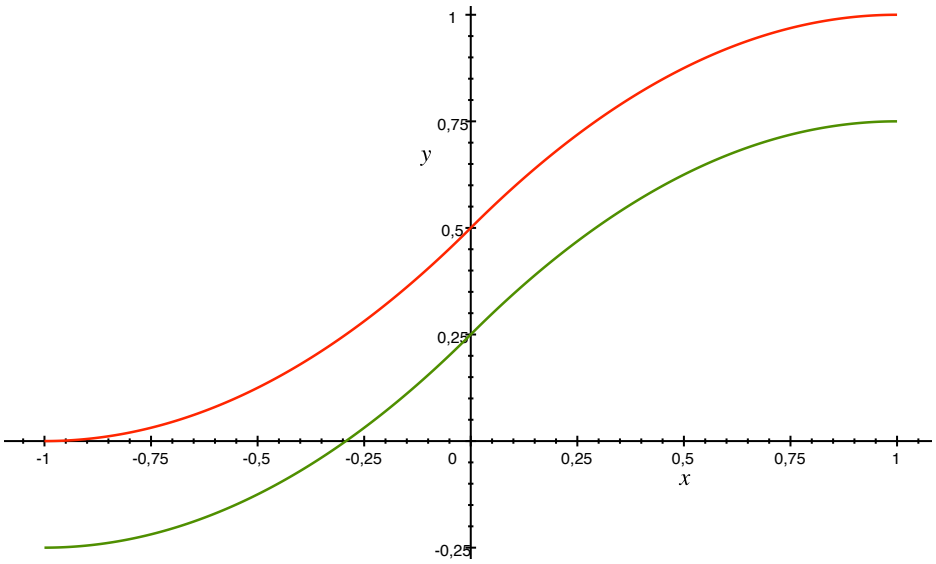


FIGURE 5.3 – Newton-Kantorovich : en vert, convergence quadratique (en partant de 0), en rouge, convergence linéaire.

5.4 Retour à la méthode d'Euler implicite

On s'intéresse à l'approximation de la solution d'un problème de Cauchy pour une équation différentielle ordinaire

$$\forall t \in [0,T], y'(t) = g(t,y(t)), \quad y(0) = y_0,$$

où $g\colon [0,T]\times\mathbb{R}^d \to \mathbb{R}^d$ et $y_0 \in \mathbb{R}^d$ sont donnés et l'on cherche la fonction inconnue $y\colon [0,T]\to\mathbb{R}^d$. Si g est continue d'une part et Lipschitzienne par rapport à y, uniformément par rapport à t d'autre part, c'est-à-dire qu'il existe une constante L telle que

$$\forall t \in [0,T], \forall y,z \in \mathbb{R}^d, \|g(t,y)-g(t,z)\| \leq L\|y-z\|,$$

alors ce problème de Cauchy admet une solution et une seule pour tout $y_0 \in \mathbb{R}^d$. C'est le théorème de Cauchy-Lipschitz global. Évidemment, il s'agit d'une question d'intérêt général majeur.

En général, il n'existe par contre aucune formule qui donne cette solution, et l'on doit l'approcher pour pouvoir en dire quelque chose de quantitatif. Pour cela, on se donne un entier N et l'on introduit un pas de temps $h = \frac{T}{N+1}$, puis on pose $t_n = nh$, $n = 0,1,\ldots,N+1$ qui sont des instants de discrétisation de l'intervalle de temps $[0,T]$.

La méthode d'Euler implicite consiste à construire la suite $y_n \in \mathbb{R}^d$ définie par récurrence par

$$y_0 = y_0, \frac{y_{n+1}-y_n}{h} = g(t_{n+1},y_{n+1}) \text{ pour } 0 \leq n \leq N.$$

On démontre que cette méthode marche et que y_n est une approximation de $y(t_n)$ d'autant meilleure que h est petit, en un sens précis que l'on ne détaillera pas ici.

Comment calculer y_{n+1} connaissant y_n pour faire progresser la récurrence ? Sauf cas particulier où g est très simple, on doit de fait trouver une racine de la fonction

$$f_n(x) = x - y_n - hg(t_{n+1},x),$$

d'où le qualificatif « implicite ».

On peut procéder par itérations de point fixe, on peut également appliquer la méthode de Newton. Voyons ce que le théorème de Kantorovich a à nous dire à ce sujet. On a

$$\nabla f_n(x) = I - h\nabla g(t_{n+1},x),$$

où la matrice jacobienne de g est prise par rapport à x, à $t = t_{n+1}$ fixé. C'est donc bien une matrice $d\times d$. On va supposer que $\nabla g(t_{n+1},\cdot)$ est lipschitzienne de constante M pour la norme matricielle subordonnée, c'est-à-dire g de classe $C^{1,1}$.

On va évidemment démarrer l'itération de Newton à $x_0 = y_n$ (on est censé approcher y_{n+1} qui est censé approcher $y(t_{n+1})$ qui est proche de $y(t_n)$ qui est lui-même approché censément par y_n, donc c'est raisonnable). On a donc

$$f_n(x_0) = -hg(t_{n+1},y_n) \text{ et } \nabla f_n(x_0)^{-1} f_n(x_0) = -h\big(I - h\nabla g(t_{n+1},y_n)\big)^{-1} g(t_{n+1},y_n)$$

L'hypothèse de L-lispchitzianité de g se traduit bien sûr par

$$|||\nabla g(t_{n+1},x)||| \leq L,$$

pour tout x et le caractère Lipschitzien de ∇g par

$$|||\nabla g(t_{n+1},x) - \nabla g(t_{n+1},y)||| \leq M\|x-y\|,$$

pour une autre constante de Lipschitz M.

On déduit de la première inégalité que $\nabla f_n(x)$ est inversible dès que $h < \frac{1}{L}$ avec

$$|||\nabla f_n(x)^{-1}||| = \Big|\Big|\Big|\big(I - h\nabla g(t_{n+1},x)\big)^{-1}\Big|\Big| \leq \frac{1}{1-hL}.$$

Il vient alors

$$\begin{aligned}\|\nabla f_n(x_0)^{-1} f_n(x_0\| &= h\Big\|\big(I - h\nabla g(t_{n+1},y_n)\big)^{-1} g(t_{n+1},y_n)\Big\| \\ &\leq \frac{hG}{1-hL},\end{aligned}$$

où l'on a posé $G = \|g(t_{n+1}, y_n)\|$ (ou un majorant indépendant de n). Pour satisfaire la première condition du théorème de Kantorovich, il suffit donc que

$$\frac{hG}{1-hL} \leq \frac{r}{2}. \tag{5.4.1}$$

De même,

$$\left|\left|\left|\left(\nabla f_n(x_0)\right)^{-1}\left(\nabla f_n(y) - \nabla f_n(z)\right)\right|\right|\right| \leq \frac{hM}{1-hL}\|y-z\|,$$

et pour satisfaire la deuxième condition du théorème de Kantorovich, il suffit donc que

$$\frac{hM}{1-hL} \leq \frac{1}{r}. \tag{5.4.2}$$

Il existe un tel r si et seulement si

$$\frac{2hG}{1-hL} \leq \frac{1-hL}{hM}.$$

Ceci a lieu si et seulement si $0 \leq h \leq \min\left(h_+, \frac{1}{L}\right)$ où h_+ est la plus petite valeur positive de l'expression $\frac{L \pm \sqrt{2MG}}{L^2 - 2MG}$ (on peut toujours supposer que $L^2 - 2MG \neq 0$). Si l'inégalité de droite est stricte, il y a alors tout un intervalle de valeurs r possibles, ce qui implique que la convergence est quadratique, soit si g est C^2, soit si l'on en croit la version horrible de Kantorovich.

Bien sûr, il s'agit de la convergence d'une itération de Newton, que l'on doit refaire avec une fonction différente pour chaque valeur de n correspondant à la discrétisation temporelle de l'EDO. En fonction de la difficulté des calculs, les considérations de coût peuvent alors devenir importantes.

5.5 Considérations de mise en œuvre pratique

Supposons que nous ayons un vrai problème $f(x) = 0$ de la vraie vie à résoudre par la méthode de Newton. On s'est donc donné un $x_0 \in \mathbb{R}^n$ (en croisant les doigts pour qu'il ne soit pas trop loin de la racine cherchée) et l'on doit calculer un nombre fini de termes de l'itération de Newton

$$x_{k+1} = x_k - \left(\nabla f(x_k)\right)^{-1} f(x_k).$$

En pratique, « calculer » va bien sûr signifier approcher les valeurs de l'itération en question. Cette approximation est une source d'erreurs supplémentaires qui vont se propager à chaque étape de l'itération. On laisse de côté ici ces erreurs en les supposant négligeables [6]. On va donc devoir à chaque itération, à partir du $x_k \in \mathbb{R}^n$ obtenu à la fin de l'itération précédente

1. évaluer $f(x_k) \in \mathbb{R}^n$,
2. évaluer la matrice $\nabla f(x_k) \in M_n(\mathbb{R})$,
3. évaluer le produit matrice-vecteur $\nabla f(x_k) x_k$,
4. évaluer la différence $\nabla f(x_k) x_k - f(x_k)$,
5. résoudre le système linéaire $\nabla f(x_k) x_{k+1} = \nabla f(x_k) x_k - f(x_k)$.

Chacune de ces opérations a un coût, c'est-à-dire va demander plus ou moins d'opérations élémentaires (additions, multiplications, divisions, stockage en mémoire), et donc prendre plus ou moins de temps sur un ordinateur donné. Pour de petites valeurs de n et des fonctions f simples, ce temps peut être tellement bref qu'il semble que le résultat tombe

6. Que ceci soit justifié ou pas...

instantanément. Néanmoins, n n'est pas forcément petit, f n'est pas forcément simple [7], et surtout il se peut que le calcul de racine soit en fait inclus dans une ou plusieurs autres boucles itératives, et que l'on doive donc le répéter un très grand nombre de fois sur des valeurs différentes.

Dans ces conditions, la question du temps de calcul peut devenir primordiale [8].

Tentons d'évaluer à la louche le nombre d'opérations nécessaires pour chacune des étapes d'une itération.

1. On a n variables scalaires à combiner entre elles pour évaluer chaque composante de f. En général, cela va imposer d'effectuer au moins n opérations élémentaires sur des nombres (probablement beaucoup plus) pour chaque composante, d'où un coût global d'au moins $O(n^2)$ opérations.
2. Même chose, mais avec n^2 composantes, d'où un coût global d'au moins $O(n^3)$ opérations.
3. Là, on sait exactement ce que cela coûte : n^2 multiplications et $n(n-1)$ additions, en l'absence de structure particulière de la matrice $\nabla f(x_k)$. D'où un coût global de $O(n^2)$ opérations.
4. L'étape la plus économique : n additions.
5. Encore une fois, en l'absence de structure particulière de la matrice $\nabla f(x_k)$, on n'a guère d'autre recours que d'utiliser la méthode de Gauss, ou une de ses déclinaisons plus modernes et sophistiquées, soit $O(n^3)$ opérations. [9]

Les étapes 3 et 4 présentent un coût négligeable par rapport aux étapes 2 et 5. Si f est « simple », il en va de même de l'étape 1. C'est sur les étapes 2 et 5, et en fait surtout 2, qu'il faut porter les efforts d'optimisation si le besoin s'en fait sentir. Cette observation conduit à l'idée des *méthodes de quasi-Newton*.

5.6 Les méthodes de quasi-Newton

Il s'agit simplement de remplacer la matrice jacobienne $\nabla f(x_k)$ par une autre matrice A_k inversible, selon une stratégie à préciser. On se donnera donc x_0, puis on itérera

$$x_{k+1} = x_k - A_k^{-1} f(x_k). \tag{5.6.1}$$

Pour que ceci soit intéressant, il est préférable que les matrices A_k soient « faciles » à calculer et conduisent à des systèmes linéaires « faciles » à résoudre, sans pour autant détruire complètement les propriétés de convergence de la méthode.

Un exemple extrêmement simple, en dimension 1, consiste à prendre $A_k = f'(x_0) \neq 0$ pour tout k, que l'on ne calcule donc qu'une seule fois. Bien sûr, on est alors en train d'itérer la fonction $g(x) = x - f(x)/f'(x_0)$ dont tout point fixe est clairement racine de f. De fait, cette méthode converge si x_0 est suffisamment proche de x, manifestement d'autant plus vite que x_0 est proche de x.

7. Par exemple, il se peut que l'on n'y ait accès qu'à travers un programme informatique lui-même complexe.

8. Être capable de prédire le temps qu'il fera demain après un calcul qui prend un mois ne présente qu'un intérêt météorologique limité. Ceci dit, j'ignore si Météo France utilise la méthode de Newton ou non pour alimenter les bulletins météo des chaînes de télévision.

9. Peut-être un poil moins pour certaines méthodes.

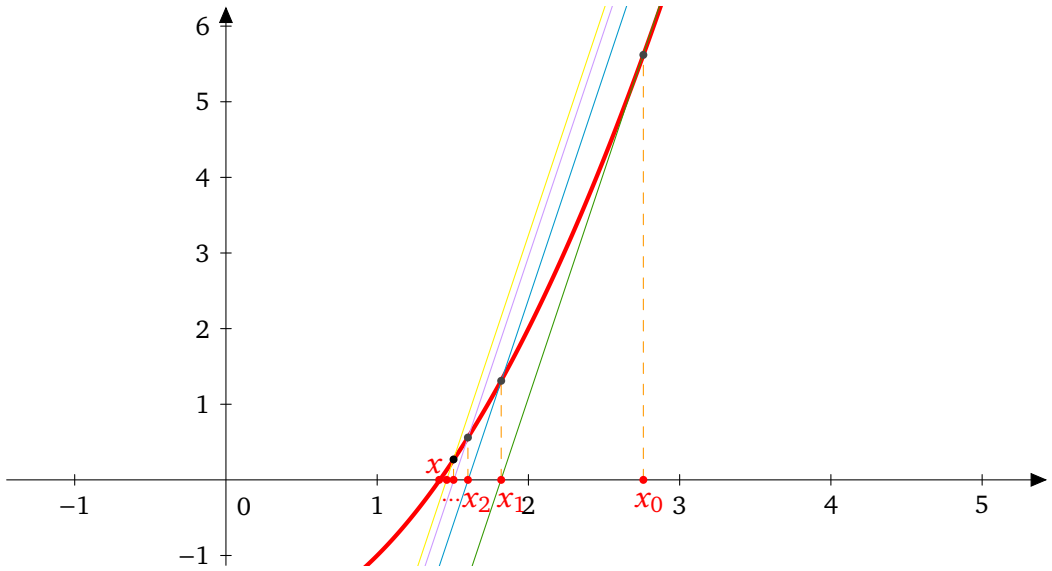


FIGURE 5.4 – Une méthode de quasi-Newton simpliste.

Dans cet exemple, on obtient semble-t-il une convergence linéaire, mais qui peut en pratique se révéler assez satisfaisante. On reviendra plus loin sur d'autres stratégies plus intelligentes. Montrons d'abord un résultat de convergence pour ces méthodes, en recopiant plus ou moins ce que l'on a déjà fait pour Newton.

Proposition 5.6.1 *Soit f est de classe C^1 sur U, $x \in U$ une racine de f et A_k une suite de matrices de $GL_n(\mathbb{R})$ telle que la suite A_k^{-1} soit bornée. Soit M un majorant de $\{|||A_k^{-1}|||; k \in \mathbb{N}\}$. On suppose qu'il existe $r > 0$ et $0 \leq \beta < 1$ tels que pour tout $y \in B(x,r) \cap U$,*

$$|||\nabla f(y) - A_k||| \leq \frac{\beta}{M}. \tag{5.6.2}$$

Alors il existe une boule centrée en x telle que pour toute donnée initiale dans cette boule, la suite des itérations de quasi-Newton est bien définie et converge au moins linéairement vers x, qui est de plus l'unique racine dans cette boule.

Démonstration. On pose $g_k(y) = y - A_k^{-1} f(y)$ et $h_k(y) = f(y) - A_k(y - x)$, qui sont définies sur U. En particulier, $x_{k+1} = g_k(x_k)$. Soit $B(x,r)$ la boule sur laquelle l'estimation (5.6.2) est supposée avoir lieu. En diminuant éventuellement r, on peut supposer que son adhérence, c'est-à-dire la boule fermée, est incluse dans U. Pour montrer que la suite est alors bien définie pour tout $x_0 \in B(x,r)$, il suffit de montrer que $g_k(B(x,r)) \subset B(x,r)$ pour tout k.

Comme $h_k(x) = 0$, on peut écrire pour tout $y \in B(x,r)$,

$$\begin{aligned} g_k(y) - x &= y - x - A_k^{-1} f(y) \\ &= A_k^{-1}\big(A_k(y-x) - f(y)\big) \\ &= A_k^{-1}\big(h_k(y) - h_k(x)\big). \end{aligned}$$

On a $\nabla h_k(y) = \nabla f(y) - A_k$. Par conséquent, par l'inégalité des accroissements finis,

$$\begin{aligned}\|g_k(y) - x\| &\leq |||A_k^{-1}||| \, \|h_k(y) - h_k(x)\| \\ &\leq |||A_k^{-1}||| \sup_{B(x,r)} |||\nabla h_k(y)||| \, \|y - x\| \\ &\leq \beta \|y - x\| < r,\end{aligned}$$

et cela pour tout k. On a donc bien $g_k(y) \in B(x, r)$ et la suite est bien définie.

Par un calcul très voisin, on note qu'alors

$$\begin{aligned}f(x_k) &= f(x_k) - f(x_{k-1}) - A_{k-1}(x_k - x_{k-1}) \\ &= h_{k-1}(x_k) - h_{k-1}(x_{k-1}),\end{aligned}$$

pour tout $k \geq 1$. Il vient donc

$$\|f(x_k)\| \leq \frac{\beta}{M} \|x_k - x_{k-1}\|. \tag{5.6.3}$$

Comme par ailleurs,

$$x_{k+1} - x_k = -A_k^{-1} f(x_k)$$

on en déduit que

$$\|x_{k+1} - x\| \leq |||A_k^{-1}||| \, \|f(x_k)\| \leq \beta \|x_k - x_{k-1}\|$$

pour tout $k \geq 1$. Il s'ensuit, comme dans la preuve du théorème de point fixe de Banach, que la suite x_k est de Cauchy. Elle converge donc vers un certain $\bar{x}$ dans l'espace complet $\overline{B(x, r)}$, et la convergence est linéaire, encore une fois comme dans la preuve du théorème de point fixe de Banach. Par l'estimation (5.6.3) et en utilisant le fait que f est continue, on en déduit que $f(\bar{x}) = 0$.

Pour conclure, on remarque que

$$x - \bar{x} = -A_0^{-1}\big(f(x) - f(\bar{x}) - A_0(x - \bar{x})\big) = -A_0^{-1}\big(h_0(x) - h_0(\bar{x})\big),$$

d'où à nouveau

$$\|x - \bar{x}\| \leq \beta \|x - \bar{x}\|,$$

ce qui implique que $\bar{x} = x$ puisque $\beta < 1$. ◊

Remarque 5.6.1 i) La méthode de Newton consiste à prendre $A_k = \nabla f(x_k)$. L'estimation (5.6.2) est alors assurée sur une boule assez petite par continuité uniforme de ∇f sur toute boule fermée. On établit donc ici la convergence de la méthode de Newton pour f de classe C^1, mais pas nécessairement C^2. Par contre, dans ce cas en l'absence de régularité au delà de C^1, on ne peut guère espérer récupérer la convergence quadratique.

ii) Cette condition dit plus généralement que, d'une certaine façon, A_k ne doit pas être trop éloigné des valeurs prises par ∇f dans la boule.

iii) Il existe une version plus quantitative du résultat, plus dans l'esprit du théorème de Kantorovich, théorème que l'on devrait avoir le temps de voir plus tard.

iv) Les méthodes de quasi-Newton sont très utilisées dans le contexte de l'optimisation. Soit F une fonction de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}$ que l'on cherche à maximiser ou minimiser. Si F est de classe C^1, une condition nécessaire pour avoir un point x de minimum ou de

maximum local est que $\nabla F(x) = 0$. Une stratégie d'optimisation peut donc consister à chercher les racines de la fonction $f(x) = \nabla F(x)$ qui est bien une fonction de $\mathbb{R}^n$ valeurs dans $\mathbb{R}^n$. Si F est de classe C^2, alors on pourra penser utiliser la méthode de Newton (dont la convergence quadratique est assurée localement si F est de classe C^3) ou bien des méthodes de quasi-Newton.

Dans ce cas, $\nabla f = \nabla^2 F$ est la *hessienne* de F, qui peut être très difficile à calculer pour des F compliquées, et conduire à des systèmes linéaires difficiles à résoudre. Les méthodes de quasi-Newton prennent tout leur intérêt ici, d'où leur popularité en optimisation. En effet, on dispose de très nombreuses façons de se donner directement les matrices $B_k = A_k^{-1}$ à partir des quantités antérieurement calculées, ce qui élimine totalement l'étape de résolution de système linéaire, qui est alors remplacée par un produit matrice-vecteur. Il arrive souvent que la perte de vitesse de convergence par rapport à la méthode de Newton soit plus que compensée par la simplification des autres calculs, ce qui rend *in fine* les méthodes de quasi-Newton plus performantes dans ce contexte que la méthode de Newton.

v) Pour revenir au contexte général de l'équation $f(x) = 0$, une stratégie possible consiste à prendre $A_{k+p} = \nabla f(x_k)$ pour $1 \leq p \leq m$, puis de passer à la matrice jacobienne suivante $A_{k+m+p} = \nabla f(x_{k+m})$ et ainsi de suite. L'intérêt de cette version de quasi-Newton est de quand même adapter la matrice de l'itération à la matrice jacobienne, mais seulement de temps à autre. Entre les deux, on ne recalcule pas la matrice, et on peut de plus en utiliser une factorisation de type LU pour que toutes les résolutions de systèmes linéaires intermédiaires se fassent en seulement $O(n^2)$ opérations. En choisissant bien m, voire en le faisant varier, on peut aboutir à des compromis performance/coût intéressants.

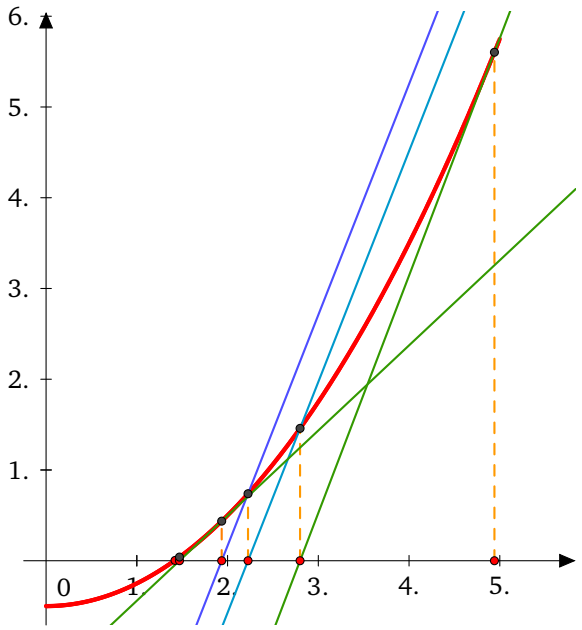


FIGURE 5.5 – Illustration de la dernière méthode de quasi-Newton.

Chapitre 6

Schémas d'ordre élevé

On présente ici plusieurs familles de schémas construits selon différents principes et d'ordres divers et variés. On suppose les solutions aussi régulières que nécessaire.

6.1 Schémas de type Taylor

Fixons un entier $p \geq 1$ et écrivons le développement de Taylor de y à l'ordre p au point t_n, avec un reste vague en O comme précédemment, en supposant comme toujours y suffisamment régulière pour cela,

$$y(t_{n+1}) = y(t_n) + \sum_{k=1}^{p} \frac{h^k}{k!} y^{(k)}(t_n) + O(h^{p+1}). \tag{6.1.1}$$

On a vu que les dérivées $y^{(k)}(t_n)$ satisfont

$$y^{(k)}(t_n) = f^{k-1}(t_n, y(t_n)),$$

où les fonctions f^k sont définies par (4.1.11). Le développement de Taylor (6.1.1). peut donc s'écrire

$$y(t_{n+1}) = y(t_n) + \sum_{k=1}^{p} \frac{h^k}{k!} f^{k-1}(t_n, y(t_n)) + O(h^{p+1}).$$

En supprimant le reste en $O(h^{p+1})$ et en remplaçant $y(t_n)$ par une approximation potentielle y_n, on obtient une famille de schémas numériques indexée par p, [1]

$$y_{n+1} = y_n + h \sum_{k=1}^{p} \frac{h^{k-1}}{k!} f^{k-1}(t_n, y_n). \tag{6.1.2}$$

Ces schémas sont de la forme (4.1.1), c'est-à-dire explicites et à un pas, avec

$$F(t, y, h) = \sum_{k=1}^{p} \frac{h^{k-1}}{k!} f^{k-1}(t, y).$$

1. On ne fait pas apparaître cette indexation par p dans la notation.

1. Pour $p = 1$, on retrouve le schéma d'Euler $y_{n+1} = y_n + hf(t_n, y_n)$.
2. Pour $p = 2$, on a obtient un nouveau schéma

$$y_{n+1} = y_n + hf(t_n, y_n) + \frac{h^2}{2}\left(\frac{\partial f}{\partial t}(t_n, y_n) + \sum_{j=1}^{m} \frac{\partial f}{\partial y_j}(t_n, y_n) f_j(t_n, y_n)\right).$$

Proposition 6.1.1 *Si f est de classe C^p et est lipschitzienne par rapport à y, uniformément par rapport à t, ainsi que toutes ses dérivées partielles successives, alors le schéma* (6.1.2) *est stable et d'ordre p. Il est donc convergent d'ordre p.*

Démonstration. La somme et le produit de deux fonctions lipschitziennes sont lipschitziens. Il est clair (et si cela n'est pas clair, alors on le démontre par récurrence) que la fonction F est polynomiale par rapport aux dérivées partielles de f et à h. Comme $h \in [0, 1]$, on déduit de la remarque qui précède que F est lipschitzienne par rapport à y, uniformément par rapport à t et h. Le schéma est donc stable. Il est par ailleurs consistant et d'ordre p par construction, puisque l'erreur de consistance n'est autre que le reste $O(h^{p+1})$ de la formule de Taylor. ◊

On s'en doute, l'inconvénient des schémas de type Taylor réside dans les calculs de dérivées qui interviennent dans le calcul des fonctions f^{k-1}. Ceux-ci peuvent être très compliqués et numériquement coûteux.

En guise d'exemple, calculons f^2 dans le cas scalaire $m = 1$. On a

$$f^1(t, y) = \frac{\partial f}{\partial t}(t, y) + \frac{\partial f}{\partial y}(t, y) f(t, y),$$

donc

$$\begin{aligned} f^2(t, y) &= \frac{\partial f^1}{\partial t}(t, y) + \frac{\partial f^1}{\partial y}(t, y) f(t, y) \\ &= \frac{\partial}{\partial t}\left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} f\right)(t, y) + \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} f\right)(t, y) f(t, y) \\ &= \left(\frac{\partial^2 f}{\partial t^2} + 2\frac{\partial^2 f}{\partial t \partial y} f + \frac{\partial f}{\partial y}\frac{\partial f}{\partial t} + \frac{\partial^2 f}{\partial y^2} f^2 + \left(\frac{\partial f}{\partial y}\right)^2 f\right)(t, y). \end{aligned}$$

Noter cependant que dans le cas simple $f(t, y) = \lambda y$, λ étant une constante, la relation (6.1.2) se simplifie en

$$y_{n+1} = \left(\sum_{k=0}^{p} \frac{(\lambda h)^k}{k!}\right) y_n. \tag{6.1.3}$$

Récemment, les méthodes de Taylor ont connu un nouvel essor avec le développement de la différentiation automatique (DA). Ce terme regroupe des logiciels prenant en entrée un programme écrit dans un langage donné et destiné à calculer numériquement une fonction donnée, et produisant en sortie un autre programme qui lui calcule numériquement les dérivées de cette fonction. [2] La différentiation automatique permet dans une certaine mesure de pallier les inconvénients des méthodes de Taylor : on ne calcule pas les dérivées nécessaires à la main ni ne les implémente, mais on laisse le programme de DA produire à partir d'un programme calculant f un autre programme qui va s'en charger automatiquement. C'est peut-être plus facile à écrire qu'à réaliser en pratique...

2. À ne pas confondre avec la différentiation numérique, qui utilise des quotients différentiels pour approcher les dérivées, ni avec le calcul formel qui manipule formellement les expressions mathématiques.

6.2 Schémas de Runge-Kutta.

Les méthodes de Runge-Kutta [3] sont fondées sur l'intégration de l'EDO entre t_n et t_{n+1}. En plus de la valeur en t_n, nous allons nous servir d'un certain nombre de valeurs intermédiaires correspondant à des instants intermédiaires entre t_n et t_{n+1}. Ces valeurs intermédiaires sont combinées entre elles pour construire l'approximation suivante à l'instant t_{n+1}. On ne les réutilise plus, elles sont donc jetées à l'itération suivante, qui va elle faire intervenir des instants intermédiaires entre t_{n+1} et t_{n+2}, et ainsi de suite. Contrairement peut-être aux apparences, ce sont donc des schémas à 1 pas, simplement le passage d'un pas au suivant est un peu tortueux. En particulier, elles s'initialisent d'elles-mêmes à partir de la donnée initiale du problème de Cauchy.

Pour simplifier, on ne va décrire les méthodes de Runge-Kutta que dans le cas scalaire, $m = 1$, mais elles sont également applicables dans le cas vectoriel, $m \geq 1$. Dans ce qui suit, f, y, y_i, etc. sont donc à valeurs réelles.

On se donne pour commencer $q \geq 1$ réels $0 \leq c_1 \leq \ldots \leq c_q \leq 1$, q réels $b_1, \ldots, b_q$ et une formule de quadrature, ou d'intégration numérique, [4] sur l'intervalle $[0, 1]$

$$\int_0^1 \varphi(s)\, ds \approx \sum_{j=1}^{q} b_j \varphi(c_j), \tag{6.2.1}$$

dont les c_j sont les nœuds et les b_j les poids – voir le chapitre 7 pour une étude systématique. On suppose que cette formule est exacte pour les fonctions constantes, c'est-à-dire que

$$\sum_{j=1}^{q} b_j = 1. \tag{6.2.2}$$

Les c_i servent à définir q instants intermédiaires $t_{n,i}$ (pas forcément distincts) entre t_n et t_{n+1},

$$t_{n,i} = t_n + c_i h, \quad 1 \leq i \leq q. \tag{6.2.3}$$

Pour chaque instant intermédiaire $t_{n,i}$ (ou à chaque c_i), on choisit une formule de quadrature utilisant les nœuds c_j de la façon suivante

$$\int_0^{c_i} \varphi(s)\, ds \approx \sum_{j=1}^{q} a_{ij} \varphi(c_j), \tag{6.2.4}$$

qu'on supposera aussi exacte pour les constantes

$$\sum_{j=1}^{q} a_{ij} = c_i. \tag{6.2.5}$$

En intégrant l'EDO (2.1.1) entre les instants t_n et $t_{n,i}$, on obtient

$$y(t_{n,i}) - y(t_n) = \int_{t_n}^{t_{n,i}} f\big(t, y(t)\big)\, dt = h \int_0^{c_i} f\big(t_n + sh, y(t_n + sh)\big)\, ds, \tag{6.2.6}$$

3. Carl David Tolmé Runge, 1856–1927 ; Martin Wilhelm Kutta, 1867–1944.

4. C'est-à-dire une formule d'approximation numérique de l'intégrale. La nature regorge de formules de quadrature.

via le changement de variable qui s'impose, $t = t_n + sh$. L'utilisation de la formule de quadrature (6.2.4) pour discrétiser l'intégrale apparaissant dans (6.2.6) conduit aux relations

$$y(t_{n,i}) \approx y(t_n) + h \sum_{j=1}^{q} a_{ij} f\big(t_{n,j}, y(t_{n,j})\big).$$

Comme d'habitude, on remplace alors les valeurs exactes $y(t_n)$ et $y(t_{n,i})$ que nous ne connaissons pas, par des approximations potentielles y_n et $y_{n,i}$, et l'on pose alors

$$y_{n,i} = y_n + h \sum_{j=1}^{q} a_{ij} f\big(t_{n,j}, y_{n,j}\big), \tag{6.2.7}$$

pour $i = 1, \dots, q$. Notons que si $a_{ij} = 0$ pour $i \leq j$, les relations (6.2.7) définissent $y_{n,i}$ de façon explicite : d'abord $y_{n,1} = y_n$, puis $y_{n,2} = y_n + ha_{21} f(t_{n,1}, y_{n,1})$, etc. Dans le cas général, il faut calculer les $y_{n,i}$ en résolvant un système non linéaire de q équations à q inconnues

$$\begin{pmatrix} y_{n,1} \\ \vdots \\ y_{n,q} \end{pmatrix} = \begin{pmatrix} y_n \\ \vdots \\ y_n \end{pmatrix} + hA \begin{pmatrix} f(t_{n,1}, y_{n,1}) \\ \vdots \\ f(t_{n,q}, y_{n,q}) \end{pmatrix}, \tag{6.2.8}$$

où $A = (a_{ij})_{1 \leq i,j \leq q}$ est une matrice de taille $q \times q$ que l'on s'est donnée. Les $y_{n,i}$, s'ils existent, sont donc fonction de y_n (et de t_n et de h, bien sûr). On rappelle que la méthode de Newton vue au chapitre 5 peut être utilisée pour approcher numériquement ces solutions.

En intégrant enfin l'EDO entre les instants t_n et t_{n+1}, on obtient

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f\big(t, y(t)\big)\, dt = h \int_0^1 f\big(t_n + sh, y(t_n + sh)\big)\, ds.$$

On discrétise cette dernière intégrale par la formule de quadrature (6.2.1), on remplace les valeurs exactes par les approximations supposées et l'on obtient le schéma de Runge-Kutta

$$y_{n+1} = y_n + h \sum_{j=1}^{q} b_j f\big(t_{n,j}, y_{n,j}\big), \tag{6.2.9}$$

les valeurs intermédiaires $y_{n,j}$ étant, rappelons le, calculées par (6.2.7). On voit que in fine, y_{n+1} est seulement fonction de y_n (et de t_n et de h), à travers les $y_{n,j}$, et le schéma est bien à un seul pas. On oublie alors les $y_{n,j}$ et l'on recommence en t_{n+1} à partir de y_{n+1}.

Un schéma de Runge-Kutta à q valeurs intermédiaires est donc déterminé par la donnée des q instants intermédiaires via les valeurs c_i, et de la matrice A et des valeurs b_j qui donnent les poids des $q + 1$ formules de quadrature utilisées. Ces paramètres sont ajustés de façon à rendre les schémas d'ordre le plus élevé possible, voire également avoir d'autres propriétés désirables plus subtiles dont nous ne parlerons pas ici. On présente souvent ces schémas de façon compacte comme des tableaux de la forme 6.1, appelés *tableaux de Butcher* [5].

5. John Charles Butcher, 1933–.

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1q} \\ \vdots & \vdots & & \vdots \\ c_q & a_{q1} & \dots & a_{qq} \\ \hline & b_1 & \dots & b_q \end{array}$$

TABLE 6.1 – Tableau de Butcher d'un schéma de Runge-Kutta

Revenons sur le calcul des valeurs intermédiaires dans le cas général. Le vecteur dont les composantes sont ces valeurs est donc un point fixe de l'application de $\mathbb{R}^q$ dans $\mathbb{R}^q$ définie par

$$\psi : z = \begin{pmatrix} z_1 \\ \vdots \\ z_q \end{pmatrix} \longmapsto \psi(z) = y_n \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + hA \begin{pmatrix} f(t_{n,1}, z_1) \\ \vdots \\ f(t_{n,q}, z_q) \end{pmatrix}.$$

Prolongeons cette fonction à y et t (qui sont des paramètres réels) quelconques en posant

$$\psi_y(z) = y \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + hA \begin{pmatrix} f(t + c_1 h, z_1) \\ \vdots \\ f(t + c_q h, z_q) \end{pmatrix} = ye + hAG(z),$$

où

$$e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^q \quad \text{et} \quad G(z) = \begin{pmatrix} f(t + c_1 h, z_1) \\ \vdots \\ f(t + c_q h, z_q) \end{pmatrix} \in \mathbb{R}^q.$$

Dans le cas où $y = y_n$ et $t = t_n$, c'est bien la même fonction. Elle dépend bien sûr également de t et de h, même si l'on ne le fait pas apparaître dans la notation.

Proposition 6.2.1 *Soit $f : [0, T] \times \mathbb{R} \to \mathbb{R}$ lipschitzienne par rapport à y uniformément par rapport à t de constante de Lipschitz L. On se donne une méthode de Runge-Kutta pour le problème de Cauchy associé à f, définie par les coefficients c_i, b_j et la matrice $A = (a_{ij})$. Soit $0 < h_0 < \frac{1}{L\|A\|_\infty}$, où $\|A\|_\infty$ désigne la norme matricielle subordonnée à la norme infinie sur $\mathbb{R}^q$. Alors pour tout $h \leq h_0$, ce schéma de Runge-Kutta est bien défini. Il est en outre stable sous la même condition.*

Démonstration. Il est commode d'utiliser ici la norme infinie $\|z\|_\infty = \max_i |z_i|$ sur $\mathbb{R}^q$. On note $\|A\|_\infty$ la norme matricielle subordonnée. [6] Fixons y, t et h. On a donc pour tous z et $\tilde{z}$ dans $\mathbb{R}^q$,

$$\begin{aligned} \|\psi_y(z) - \psi_y(\tilde{z})\|_\infty &= h\|A(G(z) - G(\tilde{z})\|_\infty \\ &\leq h\|A\|_\infty \|G(z) - G(\tilde{z})\|_\infty \\ &= h\|A\|_\infty \max_{1 \leq i \leq q} |f(t + c_i h, z_i) - f(t + c_i h, \tilde{z}_i)| \\ &\leq h\|A\|_\infty L \max_{1 \leq i \leq q} |z_i - \tilde{z}_i| \\ &= hL\|A\|_\infty \|z - \tilde{z}\|_\infty. \end{aligned}$$

6. Cette norme est donnée par $\|A\|_\infty = \max_i \sum_j |a_{ij}|$.

Par conséquent, pour $h < \frac{1}{L\|A\|_\infty}$, l'application ψ_y est strictement contractante et admet donc un point fixe unique dans $\mathbb{R}^q$ qui est le vecteur des valeurs intermédiaires recherché. Ceci montre que le schéma de Runge-Kutta est bien défini pour ces valeurs de h en prenant $y = y_n$ et $t = t_n$. Montrons maintenant que le schéma est stable. Il faut donc montrer que l'application $F(t, y, h)$ qui définit le schéma est lipschitzienne par rapport à y, uniformément par rapport à t et h (pour $h \in [0, h_0]$). Pour cela, notons $\mathrm{int}\,(y) \in \mathbb{R}^q$ le vecteur des valeurs intermédiaires précédemment obtenu par point fixe. [7] Au vu de (6.2.9), on a

$$F(t, y, h) = \sum_{j=1}^{q} b_j f(t + c_j h, \mathrm{int}\,(y)_j), \tag{6.2.10}$$

d'où pour tout couple de réels $(y, \tilde{y})$,

$$\begin{aligned}
|F(t, y, h) - F(t, \tilde{y}, h)| &\leq \sum_{j=1}^{q} |b_j||f(t + c_j h, \mathrm{int}\,(y)_j) - f(t + c_j h, \mathrm{int}\,(\tilde{y})_j)| \\
&\leq L \sum_{j=1}^{q} |b_j||\mathrm{int}\,(y)_j - \mathrm{int}\,(\tilde{y})_j| \\
&\leq L \max_j |\mathrm{int}\,(y)_j - \mathrm{int}\,(\tilde{y})_j| \sum_{j=1}^{q} |b_j| \\
&= L\|b\|_1 \|\mathrm{int}\,(y) - \mathrm{int}\,(\tilde{y})\|_\infty.
\end{aligned}$$

Or on a, par la propriété de point fixe,

$$\mathrm{int}\,(y) = ye + hAG(\mathrm{int}\,(y)) \text{ et } \mathrm{int}\,(\tilde{y}) = \tilde{y}e + hAG(\mathrm{int}\,(\tilde{y})).$$

Par conséquent,

$$\|\mathrm{int}\,(y) - \mathrm{int}\,(\tilde{y})\|_\infty \leq |y - \tilde{y}| + hL\|A\|_\infty \|\mathrm{int}\,(y) - \mathrm{int}\,(\tilde{y})\|_\infty,$$

par le même calcul que plus haut, et donc

$$\|\mathrm{int}\,(y) - \mathrm{int}\,(\tilde{y})\|_\infty \leq \frac{1}{1 - hL\|A\|_\infty} |y - \tilde{y}|$$

dès que $h < \frac{1}{L\|A\|_\infty}$. On en déduit donc que

$$|F(t, y, h) - F(t, \tilde{y}, h)| \leq \frac{L\|b\|_1}{1 - h_0 L\|A\|_\infty} |y - \tilde{y}|,$$

dès que $h \leq h_0 < \frac{1}{L\|A\|_\infty}$, d'où la stabilité. [8] ◊

7. Ce vecteur dépend aussi de t et de h, mais ce n'est pas ce qui nous importe, donc on ne l'écrit pas. Toutes les estimations sont uniformes par rapport à t et h.

8. On n'a pas le caractère uniforme pour $h \in [0, 1]$, mais seulement pour $h \in [0, h_0]$, mais cela n'a clairement aucune importance pour la stabilité.

Naturellement, dans le cas d'un schéma explicite, $a_{ij} = 0$ pour $i \leq j$, il n'y a en réalité pas de restriction sur le pas, c'est juste que l'analyse précédente n'est pas optimale, à ce propos voir proposition 6.2.3 page 103. Remarquons aussi que si les coefficients b_j sont tous positifs, alors $\|b\|_1 = 1$.

Notons à ce sujet que la classification générale schéma explicite/schéma implicite ne s'applique pas telle quelle aux schémas de Runge-Kutta. Un schéma de Runge-Kutta ne fait jamais intervenir y_{n+1} dans une équation implicite à résoudre. Le caractère implicite ou explicite d'un schéma de Runge-Kutta dépend uniquement du calcul des valeurs intermédiaires, selon que celui-ci nécessite la résolution d'un système non linéaire ou pas. Un schéma de Runge-Kutta s'écrit donc bien sous la forme $y_{n+1} = y_n + hF(t_n, y_n, h)$, mais la fonction F n'est une formule explicite que si les étapes intermédiaires se déroulent explicitement, sans résolution d'équation. Sinon, on n'a pas de formule pour F, voir la note de bas de page numéro 2, page 45, à ce propos. Une fois les valeurs intermédiaires calculées, il n'y a plus d'équation à résoudre pour obtenir le pas suivant.

Remarque 6.2.1 Pour voir que les valeurs aux instants intermédiaires $y_{n,i}$ ne sont pas vraiment essentielles dans un schéma de Runge-Kutta, notons que l'on peut réécrire celui-ci sous la forme suivante, où les inconnues intermédiaires deviennent plutôt les $f_{n,i} = f(t_{n,i}, y_{n,i})$. En effet,

$$f_{n,i} = f(t_{n,i}, y_{n,i}) = f\left(t_{n,i}, y_n + h\sum_{j=1}^{q} a_{ij} f(t_{n,j}, y_{n,j})\right),$$

si bien que la partie intermédiaire du schéma peut se réécrire sous la forme

$$f_{n,i} = f\left(t_{n,i}, y_n + h\sum_{j=1}^{q} a_{ij} f_{n,j}\right) \text{ pour } i = 1, \ldots, q.$$

Réciproquement, un q-uplet $f_{n,i}$ solution de ce système non linéaire permet de reconstruire le q-uplet des $y_{n,i}$, dont l'existence et l'unicité est assurée pour h suffisamment petit. On a donc affaire à une formulation équivalente de cette étape du schéma.

La deuxième étape du schéma s'écrit alors

$$y_{n+1} = y_n + h\sum_{j=1}^{q} b_j f_{n,j},$$

et l'on n'a jamais fait allusion aux valeurs intermédiaires $y_{n,i}$.

Dans ce cas explicite au sens des schémas de Runge-Kutta, c'est-à-dire quand la matrice A est strictement triangulaire inférieure et que le calcul des valeurs intermédiaires ne nécessite donc pas de résoudre un système d'équations, on a une description simple de chaque itération de l'algorithme.

Algorithme de Runge-Kutta explicite

$tab(1) = f(t_n, y_n)$ $\quad$ $\{tab(j)$ contient $f_{n,j}\}$
Pour $i = 2 \nearrow q$
    $s = 0$
    Pour $j = 1 \nearrow i - 1$
        $s = s + a_{ij} * tab(j)$
    Fin j
    $tab(i) = f(t_n + c_i h, y_n + hs)$
Fin i
$s = 0$
Pour $i = 1 \nearrow q$
    $s = s + b_i * tab(i)$
Fin i
$y_{n+1} = y_n + hs$

Dans le cas implicite, il faut résoudre numériquement le système donnant les valeurs intermédiaires à l'aide d'une méthode itérative (type point fixe ou Newton).

Proposition 6.2.2 *Sous les mêmes hypothèses, un schéma de Runge-Kutta est consistant.*

Démonstration. On reprend la formule (6.2.10) en explicitant la dépendance du point fixe par rapport aux autres paramètres :

$$F(t, y, h) = \sum_{j=1}^{q} b_j f(t + c_j h, \operatorname{int}(t, y, h)_j).$$

On sait [9] qu'un point fixe d'une application strictement contractante dépendant continûment de paramètres, dépend lui-même continûment de ces paramètres. Ici, l'application $(t, y, h) \mapsto \operatorname{int}(t, y, h)$ est donc continue de $[0, T] \times \mathbb{R} \times [0, h_0]$ à valeurs dans $\mathbb{R}^q$. Il s'ensuit que F est également continue comme composition de fonctions continues.

Pour $h = 0$, l'unique point fixe est trivial, c'est $\operatorname{int}(t, y, 0) = ye$. Par conséquent,

$$F(t, y, 0) = \sum_{j=1}^{q} b_j f(t + c_j 0, \operatorname{int}(t, y, 0)_j) = \sum_{j=1}^{q} b_j f(t, y) = f(t, y),$$

puisque $\sum_{j=1}^{q} b_j = 1$, d'où la consistance du schéma. ◊

Donnons quelques exemples de schémas de Runge-Kutta qui permettent aussi de se faire une idée de pourquoi ces schémas marchent, sans entrer dans la théorie générale.

1. Cas $q = 1$. D'après (6.2.2), on a $b_1 = 1$ et il y a un seul instant intermédiaire $t_{n,1}$. Le schéma s'écrit

$$\begin{cases} y_{n,1} = y_n + a_{11} h f(t_{n,1}, y_{n,1}), \\ y_{n+1} = y_n + h f(t_{n,1}, y_{n,1}). \end{cases}$$

9. Si on ne sait pas, on le montre.

Si l'on souhaite imposer (6.2.5), il convient de prendre $a_{11} = c_1$ (cela n'a pas d'influence sur l'ordre du schéma). Il y a une infinité de choix possibles. Regardons ce qui se passe pour quelques valeurs particulières. Pour $c_1 = 0$, on obtient alors le schéma

$$\text{(RK1)} \qquad \begin{cases} y_{n,1} = y_n, \\ y_{n+1} = y_n + hf(t_n, y_n). \end{cases} \qquad \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

C'est le schéma d'Euler explicite, que nous appellerons aussi schéma RK1.

Pour $c_1 = \frac{1}{2}$, on obtient le nouveau schéma, que nous n'avons pas encore rencontré sous un autre nom,

$$\begin{cases} y_{n,1} = y_n + \dfrac{h}{2} f\left(t_n + \dfrac{h}{2}, y_{n,1}\right), \\ y_{n+1} = y_n + hf\left(t_n + \dfrac{h}{2}, y_{n,1}\right). \end{cases} \qquad \begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$$

Noter que, cette fois, le calcul de $y_{n,1}$ est implicite.

Pour $c_1 = 1$, on obtient le schéma

$$\begin{cases} y_{n,1} = y_n + hf(t_{n+1}, y_{n,1}), \\ y_{n+1} = y_n + hf(t_{n+1}, y_{n,1}). \end{cases} \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

C'est le schéma d'Euler implicite, un peu bizarrement écrit (en effet, $y_{n+1} = y_{n,1}$ par unicité du point fixe).

Tous les schémas précédents sont d'ordre 1.

2. Cas $q = 2$. Il y a un infinité de choix des deux instants intermédiaires $t_{n,1}$ et $t_{n,2}$ et des poids de quadrature, ce qui fait au total huit paramètres. Pour simplifier, nous allons considérer des schémas de Runge-Kutta explicites, lesquels correspondent au cas $a_{11} = a_{12} = a_{22} = 0$. Il reste donc à ce stade à choisir cinq paramètres : a_{21}, b_1, b_2, c_1 et c_2. D'après (6.2.2), on impose $b_1 + b_2 = 1$. D'après (6.2.5), on prend aussi $a_{21} = c_2$, ce qui n'est pas essentiel, mais diminue le nombre de paramètres. Encore pour avoir (6.2.5), il faut prendre $c_1 = 0$, c'est-à-dire $t_{n,1} = t_n$. On obtient ainsi une famille de schémas à deux paramètres, $c = c_2$ et $b = b_1 = 1 - b_2$, qui s'écrivent

$$\begin{cases} y_{n,1} = y_n \\ y_{n,2} = y_n + chf(t_n, y_{n,1}) \\ y_{n+1} = y_n + h\big[bf(t_n, y_{n,1}) + (1-b)f(t_n + ch, y_{n,2})\big], \end{cases} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ c & c & 0 \\ \hline & b & 1-b \end{array}$$

En remplaçant les pas intermédiaires par leur valeur, on voit que y_{n+1} est calculé à partir de y_n par la formule explicite

$$y_{n+1} = y_n + h\big[bf(t_n, y_n) + (1-b)f\big(t_n + ch, y_n + chf(t_n, y_n)\big)\big],$$

d'où

$$F(t, y, h) = bf(t, y) + (1-b)f\big(t + ch, y + chf(t, y)\big).$$

L'erreur de consistance du schéma est donc égale à

$$\begin{aligned} \varepsilon_n &= y(t_{n+1}) - y(t_n) - bhf(t_n, y(t_n)) - (1-b)hf\big(t_n + ch, y(t_n) + chf\big(t_n, y(t_n)\big)\big) \\ &= y(t_{n+1}) - y(t_n) - bhy'(t_n) - (1-b)hf\big(t_n + ch, y(t_n) + chy'(t_n)\big), \end{aligned} \tag{6.2.11}$$

où y est une solution assez régulière de l'EDO. Dans tous ces calculs, on remplace dès que possible tout terme de la forme $f(t_n, y(t_n))$ par sa valeur $y'(t_n)$ issue de l'EDO pour simplifier les expressions qui apparaissent. Nous allons déterminer l'ordre de la méthode en fonction des paramètres b et c. On écrit le développement de Taylor [10] de y au voisinage de t_n jusqu'à l'ordre 2 (il n'est pas réaliste d'espérer un ordre plus élevé a priori),

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

Rappelons que

$$y''(t) = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t))y'(t). \tag{6.2.12}$$

Un autre développement de Taylor à deux variables de la fonction f au voisinage de $(t_n, y(t_n))$ montre que

$$\begin{aligned} f\big(t_n + ch, y(t_n) + chy'(t_n)\big) &= f(t_n, y(t_n)) + ch\frac{\partial f}{\partial t}(t_n, y(t_n)) + chy'(t_n)\frac{\partial f}{\partial y}(t_n, y(t_n)) + O(h^2) \\ &= y'(t_n) + chy''(t_n) + O(h^2), \end{aligned}$$

à la vue de l'EDO et de celle du rappel précédent. En injectant ces deux développements de Taylor dans (6.2.11), on obtient

$$\begin{aligned} \varepsilon_n &= hy'(t_n) + \frac{h^2}{2}y''(t_n) - bhy'(t_n) - (1-b)hy'(t_n) - (1-b)ch^2y''(t_n) + O(h^3) \\ &= \frac{h^2}{2}\big(1 - 2(1-b)c\big)y''(t_n) + O(h^3). \end{aligned}$$

On voit que ces schémas sont tous d'ordre au moins égal à 1 et que l'ordre 2 est atteint pour $1 - 2(1-b)c = 0$, c'est-à-dire $b = 1 - \frac{1}{2c}$. On obtient donc une famille à un paramètre de schémas de Runge-Kutta explicites d'ordre 2

$$\begin{cases} y_{n,1} = y_n, \\ y_{n,2} = y_n + chf(t_n, y_{n,1}), \\ y_{n+1} = y_n + \dfrac{2c-1}{2c}hf(t_n, y_{n,1}) + \dfrac{1}{2c}hf(t_n + ch, y_{n,2}), \end{cases} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ c & c & 0 \\ \hline & \frac{2c-1}{2c} & \frac{1}{2c} \end{array}$$

définie pour $0 < c \leq 1$.

Pour $c = 1/2$, on a $t_{n,2} = t_n + h/2$ et

$$\begin{cases} y_{n,1} = y_n, \\ y_{n,2} = y_n + \dfrac{h}{2}f(t_n, y_{n,1}), \\ y_{n+1} = y_n + hf\left(t_n + \dfrac{h}{2}, y_{n,2}\right), \end{cases} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

C'est le schéma d'Euler modifié.

10. avec un reste en O un peu vague...

Pour $c = 1$, on obtient le *schéma de Heun* [11], que nous appellerons aussi schéma RK2,

$$\text{(RK2)} \qquad \begin{cases} y_{n,1} = y_n, \\ y_{n,2} = y_n + hf(t_n, y_{n,1}), \\ y_{n+1} = y_n + \dfrac{h}{2} f(t_n, y_{n,1}) + \dfrac{h}{2} f(t_{n+1}, y_{n,2}), \end{cases} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

On peut montrer que la famille de schémas

$$y_{n+1} = y_n + \frac{2d-1}{2d} hf(t_n, y_n) + \frac{1}{2d} hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\right)$$

est aussi une famille de schémas d'ordre 2. Les deux familles coïncident uniquement pour $d = c = \frac{1}{2}$.

Terminons le cas $q = 2$ par un exemple de schéma RK implicite. Il correspond au choix $c_1 = 0, c_2 = 1$, $a_{11} = a_{12} = 0, a_{21} = a_{22} = \frac{1}{2}$, $b_1 = b_2 = \frac{1}{2}$,

$$\begin{cases} y_{n,1} = y_n, \\ y_{n,2} = y_n + \dfrac{h}{2}[f(t_n, y_n) + f(t_{n+1}, y_{n,2})], \\ y_{n+1} = y_n + \dfrac{h}{2}[f(t_n, y_n) + f(t_{n+1}, y_{n,2})]. \end{cases} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

On reconnaît le schéma de Crank-Nicolson, écrit un peu bizarrement.

3. Pour $q = 4$, le schéma le plus utilisé est le suivant, appelé schéma RK4,

$$\text{(RK4)} \qquad \begin{cases} y_{n,1} = y_n, \\ y_{n,2} = y_n + \dfrac{h}{2} f(t_n, y_{n,1}), \\ y_{n,3} = y_n + \dfrac{h}{2} f\left(t_n + \dfrac{h}{2}, y_{n,2}\right), \\ y_{n,4} = y_n + hf\left(t_n + \dfrac{h}{2}, y_{n,3}\right), \\ y_{n+1} = y_n + \dfrac{h}{6}\left[f(t_n, y_{n,1}) + 2f\left(t_n + \dfrac{h}{2}, y_{n,2}\right) + 2f\left(t_n + \dfrac{h}{2}, y_{n,3}\right) + f(t_{n+1}, y_{n,4})\right]. \end{cases}$$

Il est explicite, d'ordre 4 et correspond au tableau de Butcher

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

C'est de ce schéma dont on parle quand on mentionne " la " méthode de Runge-Kutta sans préciser.

11. Karl Heun, 1859–1929.

La formulation alternative sans les valeurs intermédiaires du schéma RK4 s'écrit

$$\begin{cases} f_{n,1} = f(t_n, y_n), \\ f_{n,2} = f\left(t_n + \dfrac{h}{2}, y_n + \dfrac{h}{2} f_{n,1}\right), \\ f_{n,3} = f\left(t_n + \dfrac{h}{2}, y_n + \dfrac{h}{2} f_{n,2}\right), \\ f_{n,4} = f(t_n + h, y_n + h f_{n,3}), \\ y_{n+1} = y_n + \dfrac{h}{6}\big(f_{n,1} + 2f_{n,2} + 2f_{n,3} + f_{n,4}\big). \end{cases}$$

Dans le cas où $f(t, y) = g(t)$ de dépend pas de y, le schéma RK4 redonne la méthode de Simpson [12], dont il est bien connu qu'il s'agit d'une méthode d'intégration numérique d'ordre 4. En effet, dans ce cas $y_n = h\big(\frac{1}{6}g(t_0) + \frac{2}{3}\sum_{k=0}^{n-1} g(t_k + \frac{h}{2}) + \frac{1}{3}\sum_{k=1}^{n-1} g(t_k) + \frac{1}{6}g(t_n)\big)$.

Pour faire le lien avec la manière standard d'écrire les schémas explicites à un pas, récrivons le schéma RK4 sous la forme $y_{n+1} = y_n + hF(t_n, y_n, h)$. Il vient

$$\begin{aligned} y_{n+1} = y_n + \frac{h}{6}\Bigg[& f(t_n, y_n) + 2f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\right) \\ & + 2f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\right)\right) \\ & + f\left(t_n + h, y_n + hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\right)\right)\right)\Bigg], \end{aligned}$$

soit

$$\begin{aligned} F(t, y, h) = \frac{1}{6}\Bigg[& f(t, y) + 2f\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right) \\ & + 2f\left(t + \frac{h}{2}, y + \frac{h}{2} f\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right)\right) \\ & + f\left(t + h, y + hf\left(t + \frac{h}{2}, y + \frac{h}{2} f\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right)\right)\right)\Bigg], \end{aligned}$$

très clairement explicite, mais la vision sous forme de pas intermédiaires ou d'un tableau de Butcher est quand même un peu plus maniable.

Démonstration de l'ordre du schéma RK4. Montrons à la main que le schéma RK4 est bien d'ordre 4. On peut penser à utiliser la Proposition 4.1.14, mais la fonction F du schéma écrite ci-dessus n'est pas spécialement engageante, et il faut procéder par compositions successives, ce qui est compliqué pour des dérivées d'ordre élevé. De plus, le calcul des fonctions f^k, $k = 0, \dots, 3$ est déjà pénible. Procédons plutôt par développements de Taylor usuels. [13] L'idée pour simplifier au maximum et garder des calculs lisibles par l'être humain, est de remplacer aussi tôt que possible toute expression faisant intervenir f et ses dérivées partielles par des valeurs correspondantes de y et de ses dérivées. On va avoir besoin de développer $f(t_n, e_{n,1})$, $f\big(t_n + \frac{h}{2}, e_{n,2}\big)$, $f\big(t_n + \frac{h}{2}, e_{n,3}\big)$ et $f(t_{n+1}, e_{n,4})$ jusqu'à l'ordre 3, où les $e_{n,i}$ sont obtenus par les formules des pas intermédiaires

12. Thomas Simpson, 1710–1761. La formule était apparemment déjà utilisée par Kepler un bon siècle auparavant.

13. Cela revient en fait plus ou moins au même.

en remplaçant dans la première ligne y_n par $y(t_n)$, puis en descendant ainsi jusqu'à la quatrième ligne. Pour raccourcir la notation, on écrira $t_n = t$, toute dérivée partielle de f écrite sans argument est prise au point $(t, y(t))$ et toute dérivée de y écrite sans argument est prise au point t. Collationnons d'abord les dérivées successives de y par dérivation des fonctions composées.

$$\begin{aligned}
y' &= f,\\
y'' &= \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}y',\\
y''' &= \frac{\partial^2 f}{\partial t^2} + 2\frac{\partial^2 f}{\partial t\partial y}y' + \frac{\partial^2 f}{\partial y^2}(y')^2 + \frac{\partial f}{\partial y}y'',\\
y^{(4)} &= \frac{\partial^3 f}{\partial t^3} + 3\frac{\partial^3 f}{\partial t^2\partial y}y' + 3\frac{\partial^3 f}{\partial t\partial y^2}(y')^2 + \frac{\partial^3 f}{\partial y^3}(y')^3 + 3\frac{\partial^2 f}{\partial t\partial y}y'' + 3\frac{\partial^2 f}{\partial y^2}y'y'' + \frac{\partial f}{\partial y}y'''.
\end{aligned} \tag{6.2.13}$$

Pour s'entraîner, on pourra commencer par le cas où f ne dépend pas de t, ce qui divise par deux la longueur des expressions ci-dessus, mais allons-y dans le cas général.

On va tout baser sur le développement de Taylor à l'ordre 3 de la fonction $s \mapsto f(t + sh\alpha, y + sh\beta)$ entre 0 et 1, pour diverses valeurs de α et β. Il vient donc

$$\begin{aligned}
f(t + h\alpha, y + h\beta) = f &+ h\left[\frac{\partial f}{\partial t}\alpha + \frac{\partial f}{\partial y}\beta\right] + \frac{h^2}{2}\left[\frac{\partial^2 f}{\partial t^2}\alpha^2 + 2\frac{\partial^2 f}{\partial t\partial y}\alpha\beta + \frac{\partial^2 f}{\partial y^2}\beta^2\right]\\
&+ \frac{h^3}{6}\left[\frac{\partial^3 f}{\partial t^3}\alpha^3 + 3\frac{\partial^3 f}{\partial t^2\partial y}\alpha^2\beta + 3\frac{\partial^3 f}{\partial t\partial y^2}\alpha\beta^2 + \frac{\partial^3 f}{\partial y^3}\beta^3\right] + O(h^4).
\end{aligned} \tag{6.2.14}$$

On a donc d'abord $e_{n,1} = y$. Là c'est facile avec $\alpha = \beta = 0$,

$$f(t, e_{n,1}) = f = y' \tag{6.2.15}$$

pour le premier terme à développer. Ensuite vient par définition du schéma

$$e_{n,2} = y + \frac{h}{2}y', \tag{6.2.16}$$

d'où, prenant $\alpha = \frac{1}{2}$, $\beta = \frac{y'}{2}$,

$$\begin{aligned}
f\left(t + \frac{h}{2}, e_{n,2}\right) &= f + \frac{h}{2}\left[\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}y'\right] + \frac{h^2}{8}\left[\frac{\partial^2 f}{\partial t^2} + 2\frac{\partial^2 f}{\partial t\partial y}y' + \frac{\partial^2 f}{\partial y^2}(y')^2\right]\\
&\quad + \frac{h^3}{48}\left[\frac{\partial^3 f}{\partial t^3} + 3\frac{\partial^3 f}{\partial t^2\partial y}y' + 3\frac{\partial^3 f}{\partial t\partial y^2}(y')^2 + \frac{\partial^3 f}{\partial y^3}(y')^3\right] + O(h^4)\\
&= y' + \frac{h}{2}y'' + \frac{h^2}{8}\left[y''' - \frac{\partial f}{\partial y}y''\right]\\
&\quad + \frac{h^3}{48}\left[y^{(4)} - 3\frac{\partial^2 f}{\partial t\partial y}y'' - 3\frac{\partial^2 f}{\partial y^2}y'y'' - \frac{\partial f}{\partial y}y'''\right] + O(h^4).
\end{aligned} \tag{6.2.17}$$

Par conséquent,

$$e_{n,3} = y + \frac{h}{2}y' + \frac{h^2}{4}y'' + \frac{h^3}{16}\left[y''' - \frac{\partial f}{\partial y}y''\right] + O(h^4), \tag{6.2.18}$$

(remarquons que l'on doit heureusement rejeter le gros terme compliqué dans le reste à chaque multiplication par h) que l'on réutilise pour développer avec $\alpha = \frac{1}{2}$, $\beta = \frac{1}{2}\left(y' + \frac{h}{2}y'' + \frac{h^2}{8}\left[y''' - \frac{\partial f}{\partial y}y''\right]\right) + O(h^3)$,

$$\begin{aligned}
f\left(t + \frac{h}{2}, e_{n,3}\right) = f &+ \frac{h}{2}\left[\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}\left(y' + \frac{h}{2}y'' + \frac{h^2}{8}\left(y''' - \frac{\partial f}{\partial y}y''\right)\right)\right]\\
&+ \frac{h^2}{8}\left[\frac{\partial^2 f}{\partial t^2} + 2\frac{\partial^2 f}{\partial t\partial y}\left(y' + \frac{h}{2}y''\right) + \frac{\partial^2 f}{\partial y^2}\left((y')^2 + hy'y''\right)\right]
\end{aligned}$$

(en prenant soin de ne pas développer inutilement trop loin)

$$+\frac{h^3}{48}\left[\frac{\partial^3 f}{\partial t^3}+3\frac{\partial^3 f}{\partial t^2\partial y}y'+3\frac{\partial^3 f}{\partial t\partial y^2}(y')^2+\frac{\partial^3 f}{\partial y^3}(y')^3\right]+O(h^4)$$

(même soin)

(6.2.19)

Réarrangeons les puissances de h dans cette dernière expression. Il vient

$$\begin{aligned}f\left(t+\frac{h}{2},e_{n,3}\right)&=f+\frac{h}{2}\left[\frac{\partial f}{\partial t}+\frac{\partial f}{\partial y}y'\right]\\&\quad+\frac{h^2}{8}\left[\frac{\partial^2 f}{\partial t^2}+2\frac{\partial^2 f}{\partial t\partial y}y'+\frac{\partial^2 f}{\partial y^2}(y')^2+2\frac{\partial f}{\partial y}y''\right]\\&\quad+\frac{h^3}{48}\left[\frac{\partial^3 f}{\partial t^3}+3\frac{\partial^3 f}{\partial t^2\partial y}y'+3\frac{\partial^3 f}{\partial t\partial y^2}(y')^2+\frac{\partial^3 f}{\partial y^3}(y')^3\right.\\&\quad\left.+3\frac{\partial f}{\partial y}\left(y'''-\frac{\partial f}{\partial y}y''\right)+6\frac{\partial^2 f}{\partial t\partial y}y''+6\frac{\partial^2 f}{\partial y^2}y'y''\right]+O(h^4).\end{aligned}\tag{6.2.20}$$

Relisant alors les formules (6.2.13), on en déduit

$$\begin{aligned}f\left(t+\frac{h}{2},e_{n,3}\right)&=y'+\frac{h}{2}y''+\frac{h^2}{8}\left[y'''+\frac{\partial f}{\partial y}y''\right]\\&\quad+\frac{h^3}{48}\left[y^{(4)}+3\frac{\partial^2 f}{\partial t\partial y}y''+3\frac{\partial^2 f}{\partial y^2}y'y''+2\frac{\partial f}{\partial y}y'''-3\left(\frac{\partial f}{\partial y}\right)^2y''\right]+O(h^4).\end{aligned}\tag{6.2.21}$$

On refait le tri dans les puissances de h,

$$e_{n,4}=y+hy'+\frac{h^2}{2}y''+\frac{h^3}{8}\left[y'''+\frac{\partial f}{\partial y}y''\right]+O(h^4).\tag{6.2.22}$$

Il reste un dernier développement de f à faire avec $\alpha=1$ et $\beta=y'+\frac{h}{2}y''+\frac{h^2}{8}\left[y'''+\frac{\partial f}{\partial y}y''\right]+O(h^3)$

$$\begin{aligned}f\left(t+h,e_{n,4}\right)&=f+h\left[\frac{\partial f}{\partial t}+\frac{\partial f}{\partial y}\left(y'+\frac{h}{2}y''+\frac{h^2}{8}\left[y'''+\frac{\partial f}{\partial y}y''\right]\right)\right]\\&\quad+\frac{h^2}{2}\left[\frac{\partial^2 f}{\partial t^2}+2\frac{\partial^2 f}{\partial t\partial y}\left(y'+\frac{h}{2}y''\right)+\frac{\partial^2 f}{\partial y^2}\left((y')^2+hy'y''\right)\right]\\&\quad+\frac{h^3}{6}\left[\frac{\partial^3 f}{\partial t^3}+3\frac{\partial^3 f}{\partial t^2\partial y}y'+3\frac{\partial^3 f}{\partial t\partial y^2}(y')^2+\frac{\partial^3 f}{\partial y^3}(y')^3\right]+O(h^4)\end{aligned}\tag{6.2.23}$$

Réarrangeons,

$$\begin{aligned}f\left(t+h,e_{n,4}\right)&=f+h\left[\frac{\partial f}{\partial t}+\frac{\partial f}{\partial y}y'\right]\\&\quad+\frac{h^2}{2}\left[\frac{\partial^2 f}{\partial t^2}+2\frac{\partial^2 f}{\partial t\partial y}y'+\frac{\partial^2 f}{\partial y^2}(y')^2+\frac{\partial f}{\partial y}y''\right]\\&\quad+\frac{h^3}{6}\left[\frac{\partial^3 f}{\partial t^3}+3\frac{\partial^3 f}{\partial t^2\partial y}y'+3\frac{\partial^3 f}{\partial t\partial y^2}(y')^2+\frac{\partial^3 f}{\partial y^3}(y')^3\right.\\&\quad\left.+\frac{3}{4}\frac{\partial f}{\partial y}\left(y'''+\frac{\partial f}{\partial y}y''\right)+3\left(\frac{\partial^2 f}{\partial t\partial y}y''+\frac{\partial^2 f}{\partial y^2}y'y''\right)\right]+O(h^4).\end{aligned}\tag{6.2.24}$$

Relisons la liste des dérivées de y,

$$f\left(t+h,e_{n,4}\right)=y'+hy''+\frac{h^2}{2}y'''+\frac{h^3}{6}\left[y^{(4)}-\frac{1}{4}\frac{\partial f}{\partial y}y'''+\frac{3}{4}\left(\frac{\partial f}{\partial y}\right)^2y''\right]+O(h^4).\tag{6.2.25}$$

Nous pouvons maintenant combiner les développements (6.2.15), (6.2.17), (6.2.21) et (6.2.25) pour obtenir

$$\begin{aligned}
f(t,e_{n,1})+2f\left(t+\frac{h}{2},e_{n,2}\right)+2f\left(t+\frac{h}{2},e_{n,3}\right)+f(t+h,e_{n,4}) = \\
y'+2\left(y'+\frac{h}{2}y''+\frac{h^2}{8}\left[y'''-\frac{\partial f}{\partial y}y''\right]+\frac{h^3}{48}\left[y^{(4)}-3\frac{\partial^2 f}{\partial t\partial y}y''-3\frac{\partial^2 f}{\partial y^2}y'y''-\frac{\partial f}{\partial y}y'''\right]\right) \\
+2\left(y'+\frac{h}{2}y''+\frac{h^2}{8}\left[y'''+\frac{\partial f}{\partial y}y''\right]+\frac{h^3}{48}\left[y^{(4)}+3\frac{\partial^2 f}{\partial t\partial y}y''+3\frac{\partial^2 f}{\partial y^2}y'y''+2\frac{\partial f}{\partial y}y'''-3\left(\frac{\partial f}{\partial y}\right)^2 y''\right]\right) \\
+y'+hy''+\frac{h^2}{2}y'''+\frac{h^3}{6}\left[y^{(4)}-\frac{1}{4}\frac{\partial f}{\partial y}y'''+\frac{3}{4}\left(\frac{\partial f}{\partial y}\right)^2 y''\right]+O(h^4) \\
=6y'+3hy''+h^2y'''+\frac{h^3}{4}y^{(4)}+O(h^4). \quad (6.2.26)
\end{aligned}$$

On obtient donc l'erreur de consistance

$$\varepsilon_n = y(t_{n+1})-y(t_n)-hy'(t_n)-\frac{h^2}{2}y''(t_n)-\frac{h^3}{6}y'''(t_n)-\frac{h^4}{24}y^{(4)}(t_n)+O(h^5)=O(h^5), \quad (6.2.27)$$

par le développement de Taylor de $y(t_{n+1})$ à l'ordre 4 en t_n, d'où l'ordre au moins 4 de la méthode RK4. ♢

Remarque 6.2.2 Bien sûr, le calcul précédent paraît miraculeux quand tous les termes baroques s'éliminent à la fin et n'explique pas vraiment de façon profonde pourquoi la méthode est au moins d'ordre 4, ni comment Runge (1895) et Kutta (1901) ont bien pu faire pour l'obtenir. On a quand même la satisfaction du travail manuel bien fait. Évidemment, le calcul ne montre pas que la méthode est exactement d'ordre 4, car il aurait fallu développer tous les pas intermédiaires jusqu'à l'ordre 4 pour s'assurer que le terme suivant d'ordre 5 dans l'erreur de consistance ne s'annule pas. C'est un peu décourageant, à la réflexion. On pourrait néanmoins penser s'aider de logiciels de calcul formel comme `Maxima` ou `xcas` pour rester dans le cadre des logiciels libres. ♢

Pour énoncer les résultats sur la stabilité et l'ordre de convergence des schémas de Runge-Kutta, on complète la définition matricielle de ces schémas en introduisant, en plus de la matrice $A=(a_{ij})$, les notations $|A|=(|a_{ij}|)$, la matrice des valeurs absolues des éléments de A, et $\rho(|A|)$ pour son rayon spectral (le plus grand des modules de ses valeurs propres, cf. lemme 3.3.2). On définit également

$$C=\begin{pmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & c_q \end{pmatrix}, \quad b=\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_q \end{pmatrix}, \quad e=\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

On admet les résultats suivants (voir [3], chapitre 5, paragraphe 5 pour les démonstrations). Dans ce qui suit L désigne comme d'habitude la constante de Lipschitz de f et $(t,y,h)\to F(t,y,h)$ désigne la fonction qui définit le schéma à un pas.

Proposition 6.2.3 *Sous l'hypothèse $hL\rho(|A|)<1$, le schéma de Runge-Kutta admet une solution unique et est stable. Si la fonction f est k fois continûment différentiable, alors la fonction F est aussi k fois continûment différentiable.*

Dans le cas explicite, la matrice $|A|$ est strictement triangulaire inférieure, donc son rayon spectral est nul. On en déduit que les méthodes de Runge-Kutta explicites sont stables sans restriction sur le pas h, ce qui était plus ou moins évident par composition de fonctions lipschitziennes. Ce résultat est donc un peu plus précis que celui de la proposition 6.2.1, puisque $\rho(|A|)\le\|A\|_\infty$ pour tout matrice A.

Proposition 6.2.4 *Une condition nécessaire et suffisante pour qu'une méthode de Runge-Kutta soit au moins d'ordre 1 (ou consistante) s'écrit $b^Te=1$.*

Une condition nécessaire et suffisante pour qu'une méthode de Runge-Kutta soit au moins d'ordre 2 *s'écrit*

$$b^T e = 1 \quad et \quad b^T Ce = b^T Ae = \frac{1}{2}.$$

Une condition nécessaire et suffisante pour qu'une méthode de Runge-Kutta soit au moins d'ordre 3 *s'écrit*

$$b^T e = 1, \quad b^T Ce = \frac{1}{2}, \quad b^T C^2 e = \frac{1}{3} \quad et \quad b^T ACe = \frac{1}{6}.$$

Pour que la méthode soit au moins d'ordre 4*, il faut et il suffit qu'on ait en plus*

$$b^T C^3 e = \frac{1}{4}, \quad b^T AC^2 e = \frac{1}{12}, \quad b^T A^2 Ce = \frac{1}{24} \quad et \quad b^T CACe = \frac{1}{8}.$$

Rappelons que nous n'avons considéré ici que le cas scalaire, $m = 1$, mais que les méthodes de Runge-Kutta marchent tout aussi bien pour les systèmes avec m quelconque, avec essentiellement aucune modification.

6.3 Méthodes d'Adams

Les méthodes d'Adams [14] sont des méthodes à pas multiples qui sont fondées sur l'interpolation polynomiale de Lagrange (voir chapitre 7). Remarquons que nous n'avons pas traité la théorie générale des schémas à pas multiples, consistance, ordre, stabilité et convergence. On fera au cas par cas. D'abord un petit rappel sur l'interpolation de Lagrange.

Théorème 6.3.1 *Pour tout choix de $q+1$ points $\alpha_0, \alpha_1, \ldots, \alpha_q$ distincts et tout jeu de $q+1$ valeurs β_j, il existe un unique polynôme P de degré inférieur ou égal à q qui vérifie*

$$\forall j \in \{0, 1, \ldots, q\}, \quad P(\alpha_j) = \beta_j.$$

Le polynôme P est appelé le polynôme d'interpolation de Lagrange *des valeurs β_j aux points α_j, $0 \le j \le q$. Il s'écrit :*

$$P(t) = \sum_{j=0}^{q} \beta_j \prod_{k \neq j} \left(\frac{t - \alpha_k}{\alpha_j - \alpha_k} \right). \tag{6.3.1}$$

Les méthodes d'Adams entrent dans la deuxième catégorie de schémas présentée au Chapitre 2, c'est-à-dire ceux reposant sur une approximation de l'intégrale dans la formule

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt.$$

Pour approcher l'intégrale $\int_{t_n}^{t_{n+1}} f(s, y(s)) ds$, on commence par utiliser le Théorème 6.3.1 pour déterminer le polynôme $P_{n,q}$ de degré inférieur à q qui interpole les valeurs $f(t_j, y(t_j))$ aux $q+1$ instants successifs $t_{n-g}, \ldots, t_{n+d}$ avec $d + g = q$. On se limitera aux deux cas $d = 0$ et $d = 1$. On remplace ensuite le calcul de (2.2.8) par le calcul de l'intégrale de $P_{n,q}$, en commettant une certaine erreur. On effectue exactement ce calcul d'intégrale à l'aide de la formule (7.1.1). Le résultat est une combinaison linéaire des valeurs $f(t_j, y(t_j))$, dont il faut déterminer les coefficients, qui ne dépendent ni de f ni de n. Enfin, dans une dernière

14. John Couch Adams, 1819–1892.

étape, on remplace les $y(t_j)$ inconnus par des approximations y_j, comme d'habitude, afin de définir effectivement les schémas numériques. En fait, on est en train de calculer dans cette dernière étape, l'intégrale d'un autre polynôme d'interpolation de Lagrange, toujours noté $P_{n,q}$, qui interpole les valeurs $f_j = f(t_j, y_j)$.

L'avantage de ces méthodes est que, les coefficients en question étant calculés une fois pour toutes de façon à produire des schémas d'ordre élevé, leur utilisation est relativement économique. En effet, elle ne demande que des évaluations de $f(t_j, y_j)$, qu'il faut faire de toutes façons au minimum au moins une fois, et que l'on réutilise autant de fois que nécessaire dans ces combinaisons linéaires très bon marché. En effet, ce sont les évaluations de f qui sont a priori les opérations les plus coûteuses en temps de calcul.

Les schémas d'Adams s'écrivent donc

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} P_{n,q}(t)\,dt. \tag{6.3.2}$$

où $P_{n,q}$ désigne donc le polynôme de degré inférieur ou égal à q qui interpole les valeurs $f_j = f(t_j, y_j)$ aux points t_j pour $n-q+d \le j \le n+d$. Noter que si $d = 0$, on aura uniquement besoin des valeurs de la solution approchée aux instants antérieurs à t_n. Cela conduira à des schémas explicites appelés schémas d'Adams-Bashforth [15]. Si $d = 1$, les schémas obtenus, appelés schémas d'Adams-Moulton [16], sont implicites.

1. Le schéma d'Adams-Bashforth pour $q = 0$, $d = 0$, est

$$\text{(AB1)} \qquad y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(t_n, y_n)\,dt = y_n + hf(t_n, y_n).$$

En effet, la valeur $f_n = f(t_n, y_n)$ est interpolée en t_n par le polynôme constant $t \mapsto f_n$. On retrouve le schéma d'Euler explicite, il est donc d'ordre un. Le schéma d'Adams-Moulton correspondant est

$$\text{(AM1)} \qquad y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(t_{n+1}, y_{n+1})\,dt = y_n + hf(t_{n+1}, y_{n+1}),$$

c'est le schéma d'Euler implicite, qui est aussi d'ordre un.

2. Le schéma d'Adams-Bashforth pour $q = 1$ consiste à interpoler les valeurs f_n et f_{n-1} en t_n et t_{n-1}. Il correspond à l'intégrale de

$$P_{n,1}(t) = f_n + (f_n - f_{n-1})\frac{t - t_n}{h}.$$

On obtient donc en faisant le changement de variable $s = \frac{t-t_n}{h}$,

$$\int_{t_n}^{t_{n+1}} P_{n,1}(t)\,dt = h\int_0^1 \big(f_n + (f_n - f_{n-1})s\big)\,ds = h\left(f_n\left[1 + \frac{s^2}{2}\right]_0^1 - f_{n-1}\left[\frac{s^2}{2}\right]_0^1\right),$$

15. Francis Bashforth, 1819–1912. Les méthodes d'Adams-Bashforth sont apparemment uniquement dues à Adams, Bashforth n'ayant fait que les mentionner dans un livre. Mais le vocabulaire s'est imposé.

16. Forest Ray Moulton, 1872–1952. Il semble que l'apport de Moulton ait consisté à remarquer que les méthodes d'Adams implicites pouvaient être utilisées dans un autre contexte, celui des schémas *prédicteur-correcteur*.

d'où le schéma

$$\text{(AB2)} \qquad y_{n+1} = y_n + h\left(\frac{3}{2}f_n - \frac{1}{2}f_{n-1}\right). \tag{6.3.3}$$

C'est un schéma à deux pas. Pour le démarrer, il faut donc calculer y_1, avec un autre schéma, à un pas. Ensuite, une fois calculé $f_n = f(t_n, y_n)$, on l'utilise pour le calcul de y_{n+1} avec le coefficient $\frac{3}{2}$, puis on le stocke pour l'utiliser à nouveau avec le coefficient $-\frac{1}{2}$ pour le calcul de y_{n+2}.

Proposition 6.3.2 *Le schéma* AB2 *est d'ordre deux.*

Démonstration. Comme toujours pour ces questions d'ordre de schémas, il s'agit de développements de Taylor plus ou moins longs.

$$\begin{aligned}
\varepsilon_n &= y(t_{n+1}) - y(t_n) - \frac{h}{2}(3f(t_n, y(t_n)) - f(t_{n-1}, y(t_{n-1}))) \\
&= y(t_n + h) - y(t_n) - \frac{h}{2}\big(3y'(t_n) - y'(t_n - h)\big) \\
&= hy'(t_n) + \frac{h^2}{2}y''(t_n) + O(h^3) - \frac{h}{2}\Big[3y'(t_n) - \Big(y'(t_n) - hy''(t_n) + O(h^2)\Big)\Big] \\
&= O(h^3).
\end{aligned}$$

On imagine bien que ce n'est pas par hasard si les coefficients tombent juste pour annuler tous les termes d'ordre inférieur à 2... ◊

Le schéma d'Adams-Moulton pour $q = 1$ correspond à l'intégrale de

$$P_{n,1}(t) = f_n + (f_{n+1} - f_n)\frac{t - t_n}{h}.$$

Il s'écrit

$$y_{n+1} = y_n + \frac{h}{2}(f_n + f_{n+1}).$$

On retrouve le schéma de Crank-Nicolson, d'ordre 2 également, mais qui est aussi un schéma à un pas.

3. Pour $q = 2$, on obtient un schéma explicite en déterminant le polynôme de degré au plus 2 tel que $P_{n,2}(t_{n-j}) = f_{n-j} = f(t_{n-j}, y_{n-j})$ pour $j = 0, 1, 2$. Le premier polynôme de base correspondant à l'interpolation au point t_n est donné par

$$L_1(t) = \frac{(t - t_{n-1})(t - t_{n-2})}{(t_n - t_{n-1})(t_n - t_{n-2})} = \frac{1}{2h^2}(t - t_{n-1})(t - t_{n-2}).$$

Pour obtenir le coefficient de f_n, on intègre ce polynôme entre t_n et t_{n+1}, en utilisant le changement de variable $t = t_n + sh$,

$$\int_{t_n}^{t_{n+1}} L_1(t)\,dt = \frac{h}{2h^2}\int_0^1 (sh + h)(sh + 2h)\,ds = \frac{h}{2}\int_0^1 (s^2 + 3s + 2)\,ds = \frac{23}{12}h.$$

Procédant de même pour les deux autres points d'interpolation, on obtient le schéma AB3

$$\text{(AB3)} \qquad y_{n+1} = y_n + h\left(\frac{23}{12}f_n - \frac{4}{3}f_{n-1} + \frac{5}{12}f_{n-2}\right). \tag{6.3.4}$$

Proposition 6.3.3 *Le schéma* AB3 *est d'ordre trois.*

Démonstration. Comme toujours pour ces questions d'ordre de schémas, [...],

$$\begin{aligned}
\varepsilon_n &= y(t_{n+1}) - y(t_n) - \frac{h}{12}(23f(t_n, y(t_n)) - 16f(t_{n-1}, y(t_{n-1})) + 5f(t_{n-2}, y(t_{n-2})) \\
&= y(t_n + h) - y(t_n) - \frac{h}{12}\big(23y'(t_n) - 16y'(t_n - h) + 5y'(t_n - 2h)\big) \\
&= hy'(t_n) + \frac{h^2}{2}y''(t_n) + \frac{h^3}{6}y'''(t_n) + O(h^4) \\
&\qquad - \frac{h}{12}\Big[23y'(t_n) - 16\Big(y'(t_n) - hy''(t_n) + \frac{h^2}{2}y'''(t_n) + O(h^3)\Big) \\
&\qquad\qquad + 5\big(y'(t_n) - 2hy''(t_n) + 2h^2y'''(t_n) + O(h^3)\big)\Big] \\
&= O(h^4).
\end{aligned}$$

On imagine bien que [...] d'ordre inférieur à 3... ◊

4. Le schéma implicite Adams-Moulton sur 3 points est défini à partir de l'interpolation $P_{n,2}(t_{n-j}) = f_{n-j} = f(t_{n-j}, y_{n-j})$ pour $j = -1, 0, 1$. Par la même méthode, on obtient le schéma

$$\text{(AM3)} \qquad y_{n+1} = y_n + h\left(\frac{5}{12}f(t_{n+1}, y_{n+1}) + \frac{2}{3}f(t_n, y_n) - \frac{1}{12}f(t_{n-1}, y_{n-1})\right).$$

Ce schéma implicite fonctionne si l'équation non linéaire $\varphi(y_{n+1}) = y_{n+1}$ avec

$$\varphi(z) = y_n + \frac{h}{12}\Big(5f(t_{n+1}, z) + 8f(t_n, y_n) - f(t_{n-1}, y_{n-1})\Big)$$

a une solution à chaque pas de temps. Une condition suffisante pour cela est que φ soit strictement contractante, ce qui est clairement le cas si f est $L-$lipschitzienne et si $h < \frac{12}{5L}$.

Proposition 6.3.4 *Le schéma* AM3 *est d'ordre trois.*

Démonstration. Idem que pour AB3. ◊

On peut définir des méthodes d'Adams d'ordre arbitrairement élevé. Leur avantage est leur simplicité et leur économie de calcul. Leur désavantage est qu'il faut les initialiser à un ordre supérieur ou égal à leur ordre, sinon on perd toute la précision attendue. Cela ne peut se faire qu'avec des méthodes à un pas d'ordre élevé, comme les méthodes de Runge-Kutta qu'on vient de voir (et qui sont à un pas, mais compliqués). Une stratégie possible lorsque l'on envisage des calculs de grande envergure, comme des calculs en astronomie par exemple, ou de dynamique moléculaire, est d'initialiser une méthode d'Adams à l'aide d'une méthode de Runge-Kutta. Supposons que l'on prévoie d'utiliser une méthode d'Adams à une dizaine de pas, pour avoir un ordre élevé et donc une très bonne précision. On a besoin pour lancer la méthode d'Adams de la dizaine de premiers pas avec une précision égale ou supérieure à celle attendue globalement. Il suffit de calculer cette dizaine de premiers pas avec une méthode de Runge-Kutta ou de Taylor du même ordre que celle d'Adams prévue, puis d'embrayer sur le milliard d'itérations suivantes par la méthode d'Adams.

Chapitre 7

Intégration numérique

L'objectif de ce chapitre est d'introduire et d'analyser des méthodes pour calculer de manière approchée l'intégrale au sens de Riemann

$$I := \int_a^b f(x)dx,$$

où a et b sont deux réels et f une fonction (qu'on supposera continue). On présentera deux classes de méthodes : de type interpolation de Lagrange et de type Gauss.

7.1 Intégration de type interpolation de Lagrange

Les méthodes de type interpolation consistent à approcher la fonction f par une fonction plus simple dont l'intégrale est explicite, typiquement des polynômes. On se donne $[a, b]$ un intervalle et on note $(x_i)_{0\leq i\leq n}$ une subdivision ordonnée à $n+1$ points (non nécessairement uniforme de $[a, b]$) : $a = x_0 < x_1 < \cdots < x_n = b$.

Définition 7.1.1 *Une formule d'intégration à $n+1$ points pour calculer une valeur approchée de I est une relation de la forme*

$$J_n(f) := \sum_{k=0}^{n} A_{nk} f(x_k),$$

où les coefficients A_{nk} ne dépendent pas de f.

7.1.1 Interpolation de Lagrange

Le polynôme d'interpolation de Lagrange (déjà rencontré pour définir les méthodes d'Adams, cf. Théorème 6.3.1) permet de définir de tels coefficients A_{nk}. On rappelle ce théorème avec les notations de ce chapitre.

Théorème 7.1.2 *Soit $f : [a, b] \to \mathbb{R}$ une fonction. Pour tout choix de $n+1$ points $a = x_0 < x_1 < \cdots < x_n = b$ distincts et tout jeu de $n+1$ valeurs $f(x_j)$, il existe un unique polynôme P_f de degré inférieur ou égal à n qui vérifie*

$$\forall j \in \{0, 1, \ldots, n\}, \quad P_f(x_j) = f(x_j).$$

Le polynôme P_f est appelé le polynôme d'interpolation de Lagrange *de f aux points x_j, $0 \leq j \leq n$. Il s'écrit :*

$$P_f(x) = \sum_{j=0}^{n} f(x_j) \prod_{k \neq j} \left(\frac{x - x_k}{x_j - x_k} \right). \tag{7.1.1}$$

Démonstration. Ce résultat assure l'existence et l'unicité du polynôme interpolateur de Lagrange. Commençons par l'existence. La formule (7.1.1) définit bien un polynôme de degré n comme somme de n produits de n monômes de degré 1 (les points $(x_k)_k$ étant tous distincts, les dénominateurs sont non nuls). On remarque tout de suite que P_f satisfait bien $P_f(x_k) = f(x_k)$ par construction. Il reste à montrer l'unicité d'un tel polynôme interpolateur. Pour tout $0 \leq j \leq n$, on pose $p_j : x \mapsto \prod_{k \neq j} (\frac{x - x_k}{x_j - x_k})$, qui définit un polynôme de degré n comme produit de n monômes de degré 1 (les points $(x_k)_k$ étant tous distincts, les dénominateurs sont non nuls). Montrons qu'ils forment une famille libre. Remarquons d'abord que $p_j(x_k) = 1$ si $k = j$ et $p_j(x_k) = 0$ si $k \neq j$. Considérons alors un jeu de $n+1$ coefficients $\alpha_0, \ldots, \alpha_n$ tel que $p := \sum_{j=0}^{n} \alpha_j p_j = 0$ et montrons que $\alpha_j = 0$ pour tout $0 \leq j \leq n$, ce qui assurera que la famille est libre. En évaluant p au point x_k, on obtient bien $0 = p(x_k) = \sum_{j=0}^{n} \alpha_j p_j(x_k) = \alpha_k$. La famille de polynômes $(p_j)_{0 \leq j \leq n}$ est libre, de cardinal $n+1$ et les polynômes sont de degré n : elle forment donc une base des polynômes de degré inférieur ou égal à n. Cela implique l'unicité du polynôme interpolateur. En effet, soit f une fonction et P_1 et P_2 deux polynômes interpolateurs de Lagrange associés à f. $P_1 - P_2$ est un polynôme de degré n, qu'on peut écrire dans la base des p_j comme $P_1 - P_2 = \sum_{j=0}^{n} \alpha_j p_j$. Par définition, pour tout $0 \leq k \leq n$, $P_1(x_k) - P_2(x_k) = f(x_k) - f(x_k) = 0$ et donc $\alpha_k = 0$, ce qui implique $P_1 = P_2$ et donc l'unicité du polynôme interpolateur de Lagrange. ◊

Le méthode d'intégration numérique s'écrit alors

$$J_n(f) := \sum_{k=0}^{n} A_{nk} f(x_k), \text{ avec } A_{nk} = \int_a^b \prod_{k \neq j} \left(\frac{x - x_k}{x_j - x_k} \right) dx, \tag{7.1.2}$$

où les termes A_{nk} peuvent être calculés explicitement.

Une fois la méthode bien définie, il convient d'en analyser le comportement. La méthode est-elle convergente, a-t-on des estimations d'erreur ? En analyse, on sait bien que comparer une fonction à un polynôme est raisonnable si cette fonction est régulière. C'est ce que quantifie la proposition suivante en terme d'intégrales.

Proposition 7.1.3 *Soit $f : [a,b] \to \mathbb{R}$ une fonction C^{n+1}, $a = x_0 < x_1 < \cdots < x_n = b$ un choix de $n+1$ points distincts et P_f le polynôme d'interpolation de Lagrange associé. Il existe une fonction $\zeta : [a,b] \to [a,b]$ telle que pour tout $x \in [a,b]$ on a*

$$f(x) = P_f(x) + \frac{1}{(n+1)!} f^{(n+1)}(\zeta_x) \prod_{j=0}^{n} (x - x_j).$$

En particulier,

$$I - J_n(f) = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\zeta_x) \prod_{j=0}^{n} (x - x_j) dx$$

et

$$|I - J_n(f)| \leq \frac{\sup_{[a,b]} |f^{(n+1)}|}{(n+1)!} \int_a^b \prod_{j=0}^{n} |x - x_j| dx$$

Pour démontrer cette proposition, on utilise un corollaire du théorème de Rolle. On rappelle que le théorème de Rolle assure que si $g : [c, d] \to \mathbb{R}$ est une fonction continue sur le segment $[c, d]$ dérivable sur $]a, b[$ telle que $f(c) = f(d)$, alors il existe $e \in]c, d[$ tel que $f'(e) = 0$.

Corollaire 7.1.4 . *Soit $g \in C^{n+1}([a, b])$. Si g possède au moins $n + 2$ zéros distincts sur $[a, b]$ alors $g^{(n+1)}$ (dérivée d'ordre $n + 1$) a au moins un zéro sur $[a, b]$.*

Démonstration. Tout d'abord, on remarque que le théorème de Rolle implique le résultat suivant : soit $g : [a, b] \to \mathbb{R}$ une fonction dérivable qui admet au moins $n + 2$ zéros distincts sur $[a, b]$, alors f' admet au moins $n + 1$ zéros distincts sur $[a, b]$. Il suffit en effet d'appliquer le théorème de Rolle entre deux zéros consécutifs de g. Le corollaire suit par récurrence. ◊

On peut passer à la preuve de la Proposition 7.1.3.

Démonstration. On commence par montrer l'existence d'une fonction $\zeta : [a, b] \to [a, b]$ telle que pour tout $x \in [a, b]$ on a

$$f(x) = P_f(x) + \frac{1}{(n+1)!} f^{(n+1)}(\zeta_x) q(x),$$

où $q(x) = \prod_{j=0}^{n}(x - x_j)$ (polynôme de degré $n + 1$ qui s'annule en les points $(x_j)_{0 \leq x_j \leq n}$ uniquement). Si $x = x_j$ pour $0 \leq j \leq n$, on pose simplement $\zeta_{x_j} = x_j$. Supposons $x \notin \{x_0, \ldots, x_n\}$, auquel cas $q(x) \neq 0$ et définissons la fonction $w : [a, b] \to \mathbb{R}, y \mapsto w(y)$ par

$$w(y) = f(y) - P_f(y) - \frac{q(y)}{q(x)}(f(x) - P_f(x)).$$

Tout comme f, w est C^{n+1}. De plus w s'annule en les $n + 2$ points $\{x, x_0, \ldots, x_n\}$. Par conséquent, d'après le corollaire 7.1.4, il existe au moins un point $\zeta_x \in [a, b]$ tel que $w^{(n+1)}(\zeta_x) = 0$. Comme P_f est de degré n, $P_f^{(n+1)} \equiv 0$. Par définition, $q(y) - y^{n+1}$ est un polynôme de degré n, sa dérivée $(n + 1)$-ième est nulle et donc $q^{(n+1)} - (n + 1)! \equiv 0$. On en déduit $0 = w^{(n+1)}(\zeta_x) = \frac{(n+1)!}{q(x)}(f(x) - P_f(x))$, d'où le résultat.

En intégrant cette égalité entre a et b et en utilisant la définition (7.1.2) de $J_n(f)$ on obtient bien la formule

$$I - J_n(f) = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\zeta_x) \prod_{j=0}^{n}(x - x_j) dx.$$

En prenant ensuite la valeur absolue de l'intégrale et en bornant $|f^{(n+1)}(\zeta_x)|$ par le supremum de (la fonction continue) $|f^{(n+1)}|$, on conclut que

$$|I - J_n(f)| \leq \frac{\sup_{[a,b]} |f^{(n+1)}|}{(n+1)!} \int_a^b \prod_{j=0}^{n} |x - x_j| dx.$$

◊

7.1.2 Formules de Newton-Côtes

Soit $n \geq 1$. Les formules de Newton-Côtes (fermées) reviennent à choisir une subdivision uniforme de $[a,b]$ en $n+1$ points $x_i = a + \frac{i}{n}(b-a)$ pour $0 \leq i \leq n$, de pas $h := \frac{b-a}{n}$. Dans ce cas, en appliquant la proposition 7.1.3, on obtient l'estimation d'erreur suivante

Corollaire 7.1.5 *Si f est C^{n+1} sur $[a,b]$ alors*

$$|I - J_n(f)| \;\leq\; 2h^{n+1} \sup_{[a,b]} |f^{(n+1)}|.$$

Démonstration. Par le changement de variables $x \rightsquigarrow t = \frac{x-a}{h} = (x-a)\frac{b-a}{n}$, on obtient

$$|I - J_n(f)| \;\leq\; \frac{\sup_{[a,b]} |f^{(n+1)}|}{(n+1)!} \left(\frac{b-a}{n}\right)^{n+1} \int_0^n \prod_{j=0}^{n} |t-j| dt$$

et il reste à estimer l'intégrale. On coupe l'intégrale en n morceaux de la forme $[i, i+1]$. Sur cet intervalle, on a la borne

$$\begin{aligned} \int_i^{i+1} \prod_{j=0}^{n} |t-j| dt \;&\leq\; \prod_{j=0}^{n} (|i-j|+1) = \prod_{j=1}^{i} (i+1-j) \times \prod_{j=i+1}^{n} (j+1-i) \\ &\leq\; \Big(\prod_{k=1}^{i} k\Big)\Big(\prod_{k=1}^{n+1-i} j\Big) \;=\; i!(n+1-i)!. \end{aligned}$$

Ainsi,

$$\int_0^n \prod_{j=0}^{n} |t-j| dt \;\leq\; \sum_{i=0}^{n-1} i!(n+1-i)! \leq 2\,(n+1)!$$

(après un petit peu de travail), d'où la borne. ◊

L'analyse ci-dessus donne de bons résultats, mais la preuve n'utilise pas que les points sont équidistants (vue qu'elle est valable pour un choix général de points). Les estimations suivantes sont plus fines.

Théorème 7.1.6 *On suppose que $n \geq 1$ et on rappelle que $h = \frac{b-a}{n}$. Si f est C^{n+2} sur $[a,b]$ et n est pair, alors on a pour un $\xi \in [a,b]$*

$$I - J_n(f) \;=\; -h^{n+3} \frac{f^{(n+2)}(\xi)}{(n+2)!} \int_0^n s^2(s-1)\dots(s-n) ds.$$

Si f est C^{n+1} sur $[a,b]$ et n est impair, alors on a pour un $\xi \in [a,b]$

$$I - J_n(f) \;=\; -h^{n+2} \frac{f^{(n+1)}(\xi)}{(n+2)!} \int_0^n s(s-1)\dots(s-n) ds.$$

Ce théorème se démontre en utilisant de manière explicite que le polynôme de Lagrange interpole f aux points x_i et que les points sont équidistants.

Définition 7.1.7 *On dit qu'une méthode d'intégration est d'ordre n si elle est exacte pour des polynômes de degré n.*

On conclut ce paragraphe par quelques formules concrètes et en fait bien connues. Avant cela, on donne la définition de l'ordre d'une méthode d'intégration.

— Cas $n = 0$: Le seul point est soit a ou b, auquel cas on obtient

$$J_0(f) = (b-a)f(a) \text{ ou } J_0(f) = (b-a)f(b),$$

c'est-à-dire la formule des rectangles, qui est d'ordre 0. On a alors par la proposition 7.1.3

$$|I - J_0(f)| \leq \frac{1}{2}(b-a)^2 \sup_{[a,b]} |f'|.$$

— Cas $n = 1$: Les deux points sont a et b et on obtient

$$J_1(f) = \frac{(b-a)}{2}(f(a) + f(b)),$$

c'est-à-dire la formule des trapèzes, qui est d'ordre 1. On a alors par le théorème 7.1.6

$$|I - J_1(f)| \leq (b-a)^3 \Big(\frac{\sup_{[a,b]} |f^{(2)}|}{3!}\Big)\frac{1}{6} = \frac{1}{36}(b-a)^3 \sup_{[a,b]} |f^{(2)}|.$$

— Cas $n = 2$: Les trois points sont $\{a, \frac{a+b}{2}, b\}$ et on obtient (il faut savoir intégrer un polynôme de degré 2 !)

$$J_2(f) = \frac{(b-a)}{6}\Big(f(a) + 4f\big(\frac{a+b}{2}\big) + f(b)\Big),$$

c'est-à-dire la formule de Simpson, qui est d'ordre 3. On a alors par le théorème 7.1.6

$$|I - J_2(f)| \leq \Big(\frac{b-a}{2}\Big)^5 \Big(\frac{\sup_{[a,b]} |f^{(4)}|}{4!}\Big)\frac{4}{15} = \frac{1}{2880}(b-a)^5 \sup_{[a,b]} |f^{(4)}|.$$

Dans les estimations du théorème 7.1.6, il y a deux raisons pour lesquelles l'estimation semble s'améliorer quand n augmente (si on néglige pour cette discussion que $\sup_{[a,b]} |f^{(n)}|$ dépend aussi de n) : comme $h^{n+k} = \left(\frac{b-a}{n}\right)^{n+k}$ (pour $k = 2$ ou 3), cette quantité est décroissante par rapport à n car on divise $b - a$ par n et parce qu'on prend une puissance $n + k$ (ce qui est bien dès que $\frac{b-a}{n}$ est plus petit que 1). On peut en fait séparer ces deux effets en subdivisant d'abord l'intervalle $[a, b]$ en sous-intervalles de taille plus petite et en choisissant l'ordre de la méthode. C'est ce qu'on appelle des formules composites.

7.1.3 Formules composites

Quand le degré n du polynôme d'interpolation devient grand, le polynôme obtenu oscille beaucoup (même si la fonction qu'il interpole oscille peu), ce qui fait qu'on n'utilise que rarement des polynômes de degré élevé pour l'approximation d'intégrales. En revanche, pour traiter des grands intervalles, on utilise une approche locale : on découpe d'abord l'intervalles $[a, b]$ et on applique des formules d'ordre faible sur les sous-intervalles obtenus.

7.2 Intégration de type interpolation de Gauss

Jusqu'à présent, dans les méthodes d'intégration basées sur l'interpolation par des polynômes, on a considéré une subdivision donnée $a = x_0 < x_1 < \cdots < x_n = b$ de l'intervalle $[a, b]$ (même uniforme pour les méthodes de Newton-Côtes) et cherché les valeurs optimales des coefficients A_{nk} de manière à minimiser l'erreur entre $\int_a^b f$ et l'approximation

$$J_n(f) := \sum_{k=0}^{n} A_{nk} f(x_k),$$

pour des fonctions f générales (il est primordial que A_{nk} ne dépende pas de f !). Dans ce paragraphe, on se permet non seulement de choisir les coefficients A_{nk} mais aussi de choisir les points x_j de la subdivision. Alors que le théorème 7.1.6 montre que les méthodes de Newton-Côtes à $n+1$ points sont au plus d'ordre $n+2$ (pour n impair) et $n+3$ (pour n pair), le théoèrem 7.2.5 montre que la méthodes de Gauss-Legrendre à $n+1$ points est d'ordre $2n+1$.

7.2.1 Polynômes de Legendre

On commence par définir les polynômes de Legendre, dont les racines donneront notre choix de subdivision dans le cas $[a, b] = [-1, 1]$.

Définition 7.2.1 *On appelle polynôme de Legendre d'ordre $k \in \mathbb{N}$ le polynôme g_k défini par*

$$g_k(t) = \frac{d^k}{dt^k}(t^2 - 1)^k.$$

On a $g_0 \equiv 1$, $g_1(t) = 2t$, $g_2(t) = 12t^2 - 4$, $g_3(t) = 120t^3 - 72t$, etc. Le terme de plus haut degré de g_k est $\frac{(2k)!}{k!}t^k$. Le polynôme g_k est donc de degré k et $\{g_0, g_1, \ldots, g_n\}$ forme une base de l'espace vectoriel des polynômes de degré $n \in \mathbb{N}$.

Notre objectif est de montrer que g_k admet k racines distinctes dans $]-1, 1[$ si $k \geq 1$. On commence par démontrer un résultat fondamental sur les racines de polynômes.

Lemme 7.2.2 *Pour tous entiers $k \geq 1$ et $0 \leq i \leq k-1$, le polynôme $t \mapsto \frac{d^i}{dt^i}(t^2-1)^k$ admet 1 et -1 comme racines.*

Démonstration. On écrit $h_k(t) = (t^2-1)^k = (t-1)^k(t+1)^k$. Par récurrence, on obtient que pour tout $0 \leq i \leq k-1$, $(t-1)^{k-i}$ et $(t+1)^{k-i}$ divisent $h_k^{(i)}(t)$ et donc que 1 et -1 sont des racines de h_k. ◊

Ceci permet de démontrer la propriété fondamentale des polynômes de Legendre.

Théorème 7.2.3 *Les polynômes de Legendre sont orthogonaux pour la mesure de Lebesgue sur $[-1, 1]$ au sens où pour tous $i \neq k$ on a*

$$\langle g_i, g_k \rangle := \int_{-1}^{1} g_i(t) g_k(t) dt = 0.$$

En particulier, $\int_{-1}^{1} g_0(t)dt = 2$ et pour tout $i \geq 1$, $\int_{-1}^{1} g_i(t)dt = 0$. Par ailleurs, pour tout $n \geq 0$, g_{n+1} est orthogonal aux polynômes de degré inférieur ou égal à n.

Démonstration. On peut supposer $0 \leq i < k$. On intègre par parties une première fois

$$\begin{aligned}\int_{-1}^{1} g_i(t)g_k(t)dt &= \int_{-1}^{1} \frac{d^i}{dt^i}(t^2-1)^i \frac{d^k}{dt^k}(t^2-1)^k dt \\ &= \left[\frac{d^i}{dt^i}(t^2-1)^i \frac{d^{k-1}}{dt^{k-1}}(t^2-1)^k\right]_{-1}^{1} - \int_{-1}^{1} \frac{d^{i+1}}{dt^{i+1}}(t^2-1)^i \frac{d^{k-1}}{dt^{k-1}}(t^2-1)^k dt.\end{aligned}$$

Par le lemme 7.2.2, le premier terme est nul et on obtient alors

$$\int_{-1}^{1} g_i(t)g_k(t)dt = -\int_{-1}^{1} \frac{d^{i+1}}{dt^{i+1}}(t^2-1)^i \frac{d^{k-1}}{dt^{k-1}}(t^2-1)^k dt.$$

On peut intégrer encore i fois par parties (toujours en utilisant le lemme 7.2.2 pour montrer que les termes de bord s'annulent) et on obtient

$$\int_{-1}^{1} g_i(t)g_k(t)dt = (-1)^{i+1}\int_{-1}^{1} \frac{d^{2i+1}}{dt^{2i+1}}(t^2-1)^i \frac{d^{k-i-1}}{dt^{k-i-1}}(t^2-1)^k dt,$$

terme qui est nul car la dérivée d'ordre $2i+1$ d'un polynôme de degré $2i$ est identiquement nulle.

Comme $g_0 \equiv 1$, pour tout $i \geq 1$, $\int_{-1}^{1} g_i(t)dt = \int_{-1}^{1} g_i(t)g_0(t)dt = \langle g_0, g_i\rangle = 0$.

Enfin, comme $\{g_0, \ldots, g_n\}$ est une base de l'espace $\mathcal{P}_n$ des polynômes de degré inférieur ou égal à n, pour tout $p \in \mathcal{P}_n$, il existe $\alpha_0, \ldots, \alpha_n \in \mathbb{R}$ tels que $p = \sum_{i=0}^{n} \alpha_i g_i$ et donc (par linéarité de l'intégrale)

$$\langle p, g_{n+1}\rangle = \sum_{i=0}^{n} \alpha_i \langle g_i, g_{n+1}\rangle = 0.$$

◊

On conclut l'étude des polynômes de Legendre par la propriété souhaitée sur leurs racines.

Théorème 7.2.4 *Pour tout $n \geq 1$, le polynôme de Legendre g_n d'ordre n possède n racines distinctes de multiplicité 1 dans $]-1,1[$.*

Démonstration. D'après le théorème 7.2.3, $\int_1^1 g_n(t)dt = 0$, ce qui implique (comme g_n n'est pas identiquement nul) que g_n change au moins une fois de signe sur $]-1,1[$. On note $q \geq 1$ le nombre de racines g_n de multiplicité impaire (donc précisément les points où g_n change de signe sur $]-1,1[$) et on note $\xi_1, \ldots, \xi_q$ et $\beta_1, \ldots, \beta_q$ ces racines et leur multiplicité (telles que $\beta := \sum_{i=1}^{q} \beta_i \leq n$). En particulier, il existe un polynôme σ de degré $n-\beta$ et qui ne change pas de signe sur $]-1,1[$ tel que $g_n(t) = \sigma(t)\prod_{i=1}^{q}(t-x_i)^{\beta_i}$. On considère alors le polynôme $\pi(t) = \prod_{i=1}^{q}(t-x_i)$. Par construction, le polynôme $t \mapsto g_n(t)\pi(t) = \sigma(t)\prod_{i=1}^{q}(t-x_i)^{\beta_i+1}$ ne change pas de signe sur $]-1,1[$ car σ ne change pas de signe et $\beta_i + 1$ est pair pour tout $1 \leq i \leq q$. Comme ces polynômes ne sont pas identiquement nuls, cela implique que

$$\int_{-1} g_n(t)\pi(t)dt \neq 0.$$

Or, d'après le théorème 7.2.3, ceci ne peut être vrai que si $q = \deg\pi \geq n$. Comme $q \leq n$, cela implique que $q = \deg\pi = n$ et donc que g_n a exactement $q = n$ racines distinctes de multiplicité 1 dans $]-1,1[$ (en effet $q \leq \sum_{i=1}^{q} \beta_i \leq n$ implique $\beta_i = 1$ pour tout $1 \leq i \leq q$ si $q = n$). ◊

7.2.2 Méthode de Gauss-Legendre

On considère $[a, b] = [-1, 1]$. Pour tout $n \geq 1$, on note $\xi_1, \ldots \xi_n$ les n racines distinctes de g_n dans $]-1, 1[$. Ayant fixé ces racines, on cherche les poids $(w_i)_{1\leq i\leq n}$ tels que l'approximation

$$J_n(f) = \sum_{i=1}^{n} w_i f(\xi_i)$$

de l'intégrale $\int_{-1}^{1} f(t)dt$ soit exacte pour les polynômes de plus haut degré possible. Le choix de ces points est donné par le théorème suivant.

Théorème 7.2.5 *Le système linéaire*

$$\begin{pmatrix} g_0(\xi_1) & g_0(\xi_2) & \ldots & g_0(\xi_n) \\ g_1(\xi_1) & g_1(\xi_2) & \ldots & g_1(\xi_n) \\ \vdots & \vdots & \vdots & \vdots \\ g_{n-1}(\xi_1) & g_{n-1}(\xi_2) & \ldots & g_{n-1}(\xi_n) \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

admet une unique solution. Si les poids $\{w_i\}_{1\leq i\leq n}$ sont solutions de ce système linéaire alors pour tout polynôme $p \in \mathcal{P}_{2n-1}$ (degré au plus $2n-1$), on a l'égalité

$$\int_{-1}^{1} p(t)dt = \sum_{i=1}^{n} w_i p(\xi_i). \tag{7.2.1}$$

En particulier, le théorème montre que la formule de Gauss-Legendre à n points $J_n(f) = \sum_{i=1}^{n} w_i f(\xi_i)$ sur l'intervalle $[-1, 1]$ est d'ordre $2n-1$.

Remarque 7.2.1 L'erreur par la formule de Gauss-Legendre prend la forme : il existe $\eta \in]-1, 1[$ tel que pour tout $f \in C^{2n}([-1, 1])$ on ait

$$\int_{-1}^{1} f(t)dt - J_n(f) = \frac{2^{2n+1}(n!)^4}{(2n+1)((2n)!)^3} f^{(2n)}(\eta).$$

Avant de démontrer ce théorème, mentionnons deux aspects d'ordre pratique :

— Les racines des polynômes de Legendre ne sont pas explicites. Cependant, on peut les approcher avec une précision aussi grande que voulue grâce à l'algorithme de Newton.

— Pour intégrer une fonction f sur un intervalle $[a, b]$ général, il suffit de se ramener à $[-1, 1]$ par changement de variables. En effet, la fonction $g : [-1, 1], t \mapsto f\Big(- at + \frac{1}{2}(b + a)(t + 1)\Big)$ satisfait

$$\int_a^b f(t)dt \;=\; \frac{b-a}{2} \int_{-1}^{1} g(t)dt,$$

et il suffit donc d'appliquer la méthode de Gauss-Legendre à g.

Démonstration. On commence par montrer que le système admet une unique solution. Montrons pour cela que les lignes de la matrice du système sont linéairement indépendantes. Soient $c_0, c_1, \ldots, c_{n-1}$ des réels tels que pour tout $1 \leq i \leq n$,

$$\sum_{k=0}^{n-1} c_k g_k(\xi_i) = 0.$$

Comme les n points ξ_i sont distincts, cela implique que le polynôme $t \mapsto c_k g_k(t)$ admet au moins n racines distinctes sur $]-1, 1[$. Comme ce polynôme est de degré au plus $n-1$, il est identiquement nul. Comme les g_k sont linéairement indépendants (il sont de degré k), ceci implique que $c_k = 0$ pour tout $1 \leq k \leq n$. Les lignes de la matrice du système linéaire sont donc indépendantes, la matrice est inversible et la solution du système existe et est unique. On note $w_1, \ldots, w_n$ la solution du système.

On vérifie d'abord que (7.2.1) est vrai pour les polynômes de degré au plus $n-1$. Comme $\{g_0, \ldots, g_{n-1}\}$ est une base de $\mathcal{P}_{n-1}$ et par linéarité de l'intégrale, il suffit de vérifier cette relation sur les n premiers polynômes de Legendre. Par définition même du système et par le théorème 7.2.3 on a

$$\begin{aligned}
\sum_{i=1}^{n} w_i g_0(\xi_i) &= 2 = \int_1^{-1} g_0(t)dt, \\
\sum_{i=1}^{n} w_i g_k(\xi_i) &= 0 = \int_1^{-1} g_k(t)dt, \quad 0 < k \leq n-1.
\end{aligned}$$

Soit maintenant un polynôme $p \in \mathcal{P}_{2n-1}$. La division de p par g_n donne $p(t) = g_n(t)q(t) + r(t)$ avec $\deg(q) \leq n-1$ et $\deg(r) \leq n-1$ car $\deg(p) \leq 2n-1$ et $\deg(g_n) = n$. D'une part, en utilisant que g_n est orthogonal à $\mathcal{P}_{n-1}$ au sens du théorème 7.2.3, on a

$$\begin{aligned}
\int_{-1}^{1} p(t)dt &= \int_{-1}^{1} g_n(t)q(t)dt + \int_{-1}^{1} r(t)dt \\
&= \int_{-1}^{1} r(t)dt.
\end{aligned}$$

D'autre part, comme les ξ_i sont les racines de g_n on a d'autre part

$$\begin{aligned}
J_n(p) &= \sum_{i=1}^{n} w_i p(\xi_i) = \sum_{i=1}^{n} w_i \Big(g_n(\xi_i)q(\xi_i) + r(\xi_i) \Big) \\
&= \sum_{i=1}^{n} w_i r(\xi_i).
\end{aligned}$$

En combinant ces deux égalités avec (7.2.1) appliqué à r qui est de degré au plus $n-1$, on obtient

$$J_n(p) = \sum_{i=1}^{n} w_i r(\xi_i) = \int_{-1}^{1} r(t)dt = \int_{-1}^{1} p(t)dt,$$

ce qui conclut la démonstration. ◊

7.2.3 Formules composites

Tout comme les méthodes de Newton-Côtes, l'intégration de Gauss-Legendre peut être faite localement. Pour traiter des grands intervalles, on utilise une approche locale : on découpe d'abord l'intervalles $[a, b]$ et on applique l'intégration de Gauss-Legendre sur les sous-intervalles obtenus.

Chapitre 8

Précision numérique

8.1 Schémas d'ordre élevé et précision numérique

Une source d'erreur dans les calculs de solutions numériques est les erreurs d'arrondi qu'un ordinateur effectue systématiquement dès qu'il calcule en virgule flottante. L'arithmétique des erreurs d'arrondi est complexe et l'on n'en parlera pas ici. On se place à un niveau de compréhension plus grossier. Quand on implémente par exemple le schéma d'Euler

$$y_{n+1} = y_n + hf(t_n, y_n)$$

sur ordinateur, on ne calcule pas les valeurs exactes de la solution approchée $(y_n)_{n=0,\dots,N}$ mais des valeurs perturbées par l'erreur d'arrondi ρ_n commise à chaque pas de temps [1]

$$y^\star_{n+1} = y^\star_n + hf(t_n, y^\star_n) + \rho_n.$$

Posons $e^\star_n = y^\star_n - y(t_n)$. On a

$$e^\star_{n+1} = e^\star_n + h\left(f(t_n, y^\star_n) - f(t_n, y(t_n))\right) - \varepsilon_n + \rho_n$$

où ε_n est l'erreur de consistance. Ici on se souvient que pour la méthode d'Euler explicite

$$|\varepsilon_n| \leq \frac{h^2}{2} \max |y''|.$$

Donc, pour une fonction f lipschitzienne de constante L, et en supposant que l'ordinateur garantisse que $|\rho_n| \leq \rho$,

$$|e^\star_{n+1}| \leq (1 + hL)|e^\star_n| + \frac{h^2}{2} \max |y''| + \rho,$$

d'où l'on déduit, en utilisant le lemme de Grönwall discret, que

$$|e^\star_n| \leq e^{nhL}|e^\star_0| + \frac{h}{2}(nhL)e^{nhL} \max |y''| + nLe^{nhL}\rho.$$

1. Il y a aussi une erreur faite a priori sur l'évaluation de la fonction f, que l'on devrait remplacer par une fonction f^* effectivement calculée. On la laisse de côté, son effet n'est pas dominant.

En particulier, on obtient pour l'erreur finale, $n = N$,

$$|e_N^\star| \leq e^{TL}|e_0^\star| + \frac{h}{2}TLe^{TL}\max|y''| + \frac{TL}{h}e^{TL}\rho.$$

Les trois termes constituant la majoration de l'erreur effective se comportent différemment quand h tend vers 0. La contribution de l'erreur à l'instant initial $e^{TL}|e_0^\star|$ est indépendante de h. L'erreur due à l'accumulation des erreurs de consistance de la méthode est un $O(h)$ et tend donc vers zéro quand h tend vers 0. Enfin les effets de l'erreur d'arrondi se cumulent et tendent vers l'infini quand h tend vers 0.

Plus précisément, on a une majoration de la forme $|e_N^\star| \leq \varphi(h)$ avec $\varphi(h) = A + Bh + \frac{C\rho}{h}$, et il apparaît donc un pas optimal (au sens de cette majoration), $h^\star = \sqrt{\frac{C}{B}}\sqrt{\rho}$, pour lequel les erreurs de consistance et d'arrondi s'équilibrent et au dessous duquel il est inutile de descendre. Ce pas optimal, ainsi que l'erreur qu'il produit, est proportionnel à la racine carrée de l'erreur d'arrondi maximale garantie (soit un nombre a priori beaucoup plus grand que cette erreur).

La majoration précédente est évidemment pessimiste puisqu'on a considéré qu'à chaque pas de temps on faisait le maximum d'erreur d'arrondi et qu'elles se cumulaient toujours, alors qu'elles pourraient se compenser. Cependant le comportement prédit est effectivement observé : quand on diminue le pas de discrétisation, une fois atteint une erreur de l'ordre de la précision machine, l'erreur globale par rapport à la solution exacte commence à ré-augmenter. La Figure 8.1 illustre ce comportement pour le schéma de Taylor d'ordre 2 et le schéma de Runge-Kutta d'ordre 4. [2] On calcule la solution approchée du problème de Cauchy

$$y'(t) = 2(y(t) - \sin t) + \cos t, \quad y(0) = 0$$

dont la solution exacte est $y(t) = \sin t$ sur l'intervalle $[0, 1]$ avec un pas de discrétisation $h = 10^{-i}$, pour $i = 1, \ldots, 6$. Les graphes représentent l'erreur au temps $t = 1$ entre la solution approchée et la solution exacte en fonction de h, dans un repère logarithmique. Le graphe de gauche correspondant au schéma de Taylor d'ordre 2 met en évidence une erreur qui décroît comme un $O(h^2)$ en partant des grandes valeurs de h (0.1 jusqu'à 10^{-6}) où elle atteint la valeur 10^{-12} proche de la précision machine. Pour les valeurs de h inférieures à 10^{-6} l'erreur est plus élevée. Le même comportement est observé pour le schéma de Runge Kutta sur le graphe de droite. Comme l'erreur due à la méthode décroît plus rapidement avec h, l'erreur machine est atteinte pour une valeur de h plus grande (ici $h = 10^{-3}$) en deçà de laquelle l'erreur globale augmente quand h continue à diminuer. Le tableau ci-dessous résume les valeurs de l'erreur obtenue en utilisant le schéma de Runge-Kutta pour $h = 10^{-i}$.

i	erreur	temps CPU
1	3.17877 e-06	0.002
2	4.18818 e-10	0.011
3	4.11893 e-14	0.11
4	1.63869 e-13	1.036
5	5.43121 e-13	19.266
6	6.41975 e-12	212.692

2. Attention, l'analyse précédente pour le schéma d'Euler d'ordre 1 ne s'applique pas sans modification. Il faut refaire les calculs d'ordres de grandeur pour les pas et erreurs optimaux.

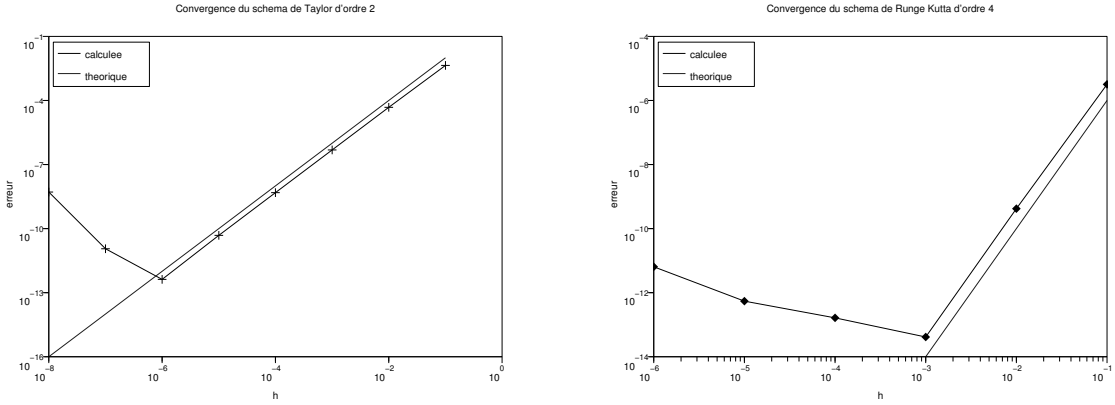


FIGURE 8.1 – Convergence de l'erreur pour les schémas de Taylor d'ordre 2 et de Runge-Kutta d'ordre 4

La conclusion de cette expérience numérique est qu'il faut d'une part connaître la précision machine de l'ordinateur qu'on utilise, d'autre part faire des tests sur des fonctions connues pour évaluer jusqu'à quelle discrétisation il est intéressant de descendre. Inutile de calculer plus pour gagner moins.

8.2 Contrôle du pas de temps

Jusqu'à présent nous avons présenté tous les schémas numériques pour un pas de discrétisation uniforme $h = t_{n+1} - t_n$. Si ce choix d'un pas constant simplifie sensiblement l'analyse numérique d'un schéma, il est en revanche loin d'être optimal en ce qui concerne les performances en temps de calcul. En effet, on a vu que l'erreur commise dépend des dérivées de la solution. Plus précisément l'erreur d'un schéma d'ordre p est en Ch^p où C est une constante dépendant de la norme infinie de certaines dérivées de la solution sur l'intervalle de calcul. Si la solution est régulière, c'est-à-dire ne varie pas beaucoup, c'est-à-dire a des dérivées de petite amplitude, on pourra utiliser un pas h plus grand et avoir la même tolérance sur l'erreur qu'avec un très petit pas pour une fonction variant beaucoup. L'idée est donc grossièrement de jouer à chaque itération sur le pas, sur la base de la constante C, qui est inconnue !

Dans cette optique, on peut faire varier le pas de discrétisation de manière adaptative en suivant l'évolution de la régularité de la solution au fur et à mesure qu'on en calcule une approximation. Mais comment faire, puisqu'on ne connaît pas C ???

Notons tout d'abord que dans le cas d'un pas variable $h_n = t_{n+1} - t_n$, toutes les définitions et tous les résultats précédemment énoncés, consistance, ordre, stabilité, convergence, estimation d'erreur, restent valables en posant $h = \max_n h_n$. On ne revient donc pas dessus. Au vu de l'analyse des différents types d'erreur de la section précédente, le seul facteur sur lequel on peut espérer jouer est l'erreur de consistance. On va négliger les autres sources d'erreur (erreur initiale, erreurs d'arrondi).

Le principe général est d'utiliser deux schémas numériques d'ordres différents $p_1 < p_2$ (dans la pratique on prendra le plus souvent $p_2 = p_1 + 1$). Le schéma d'ordre inférieur est utilisé pour le calcul de la solution approchée, et le schéma d'ordre supérieur pour l'adaptation du pas et le contrôle de l'erreur. On note y_n^1 (respectivement y_n^2) la solution

numérique calculée avec le premier (resp. deuxième) schéma au temps t_n. On a les estimations d'erreur

$$\|y^1_{n+1} - y(t_{n+1})\| \leq C_1 h^{p_1}$$
$$\|y^2_{n+1} - y(t_{n+1})\| \leq C_2 h^{p_2}.$$

On a donc

$$y^1_{n+1} - y(t_{n+1}) = y^1_{n+1} - y^2_{n+1} + y^2_{n+1} - y(t_{n+1}),$$

d'où l'on déduit en supposant h petit et la constante C_2 pas trop méchante que

$$\|y^1_{n+1} - y(t_{n+1})\| \approx \|y^1_{n+1} - y^2_{n+1}\|.$$

Supposons que l'on se soit fixé une tolérance Tol sur l'erreur à ne pas dépasser sur tout l'intervalle de temps, et que ce but soit atteint à l'itération n. On peut alors proposer la stratégie (naïve) suivante pour déterminer le prochain pas de discrétisation. On se donne $0 < \alpha < 1$.

Pas de temps adaptatif monotone

On a calculé y^1_n et y^2_n avec le dernier pas de temps h_{n-1}. On pose $h_* = h_{n-1}$.
Boucle : on calcule y^1_* et y^2_* à partir de y^1_n et y^2_n avec le pas h_*.
Si $\|y^1_* - y^2_*\| \leq Tol$, on prend $h_n = h_*$, $y^1_{n+1} = y^1_*$ et $y^2_{n+1} = y^2_*$
Sinon on prend $h_* = \alpha h_*$.

Bien sûr, il faut également imposer une borne inférieure sur le pas, qui ne peut que décroître, pour éviter les écueils mis en évidence à la section précédente ou que le schéma se bloque avec un pas descendu au zéro machine.

Cet algorithme assure plus ou moins le contrôle de l'erreur du premier schéma sur tout l'intervalle d'étude à condition que celle-ci soit rattrapable (ce qui n'est pas toujours le cas, le premier schéma peut très bien s'éloigner irrémédiablement de la solution exacte à partir d'un certain point satisfaisant la tolérance, quel que soit le pas suivant). Il n'a en fait essentiellement aucun intérêt pratique, puisqu'il implique de mener de front deux schémas d'ordres $p_1 < p_2$, pour obtenir un schéma d'ordre p_1. On aurait pu conserver le schéma d'ordre p_2, auquel on fait de plus confiance pour représenter l'erreur. De surcroît, on ne peut pas faire croître le pas à nouveau quand un pas petit n'est plus nécessaire.

On donne ci-dessous un exemple de calcul sur le problème de Cauchy

$$\begin{cases} y'(t) = -4t^3 y(t)^2, \\ y(0) = 1, \end{cases}$$

dont la solution exacte est $y(t) = \frac{1}{1+t^4}$, avec le schéma d'Euler comme schéma d'ordre 1 et celui d'Euler modifié comme schéma d'ordre 2. Le calcul est fait jusqu'à $T = 2$ avec un pas initial de 0,2, une tolérance de 0,045 et un coefficient $\alpha = 0{,}999$ (la borne inférieure sur le pas est la racine carrée du zéro machine).

Le résultat n'est pas spectaculaire (et encore, on a choisi un exemple qui marche !). On peut rendre cette stratégie plus performante en choisissant les deux schémas de telle

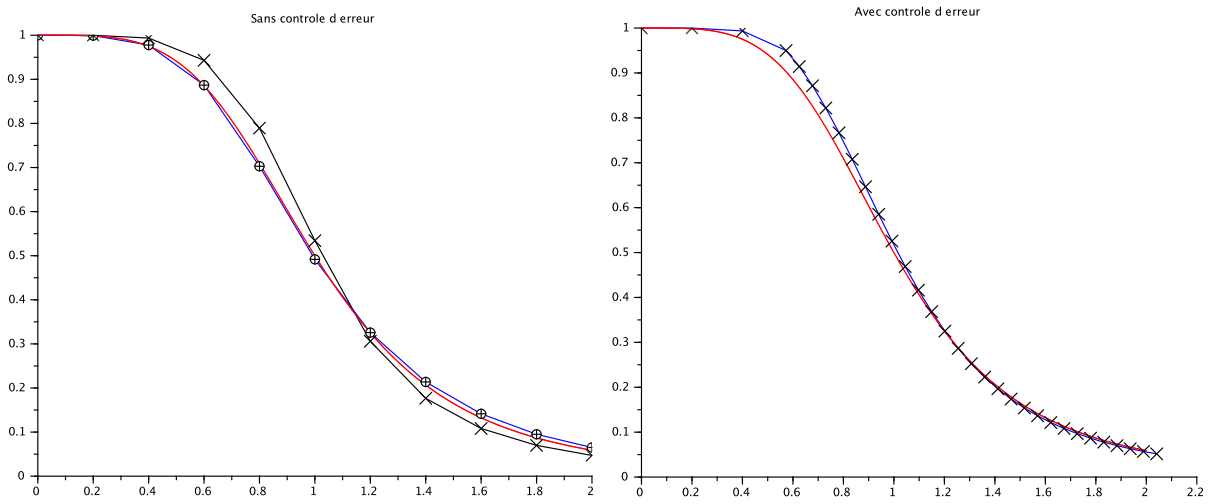


FIGURE 8.2 – Algorithme de contrôle naïf. À gauche pas constant sans contrôle : noir × Euler, bleu ⊕ Euler modifié, rouge solution exacte. À droite pas variable avec contrôle : noir × Euler, rouge solution exacte.

sorte qu'il ne coûte pas beaucoup plus cher (en temps de calcul) de mener les deux de front que le plus précis tout seul. En effet, ce qui coûte a priori cher dans un schéma, c'est l'évaluation de la fonction second membre aux différents pas de temps intermédiaires. Plusieurs algorithmes de contrôle adaptatif du pas de temps sont développés dans le cadre des schémas de Runge-Kutta, en mettant ce principe à profit.

Décrivons l'idée générale, qui est d'approcher l'erreur de consistance d'un schéma d'ordre p, donné par une fonction $\Phi(h,t,y)$, par une expression calculable à peu de frais en utilisant un autre schéma d'ordre $p+1$, donné par une fonction $\Phi^*(h,t,y)$. Évidemment, il ne faut pas calculer l'intégralité du deuxième schéma depuis l'instant initial, comme précédemment, car, disons le, c'est stupide. Et il faut si possible rentabiliser le calcul supplémentaire en le réutilisant au moins en partie lors des itérations suivantes.

Les deux erreurs de consistance sont données par

$$\begin{aligned}\varepsilon_n &= y(t_{n+1}) - y(t_n) - h_n\Phi(h_n,t_n,y(t_n)),\\ \varepsilon_n^* &= y(t_{n+1}) - y(t_n) - h_n\Phi^*(h_n,t_n,y(t_n)),\end{aligned}$$

et l'on a $\varepsilon_n = O(h^{p+1})$ et $\varepsilon_n^* = O(h^{p+2})$. On introduit un *estimateur a posteriori* de l'erreur de consistance

$$\tilde{\varepsilon}_n = h_n\big(\Phi^*(h_n,t_n,y_n) - \Phi(h_n,t_n,y_n)\big).$$

Comme on connaît y_n de l'itération précédente, cet estimateur d'erreur est calculable. De plus, le deuxième terme intervient dans le calcul de y_{n+1}, on en aura besoin de toutes façons. Le premier terme correspond à une itération du schéma d'ordre élevé à partir de la dernière valeur calculée du schéma d'ordre moins élevé. On ne calcule donc pas la solution du schéma d'ordre élevé. En quoi s'agit-il d'un estimateur de l'erreur de consistance du schéma d'ordre moins élevé ? On a

$$\begin{aligned}\varepsilon_n - \varepsilon_n^* &= h_n\big(\Phi^*(h_n,t_n,y(t_n)) - \Phi(h_n,t_n,y(t_n))\big),\\ &= \tilde{\varepsilon}_n + h_n\frac{\partial(\Phi^*-\Phi)}{\partial y}(h_n,t_n,\zeta_n)(y(t_n)-y_n),\end{aligned}$$

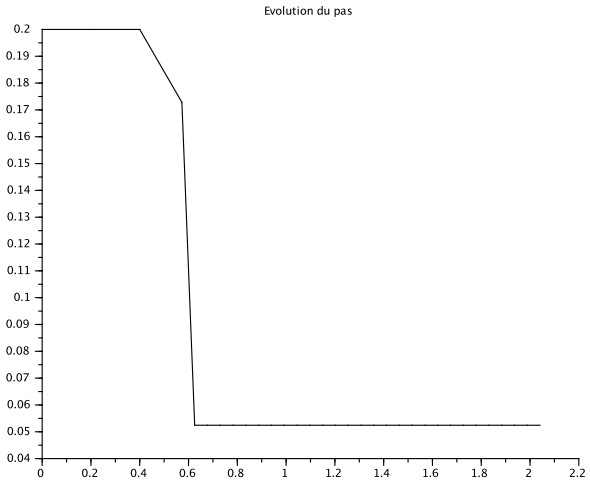


FIGURE 8.3 – Algorithme de contrôle naïf, variation du pas en fonction du temps.

pour un certain ζ_n entre $y(t_n)$ et y_n. Les deux schémas sont consistants, ce qui implique que $\Phi^*(0,t,y) - \Phi(0,t,y) = f(t,y) - f(t,y) = 0$, donc $\frac{\partial(\Phi^*-\Phi)}{\partial y}(0,t_n,\zeta_n) = 0$ et par conséquent $\frac{\partial(\Phi^*-\Phi)}{\partial y}(h_n,t_n,\zeta_n) = O(h_n)$. Comme $y(t_n) - y_n = O(h^p)$ par l'estimation d'erreur du premier schéma, on en déduit que

$$\varepsilon_n = \tilde{\varepsilon}_n + \varepsilon_n^* + O(h^{p+2}) = \tilde{\varepsilon}_n + O(h^{p+2}),$$

puisque le deuxième schéma est d'ordre $p+1$. Comme ε_n est de l'ordre de $h^{p+1} \gg h^{p+2}$, il s'ensuit que l'on peut considérer (un peu à la louche peut-être) que $\tilde{\varepsilon}_n$ en est une bonne approximation.

Prenons par exemple le schéma de Heun ou RK2, d'ordre deux, que l'on réécrit uniquement en termes des quantités à calculer

$$\begin{cases} k_1 = f(t_n, y_n), \\ k_2 = f(t_{n+1}, y_n + h_n k_1), \\ y_{n+1}^{\mathrm{rk2}} = y_n + \dfrac{h_n}{2}(k_1 + k_2), \end{cases} \tag{8.2.1}$$

et un schéma de Runge-Kutta d'ordre trois, qui commence par calculer les mêmes valeurs k_1 et k_2, puis une troisième valeur intermédiaire (on dit que ces deux schémas de Runge-Kutta sont *emboîtés*)

$$\begin{cases} k_1 = f(t_n, y_n), \\ k_2 = f(t_{n+1}, y_n + h_n k_1), \\ k_3 = f\left(t_n + \dfrac{h_n}{2}, y_n + \dfrac{h_n}{4}(k_1 + k_2)\right), \\ y_{n+1}^{\mathrm{rk3}} = y_n + \dfrac{h_n}{6}(k_1 + k_2 + 4k_3). \end{cases} \tag{8.2.2}$$

L'estimateur d'erreur est donc

$$\tilde{\varepsilon}_n = y_{n+1}^{\mathrm{rk3}} - y_{n+1}^{\mathrm{rk2}} = \frac{h_n}{3}(2k_3 - k_1 - k_2).$$

Il faut encore définir une stratégie pour contrôler l'erreur et adapter le pas dans un sens ou dans l'autre. Plusieurs choix sont possibles. On propose ici le schéma suivant

Pas de temps adaptatif RK23

Calcul de k_1, k_2, k_3 avec le pas de temps h_n
Si $h_n|2k_3 - k_1 - k_2|/3 \leq Tol$
on prend $h_{n+1} = h_n$ et $y_{n+1} = y_{n+1}^{\text{rk2}}$
Si $h_n|2k_3 - k_1 - k_2|/3 \leq Tol/10$ on multiplie h_{n+1} par 2
Sinon
on divise h_n par 2 et on recommence sans incrémenter n

On a testé cet algorithme sur le problème de Cauchy

$$\begin{cases} y'(t) = -2ty(t)^2, \\ y(0) = 1, \end{cases}$$

dont la solution exacte est $y(t) = 1/(1 + t^2)$. On commence avec un pas arbitrairement fixé à 0.05 et une tolérance fixée à 10^{-6} et on calcule la solution numérique jusqu'à $T = 10$. Le pas de temps varie de $h = 0.0125$ vers $t = 2$ jusqu'à $h = 0.2$ au temps final.

Le graphe de droite sur la Figure 8.4 montre l'évolution de l'erreur entre la solution numérique et la solution exacte (puisqu'on la connaît pour cette équation). On voit que les variations de l'erreur sont très bien corrélées avec les variations du pas de temps représentées sur le graphe de droite de la Figure 8.5. À chaque fois que l'on double le pas de temps, l'erreur commence par augmenter puis diminue jusqu'à ce que l'erreur de consistance soit inférieure à $Tol/10$ et qu'on puisse de nouveau doubler h_n.

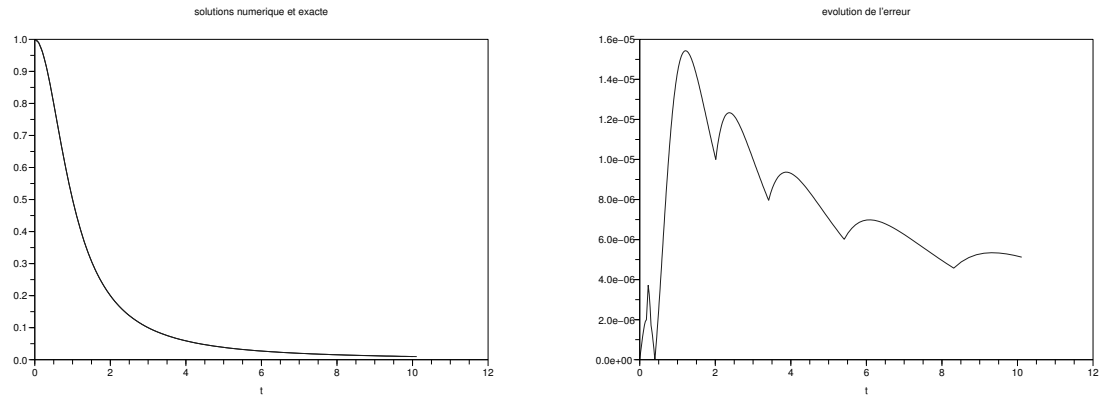


FIGURE 8.4 – Schéma RK23 avec contrôle adaptatif du pas, solution à gauche et erreur entre les solutions numériques et exactes à droite

Naturellement, toute méthode d'adaptation de pas, aussi sophistiquée soit elle, peut lamentablement échouer si l'EDO et les constantes ne coopèrent pas. C'est la question de la *robustesse* de la méthode. C'est une des raisons pour lesquelles des solveurs complexes comme ode de scilab, qui fonctionnent en boîte noire, peuvent refuser de calculer la solution. Certaines EDO sont intrinsèquement difficiles à approcher numériquement.

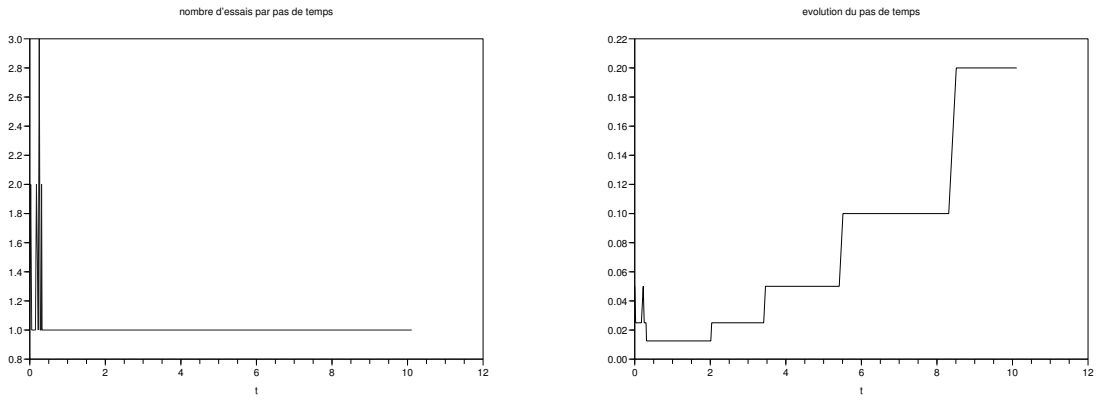


FIGURE 8.5 – Schéma RK23 avec contrôle adaptatif du pas, nombre d'essais par pas de temps à gauche et pas de temps à droite

Chapitre 9

Propriétés asymptotiques de schémas numériques

9.1 Stabilité absolue

Nous avons vu l'importance de la notion de stabilité donnée par la définition 4.1.2, qui assure que des perturbations du schéma entraînent des perturbations des solutions calculées qui restent contrôlables. Le contrôle en question peut être néanmoins difficile à réaliser en pratique dans le cas où le système comporte ce que l'on appelle une " instabilité intrinsèque ". Construisons ainsi une famille de problèmes de Cauchy à partir d'une fonction donnée g,

$$\begin{cases} y'(t) = \lambda(y(t) - g(t)) + g'(t), \\ y(0) = y_0 \end{cases} \tag{9.1.1}$$

L'équation homogène a pour solution $y(t) = e^{\lambda t}$ et $y_p(t) = g(t)$ est une solution particulière du problème non homogène. La solution du problème de Cauchy (9.1.1) est donc

$$y(t) = (y_0 - g(0))e^{\lambda t} + g(t).$$

Il est clair que si l'on ne part pas exactement de la condition initiale $g(0)$ et si $\lambda > 0$ est modérément grand, le terme exponentiel va l'emporter très rapidement sur le terme donnant la solution g. Toute méthode numérique introduisant une erreur sur la solution, cette erreur va s'amplifier de manière exponentielle avec les itérations successives. La seule issue pour traiter ce genre de problème est d'utiliser des schémas d'ordre très élevé en effectuant les calculs avec une précision également élevée, avec un pas de temps très petit. Tout cela peut être d'un coût prohibitif voire être hors de portée des ordinateurs existants.

Dans cette section, on prend $I =]0, +\infty[$ et on s'intéresse au comportement en temps long ($t_n \to +\infty$) des solutions numériques. La définition 4.1.2 n'est pas opérante, puisque placée sur un intervalle $[0, T]$ et faisant intervenir en pratique des constantes de la forme e^{CT} dans les majorations, constantes qui peuvent manifestement être énormes. Nous introduisons ici une autre notion de stabilité, qui n'a en fait pas grand chose à voir avec la précédente, si ce n'est le nom de " stabilité " malheureusement traditionnel, et qui est plus adaptée au problème qui nous intéresse ici, le comportement en temps long. [1]

1. Il y a quand même un lien entre les deux notions dans le cas des schémas à pas multiples linéaires, comme les méthodes d'Adams, que l'on n'a pas étudiés théoriquement.

Nous ne considérons dans cette section que le problème de Cauchy scalaire suivant

$$\begin{cases} y'(t) = \lambda y(t) \text{ dans } I, \\ y(0) = y_0 \neq 0, \end{cases} \tag{9.1.2}$$

dont on connaît la solution exacte $y(t) = e^{\lambda t} y_0$. Pourquoi se restreindre à une équation aussi simple ? Il se trouve que de nombreux phénomènes physiques, chimique, biologiques ou autres sont modélisés par des systèmes d'EDO. Si on linéarise un tel système au voisinage d'un instant t_0, on obtient un système linéaire à coefficients constants. Ceci justifie que l'on s'intéresse à la version matricielle de (9.1.2), $y'(t) = Ay(t)$. Mais nous avons vu qu'après diagonalisation, si celle-ci est possible, on obtient un système découplé d'équations scalaires de type (9.1.2), voir l'exemple en page **??**. Ces dernières jouent donc un rôle important.

Supposons que l'on commette une erreur sur la donnée initiale de (9.1.2) qui devient $y_0 + \varepsilon$. On obtient alors la solution $y_\varepsilon(t) = e^{\lambda t}(y_0 + \varepsilon)$. L'erreur au temps t entre les deux solutions est $\varepsilon e^{\lambda t}$. Elle tend vers $+\infty$ en valeur absolue avec t, si $\lambda > 0$. Même si ε est très petit, au bout d'un certain temps, l'erreur devient très grande et il n'y a rien à y faire. C'est la sensibilité aux conditions initiales.

Supposons maintenant qu'on parte de la donnée initiale exacte y_0 et que l'on cherche à calculer la solution de (9.1.2) avec le schéma d'Euler. On obtient une suite de valeurs $y_n = (1 + \lambda h)^n y_0$. Intéressons nous à la limite en $+\infty$ des solutions approchées. On va diminuer nos exigences et rester à un niveau qualitatif très grossier. Considérons les deux cas suivants :

1. Si $\lambda > 0$, la suite des solutions approchées y_n a le même comportement en temps long que la solution exacte, à savoir

$$\lim_{n \to +\infty} y_n = \lim_{t \to +\infty} y(t) = +\infty.$$

En effet, $n \to +\infty$ à h fixé correspond à $t_n \to +\infty$. On est donc satisfaits du point de vue qualitatif.

2. Si $\lambda < 0$, alors cette fois $\lim_{t \to +\infty} y(t) = 0$. Pour h assez petit, à savoir pour $h < -\frac{2}{\lambda}$, on a $|1 + \lambda h| < 1$ et la suite y_n tend vers 0 à h fixé, comme la solution exacte. Par contre, si $h > -\frac{2}{\lambda}$, on a $1 + \lambda h < -1$ et la suite y_n diverge au sens où $|y_n| \to +\infty$ quand $n \to +\infty$ à h fixé. Dans ce cas, la solution discrète ne reproduit pas du tout le comportement en temps long de la solution exacte. [2]

C'est à la deuxième situation que la stabilité absolue s'intéresse : la valeur $\lambda < 0$ étant donnée, comment choisir h pour que la suite des solutions approchées ait le même comportement à l'infini que la solution exacte, c'est-à-dire que $\lim_{n \to +\infty} y_n = 0$? Le cas $\lambda < 0$ correspond à des phénomènes physiques décroissants exponentiellement. Dans la suite de ce chapitre, on considère le problème plus général où $\lambda \in \mathbb{C}$ et $\Re(\lambda) < 0$ qui correspond à des phénomènes à décroissance exponentielle en module et oscillants si $\Im(\lambda) \neq 0$. Le cas le plus difficile numériquement est celui où $|\Re(\lambda)|$ est très grand, et par conséquent la solution décroît initialement très vite. C'est ce que l'on appelle les *problèmes raides*, mais il faudrait une section entière pour en parler.

Considérons maintenant un schéma à un pas (explicite ou pas) qu'on supposera donné sous la forme (4.1.1). [3] Comme l'EDO est linéaire et autonome, on suppose que la fonction F

2. Notons que ceci n'a rien à voir avec la convergence de la méthode, qui se passe sur un intervalle de temps $[0, T]$, T fixé, avec $h \to 0$.

3. Dans le cas implicite, on donne simplement un nom à la fonction implicite définissant le schéma.

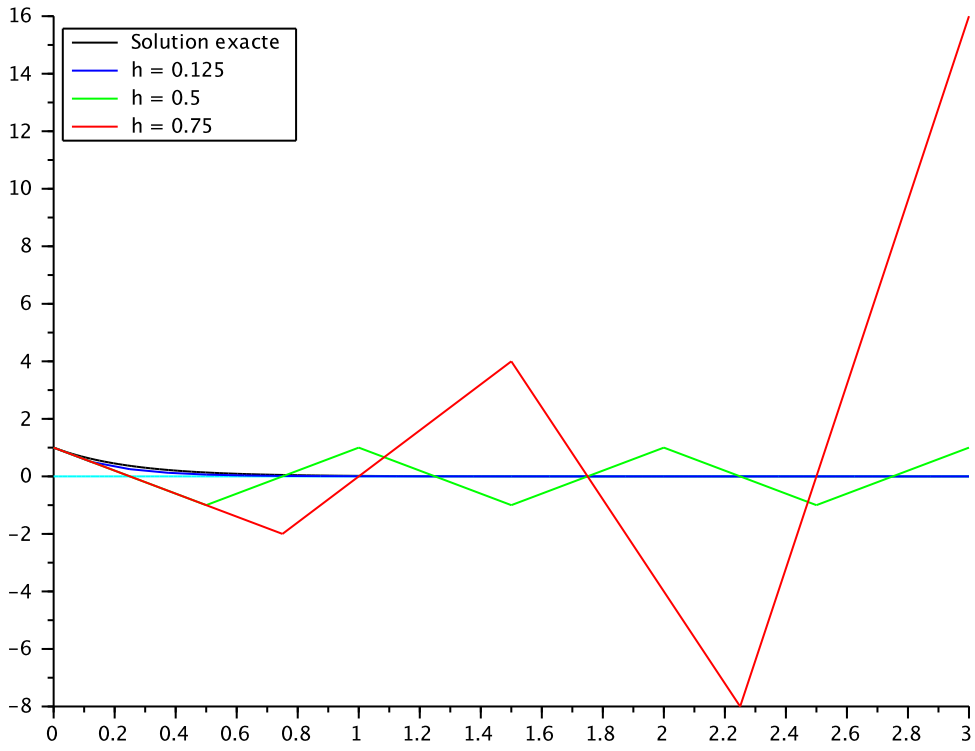


FIGURE 9.1 – Illustration de la problématique de la stabilité absolue avec $y_0 = 1$, $\lambda = -4$ et diverses valeurs de h avec le schéma d'Euler explicite. Il n'y a pas de rapport avec la stabilité au sens précédent.

est linéaire par rapport à y et indépendante de t. On a donc la relation

$$y_{n+1} = (1 + hF_h)y_n,$$

où F_h est une constante dépendant de h et telle que $F_0 = \lambda$ de façon à ce que le schéma soit consistant. En considérant les schémas que nous avons introduits jusqu'à présent appliqués dans ce cas particulier, on s'aperçoit que la relation précédente est en fait toujours de la forme

$$y_{n+1} = G(\lambda h)y_n, \tag{9.1.3}$$

ce qui implique que

$$y_n = G(\lambda h)^n y_0,$$

pour une certaine fonction G appelée *fonction d'amplification* ou *fonction de gain* du schéma. Les propriétés de décroissance vers zéro à l'infini ou non de la solution discrète vont dépendre de cette fonction. En effet, il est bien clair que $y_n \to 0$ quand $n \to +\infty$ si et seulement si $|G(\lambda h)| < 1$.

Remarquons que la solution exacte vérifie la relation analogue

$$y(t_{n+1}) = e^{\lambda h} y(t_n),$$

ces propriétés vont également être liées à la proximité de G avec l'exponentielle.

Avant d'aller plus loin, regardons quelques exemples simples de fonctions d'amplification. Pour le schéma d'Euler explicite, on a $y_{n+1} = (1 + \lambda h)y_n$, d'où $G(z) = 1 + z$. Pour le schéma d'Euler implicite, on a $y_{n+1} = y_n + \lambda h y_{n+1}$, qui se réécrit $(1 - \lambda h)y_{n+1} = y_n$, soit encore $y_{n+1} = \frac{1}{1-\lambda h} y_n$ pour $h \neq \frac{1}{\lambda}$, et donc $G(z) = \frac{1}{1-z}$. Pour le schéma d'Euler modifié, on a $y_{n+1} = y_n + \lambda h\big(y_n + \frac{h}{2}\lambda y_n\big)$, d'où $G(z) = 1 + z + \frac{z^2}{2}$. Enfin, pour le schéma de Crank-Nicolson, on obtient $G(z) = \frac{2+z}{2-z}$.

9.1.1 Domaine de stabilité

On a vu précédemment que $z_n \to 0$ si et seulement si $|G(\lambda h)| < 1$. Ceci suggère la définition suivante (on rappelle que λ est un nombre complexe) :

Définition 9.1.1 *L'ensemble des $z \in \mathbb{C}$ tels $|G(z)| < 1$ est appelé* domaine de stabilité *(ou domaine de stabilité absolue) du schéma* (9.1.3).

Remarquons que le domaine de stabilité absolue est toujours un ouvert de $\mathbb{C}$. Pour assurer la stabilité (absolue) d'un schéma, il convient donc de déterminer l'ensemble des valeurs h pour lesquelles λh appartient au domaine de stabilité du schéma. Il suffit pour cela de tracer la droite passant par l'origine et λ et déterminer son intersection avec le domaine de stabilité. On aura cependant parfois intérêt à prendre le pas h grand, pour avancer vite en temps (mais comme le domaine de stabilité est ouvert, il n'y a pas de pas le plus grand correspondant).

Pour le schéma d'Euler explicite, le domaine de stabilité est le disque du plan complexe de centre $(-1, 0)$ et de rayon 1 (voir Figure 9.2).

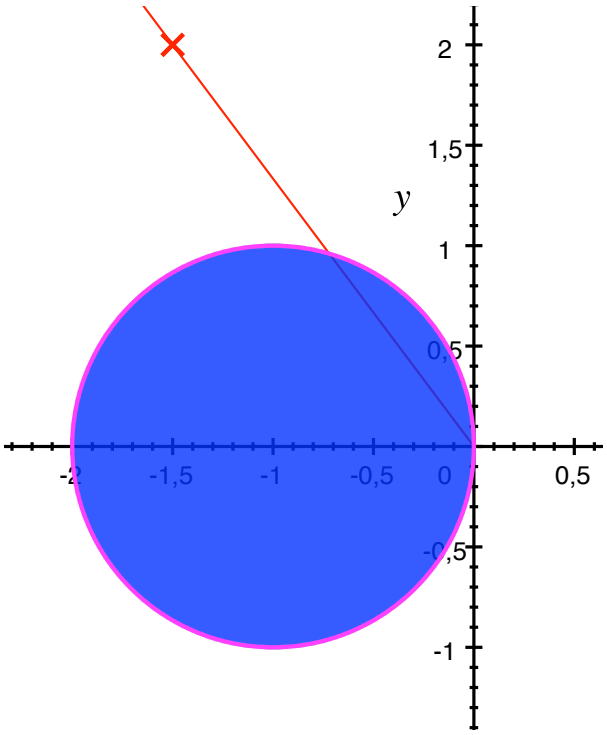


FIGURE 9.2 – Domaine de stabilité absolue du schéma d'Euler explicite.

Soit $\lambda \in \mathbb{C}$, à partie réelle $\Re(\lambda)$ négative. Pour discrétiser l'équation (9.1.2) par un schéma d'Euler explicite qui soit absolument stable, il faut choisir un pas de discrétisation h de façon que λh soit dans le disque ouvert. Au vu de la Figure 9.2, on voit que plus $\Re(\lambda)$ est petit, plus il y a de restrictions sur le pas h. Pour le cas limite $\Re(\lambda) = 0$, on ne peut plus trouver de pas h assurant une discrétisation absolument stable de (9.1.2). De même, on voit que plus λ est grand en module, plus h doit être choisi petit.

On peut quantifier la dépendance de h par rapport à λ de la façon suivante. On posant $\lambda = \varrho e^{i\theta}$ avec $\varrho > 0$ et $\cos\theta < 0$, la condition $|G(\lambda h)| < 1$ équivaut manifestement à $\varrho^2 h^2 + 2\varrho h \cos\theta < 0$. On en déduit que le schéma d'Euler explicite est stable si

$$0 < h < -2\frac{\cos\theta}{\varrho}. \tag{9.1.4}$$

Quand θ tend vers $\pm\pi/2$, h tend vers 0.

Pour le schéma d'Euler implicite, le domaine de stabilité est défini par $|1 - z| > 1$, c'est donc le plan complexe privé du disque de centre $(1, 0)$ et de rayon 1. Le demi-plan $\Re(z) < 0$ est donc entièrement inclus dans le domaine de stabilité. On en déduit que le schéma d'Euler implicite appliqué à (9.1.2), avec $\Re(\lambda) < 0$ est absolument stable, sans restriction sur le pas de discrétisation h. Rappelons que le schéma explicite est lui stable, sous réserve que la condition (9.1.4) soit vérifiée.

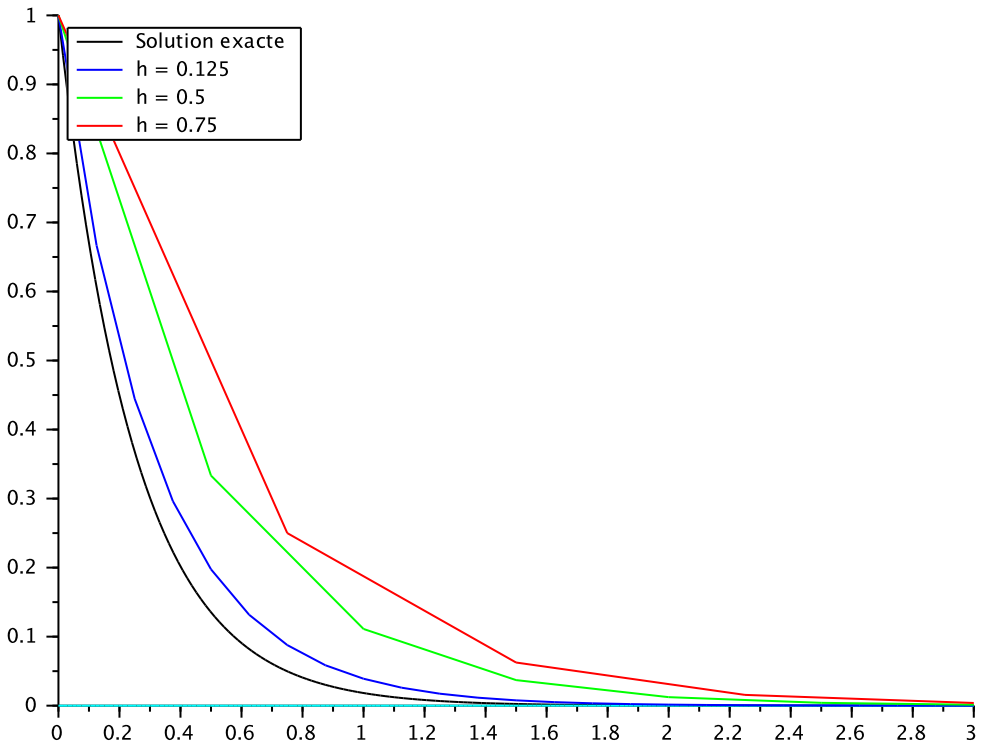


FIGURE 9.3 – Le même calcul qu'à la Figure 9.1 avec Euler implicite.

De même, le domaine de stabilité du schéma de Crank-Nicolson est défini par la condition $|2 + z| < |2 - z|$, qui est vérifiée par tous les complexes de partie réelle strictement négative.

On en déduit que le schéma de Crank-Nicolson est également absolument stable, sans restriction sur le pas de discrétisation h.

Cette situation est assez générale : les schémas implicites sont (en général) plus stables au sens de la stabilité absolue que les schémas explicites.

Enfin, le domaine de stabilité de la méthode d'Euler modifiée est donné par la condition $\left|1 + z + \frac{z^2}{2}\right| < 1$ que l'on trace à la Figure 9.4.

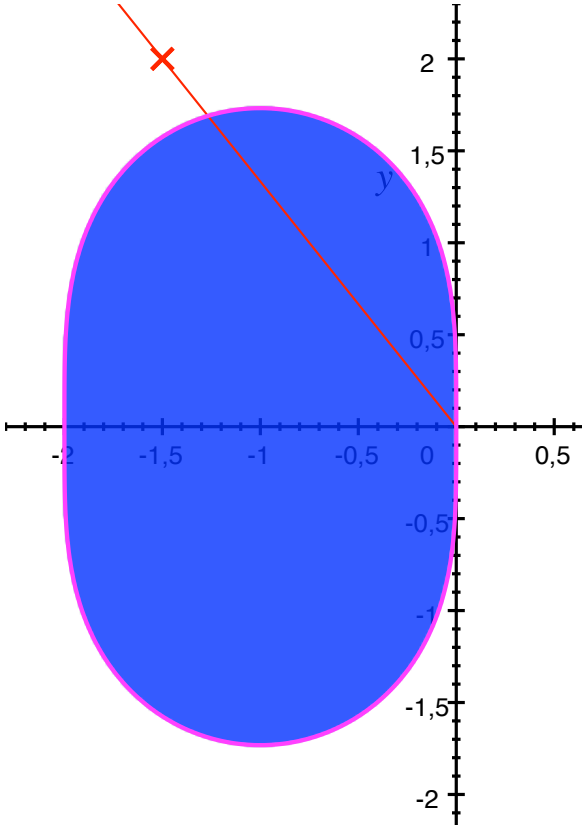


FIGURE 9.4 – Domaine de stabilité absolue du schéma d'Euler modifié.

9.1.2 Stabilité absolue de quelques schémas numériques

Domaine de stabilité des schémas RK. Nous avons vu que pour le schéma RK1, c'est-à-dire Euler explicite, on a $G(z) = 1 + z$. Pour le schéma RK2, c'est-à-dire le schéma de Heun, on a $G(z) = 1 + z + \frac{z^2}{2}$ comme pour Euler modifié (qui est aussi un schéma de Runge-Kutta d'ordre 2). Cela est dû au fait que les deux schémas coïncident sur l'équation particulière (9.1.2). [4] Pour le schéma de RK4, on a (le vérifier)

$$y_{n+1} = \left[1 + \lambda h + \frac{(\lambda h)^2}{2} + \frac{(\lambda h)^3}{6} + \frac{(\lambda h)^4}{24}\right] y_n,$$

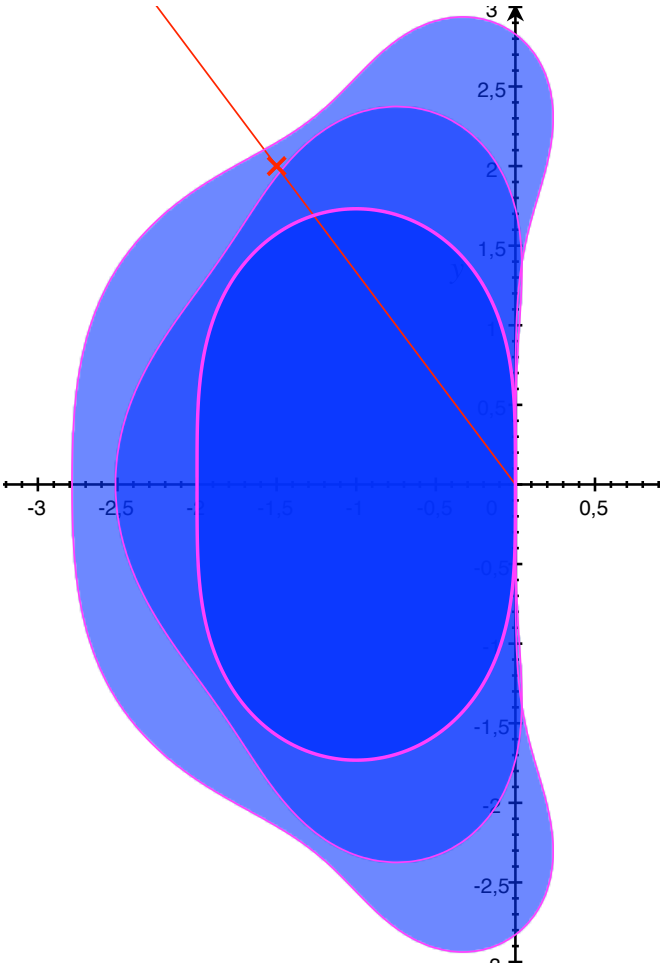


FIGURE 9.5 – Domaines de stabilité absolue des schémas RK2, RK3 et RK4. Seules les intersections avec le demi-plan $\Re(z) < 0$ sont pertinentes.

on en déduit que $G(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}$. Les domaines de stabilité des schémas de Runge-Kutta RK2, RK3 et RK4 sont tracés sur la Figure 9.5.

On observe que, quand l'ordre du schéma RK augmente, la fonction G approche de mieux en mieux l'exponentielle. On a donc intérêt, pour obtenir le schéma le plus stable du point de vue de la stabilité absolue à utiliser un schéma RK d'ordre le plus élevé possible.

Stabilité absolue des schémas à pas multiples Avant de traiter le cas général, regardons d'abord le cas d'un schéma explicite, AB2 : on initialise séparément les deux premiers pas de temps de la série perturbée, puis à chaque étape on rajoute une perturbation

$$\begin{aligned} z_0 &= y_0 + \eta_0 \\ z_1 &= y_1 + \eta_1 \\ z_{n+1} &= z_n + h\left(\frac{3}{2}f(t_n, z_n) - \frac{1}{2}f(t_{n-1}, z_{n-1})\right) + \eta_{n+1}, \quad \text{pour } n = 1, \dots, N-1. \end{aligned}$$

4. Ils coïncident aussi avec le schéma de Taylor d'ordre 2 sur cette équation. Il n'y a rien de surprenant à ce que des schémas différents donnent le même résultat sur une équation particulière.

on a donc pour $n = 1, \dots, N-1$

$$z_{n+1} - y_{n+1} = z_n - y_n + h\frac{3}{2}\Big(f(t_n, z_n) - f(t_n, y_n)\Big) - \frac{1}{2}\Big(f(t_{n-1}, z_{n-1}) - f(t_{n-1}, y_{n-1})\Big) + \eta_{n+1}, \quad \text{d'où}$$

$$\|z_{n+1} - y_{n+1}\| \leq \|z_n - y_n\| + h\frac{3}{2}\|f(t_n, z_n) - f(t_n, y_n)\| + h\frac{1}{2}\|f(t_{n-1}, z_{n-1}) - f(t_{n-1}, y_{n-1}\| + \|\eta_{n+1}\|$$

Pour majorer cette différence entre le schéma perturbé et le schéma de base, on fait l'hypothèse que la fonction second membre est globalement L-lipschitzienne par rapport à y, uniformément par rapport à t, on a donc

$$\|z_{n+1} - y_{n+1}\| \leq (1 + hL\frac{3}{2})\|z_n - y_n\| + hL\frac{1}{2}\|z_{n-1}) - y_{n-1}\| + \|\eta_{n+1}\|.$$

A cette étape, dans le cas des schémas à un pas, on utilise la formule de Gronwall discrète pour conclure. Ici, pour en faire de même on va considérer la suite

$$\theta_n = \max_{k=0,\dots,n} \|\eta_k\|.$$

Pour $n \geq 1$ on a $\|z_n - y_n\| \leq \theta_n$ et $\|z_{n-1}) - y_{n-1}\| \leq \theta_n$ donc

$$\|z_{n+1} - y_{n+1}\| \leq (1 + hL\frac{3}{2})\theta_n + hL\frac{1}{2}\theta_n + \|\eta_{n+1}\| = (1 + 2hL)\theta_n + \|\eta_{n+1}\|$$

Par ailleurs $\theta_n \leq (1 + 2hL)\theta_n$ donc

$$\theta_{n+1} = \max(\|z_{n+1} - y_{n+1}\|, \theta_n) \leq (1 + 2hL)\theta_n + \|\eta_{n+1}\|.$$

On conclut maintenant facilement que

$$\theta_n \leq e^{2LT} \sum_{k=0}^{n} \|\eta_k\|.$$

La stabilité du schéma implicite de même ordre, Adams-Moulton AM2 qui est également le schéma de Crank Nicolson, a été traitée dans la proposition 2.1.21. Pour les schémas d'Adams-Moulton d'ordre quelconque, il faut en plus utiliser la technique ci-dessus, consistant à étudier la suite θ_n plutôt que la suite $\|\eta_k\|$.

Nous allons maintenant étendre la notion de stabilité absolue au cas des schémas à pas multiples, par exemple les schémas d'Adams vus précédemment. Appliqués à l'équation différentielle linéaire (9.1.2), ils s'écrivent sous la forme générale

$$y_{n+1} = y_n + h\lambda \sum_{i=-1}^{p} b_i y_{n-i},$$

où le coefficient b_{-1} en facteur de y_{n+1} au second membre est nul pour les schémas explicites d'Adams-Bashforth et non nul pour les schémas implicites d'Adams-Moulton. Dans tous les cas, on a une relation de récurrence linéaire à $p+1$ termes à coefficients constants dont la solution générale est de la forme

$$y_n = \sum_{k=1}^{p+1} c_k \mu_k^n,$$

où les $\mu_k \in \mathbb{C}$ sont les racines de l'équation caractéristique

$$0 = P(X) = \rho(X) - z\sigma(X), \text{ avec } \rho(X) = X^{p+1} - X^p \text{ et } \sigma(X) = -\sum_{i=-1}^{p} b_i X^{p-i},$$

avec $z = h\lambda$. [5] Ces racines sont bien sûr des fonctions un peu compliquées de z, $\mu_k = \mu_k(z)$. On arrive par conséquent à la définition suivante,

Définition 9.1.2 *Le domaine de stabilité absolue d'un schéma à pas multiples est l'ensemble des $z \in \mathbb{C}$ tels que toutes les racines de l'équation caractéristique sont de module strictement inférieur à 1.*

Par exemple, le schéma d'Adams-Bashforth d'ordre deux s'écrit pour l'EDO linéaire (9.1.2)

$$y_{n+1} = y_n + h\lambda\left(\frac{3}{2}y_n - \frac{1}{2}y_{n-1}\right).$$

Son équation caractéristique est

$$\begin{aligned} P(X) &= X^2 - X - z\left(\frac{3}{2}X - \frac{1}{2}\right) \\ &= \rho(X) - z\sigma(X), \end{aligned}$$

avec

$$\rho(X) = X^2 - X \text{ et } \sigma(X) = \frac{3}{2}X - \frac{1}{2}.$$

Soit la fraction rationnelle $F(X) = \frac{\rho(X)}{\sigma(X)}$. Si $\mu(z)$ est une racine de l'équation caractéristique, on a $z = F(\mu(z))$. Par conséquent, s'il existe μ tel que $|\mu| \geq 1$ et $F(\mu) = z$, c'est que z n'est pas dans le domaine de stabilité du schéma. Celui-ci est donc le complémentaire dans $\mathbb{C}$ de l'ensemble $F(\mathbb{C} \setminus D)$ où D désigne le disque unité ouvert [6]. On a utilisé cette description pour construire la figure 9.6. La frontière du domaine de stabilité est incluse dans la courbe $\theta \mapsto F(e^{i\theta})$. [7]

Le schéma d'Adams-Moulton d'ordre deux coïncide avec le schéma de Crank-Nicolson, comme on l'a vu plus haut. Pour l'EDO linéaire (9.1.2), il s'écrit

$$y_{n+1} = y_n + \frac{h\lambda}{2}(y_{n+1} + y_n),$$

et son équation caractéristique est donc

$$P(X) = X - 1 - \frac{z}{2}(X + 1),$$

d'où

$$F(X) = 2\left(\frac{X-1}{X+1}\right).$$

5. On a supposé implicitement que ces racines sont simples, ce qui est le cas générique.

6. Attention, cela ne veut pas du tout dire que c'est l'image par F du disque D. En général, cette image est nettement plus grande.

7. Elle lui est égale dans les premiers exemples considérés ici, mais pas toujours, voir plus loin.

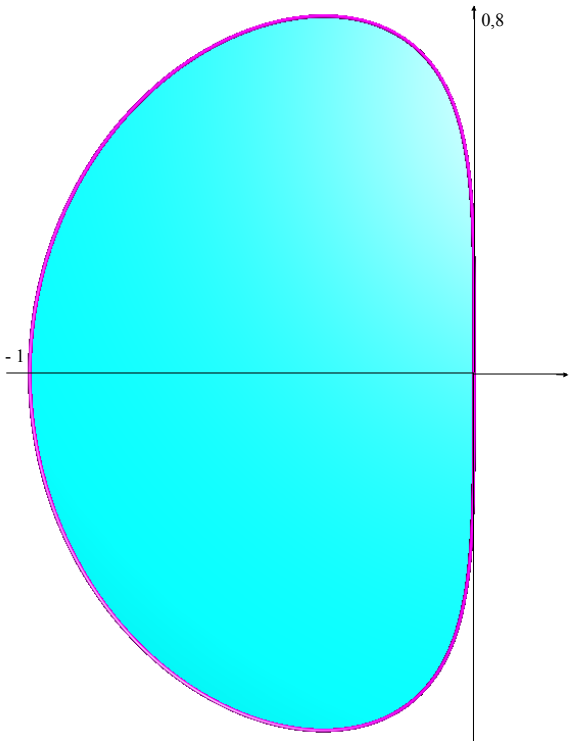


FIGURE 9.6 – Domaine de stabilité absolue du schéma d'Adams-Bashforth d'ordre 2 en bleu.

Il est bien connu que cette fonction homographique transforme l'extérieur du disque unité en le demi-plan des parties réelles positives (et si ce n'est pas bien connu, ce n'est pas bien difficile à vérifier). Le domaine de stabilité absolu du schéma est donc la totalité du demi-plan des parties réelles négatives, on l'avait déjà vu en le considérant comme un schéma à un pas. Ici aussi, à ordre égal, le domaine de stabilité absolue de la méthode implicite est plus grand que celui de la méthode explicite.

Pour les schémas d'Adams-Bashforth et d'Adams-Moulton d'ordre 3, on obtient suivant le même principe les fractions rationnelles suivantes :

$$\text{(AB3)} \quad F(X) = \frac{12(X^3 - X^2)}{23X^2 - 16X + 5},$$
$$\text{(AM3)} \quad F(X) = \frac{12(X^2 - X)}{5X^2 + 8X - 1}.$$

Les domaines de stabilité correspondants sont représentés dans les figures 9.7 et 9.8. C'est moins spectaculaire que pour l'ordre 2, mais ici encore, à ordre égal, le domaine de stabilité absolue de la méthode implicite est (beaucoup) plus grand que celui de la méthode explicite.

On a mentionné plus haut que la frontière du domaine de stabilité absolue n'est pas toujours la courbe image du cercle unité par la fraction rationnelle associée au schéma (contrairement à ce que l'on peut voir dans nombre de dessins faits dans la littérature). En voici un exemple avec le schéma AB4 pour lequel on a

$$F(X) = \frac{24(X^4 - X^3)}{55X^3 - 59X^2 + 37X - 9}.$$

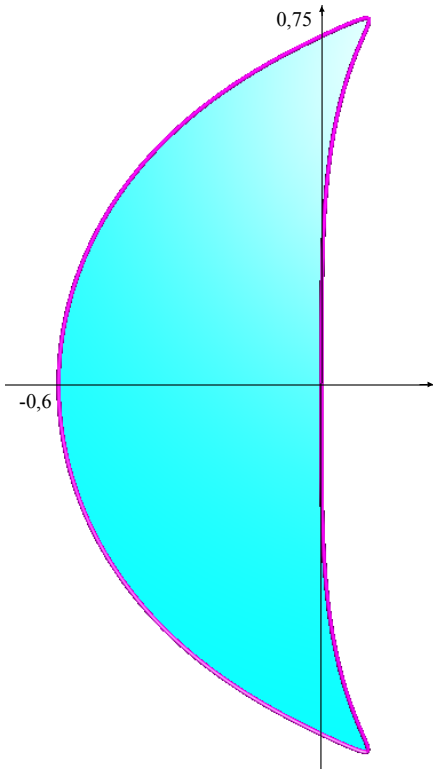


FIGURE 9.7 – Domaine de stabilité absolue du schéma d'Adams-Bashforth d'ordre 3 (seul le côté à partie réelle négative compte).

On a dessiné dans la figure 9.9, l'image de l'extérieur du disque unité en bleu (donc convention inverse des dessins précédents, ici le domaine de stabilité apparaît en blanc, ou plutôt, le domaine de stabilité est l'intersection de la partie blanche et du demi-plan à partie réelle négative) et la courbe image du cercle unité. Celle-ci n'est pas une courbe simple, elle comporte deux points doubles. En plus de ne pas être dans le demi-plan voulu, les deux boucles ainsi formées ne font pas partie du bord de l'image de l'extérieur du disque.

9.2 Schémas numériques pour les systèmes hamiltoniens

Dans cette section on va étudier une classe de schémas numériques particulièrement adaptés à la discrétisation de systèmes hamiltoniens. Un exemple simple de tels systèmes est le pendule déjà rencontré plusieurs fois auparavant. On a vu que les schémas numériques classiques ne permettaient pas de respecter les propriétés conservatives de ce système d'équations. Dans les applications astronomiques, comme le problème des N corps présenté dans l'exemple **??** ces propriétés sont cruciales, en particulier quand l'étude se fait sur un temps long. Laskar *et al* étudient par exemple le système solaire sur une très longue période, de -250 millions d'années à $+250$ millions d'années, pour calibrer les données paléoclimatologiques en fonction des variations de l'insolation de la Terre, [8]. Il s'agit ici de simulations avec un pas de temps de l'ordre de 1,8 jour poursuivies sur 500 millions d'années, où l'on calcule les mouvements combinés des huit planètes, plus la Lune et Pluton

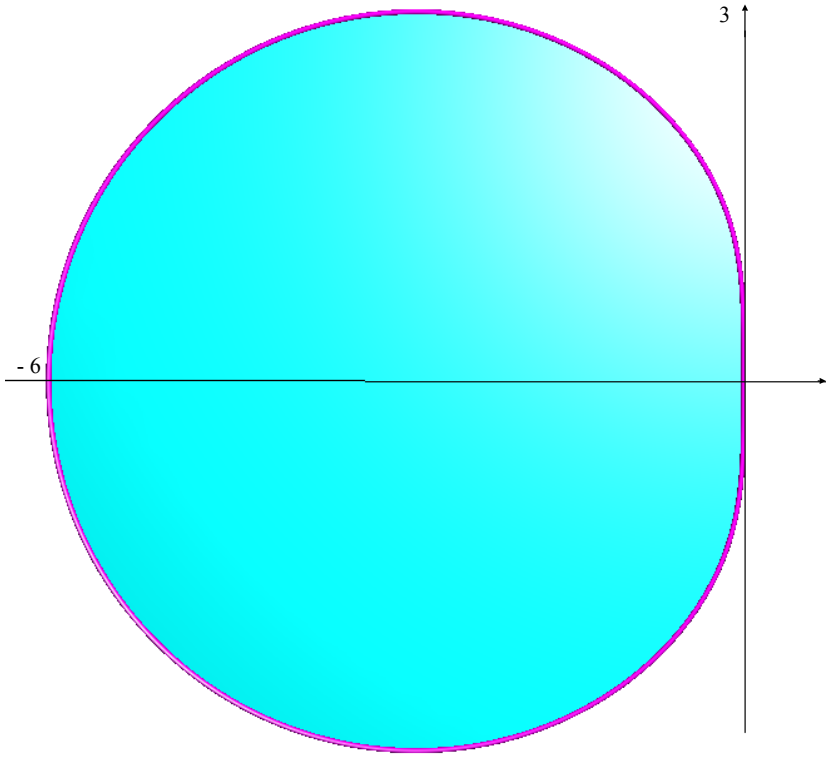


FIGURE 9.8 – Domaine de stabilité absolue du schéma d'Adams-Moulton d'ordre 3

autour du Soleil.

Considérons un problème de Cauchy admettant une unique solution locale définie sur un intervalle I_{y_0}, où y_0 désigne la donnée initiale. On considère un réel $T > 0$ et un ouvert $U \subset \mathbb{R}^m$ tels que pour tout $y_0 \in U$, $[0,T] \subset I_{y_0}$ (il en existe). Pour tout $t \in [0,T]$, on définit alors le *flot* $\varphi_t : U \to \mathbb{R}^m$, qui à $y_0 \in U$ associe $\varphi_t(y_0) = y(t)$, où y est la solution du problème de Cauchy pour la donnée initiale y_0. Le flot à l'instant t transporte donc toutes les conditions initiales appartenant à U en leur position à l'instant t en suivant l'EDO, le mot flot étant de ce point de vue particulièrement bien choisi. On montre qu'il s'agit d'un difféomorphisme de classe C^1 de U sur son image.

Introduisons quelques notions d'algèbre essentielles pour les systèmes différentiels hamiltoniens. Dans le cas d'un système hamiltonien, la dimension m est paire, soit $m = 2d$ avec $d \in \mathbb{N}^*$. Soit la matrice par blocs [8]

$$J = \begin{pmatrix} 0_d & I_d \\ -I_d & 0_d \end{pmatrix},$$

où I_d désigne la matrice identité $d \times d$ et 0_d la matrice nulle $d \times d$. On remarque que $J^2 = -I_{2d}$ et donc que $J^{-1} = -J = J^T$.

Définition 9.2.1 *Une matrice $A \in M_{2d}(\mathbb{R})$ est dite* symplectique *si elle vérifie la relation*

$$A^T J A = J.$$

8. On l'a déjà rencontrée sous forme d'opérateur page 192.

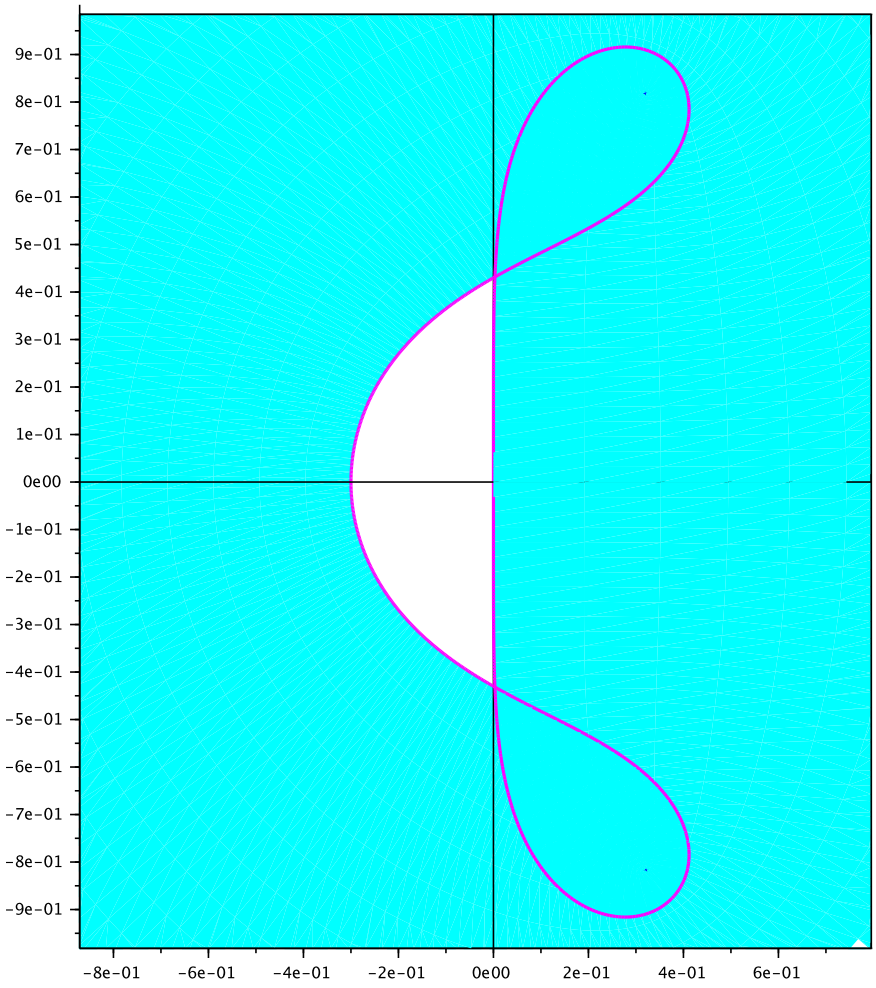


FIGURE 9.9 – Domaine de stabilité absolue du schéma d'Adams-Bashforth d'ordre 4.

Proposition 9.2.2 *Une matrice A écrite sous la forme de quatre blocs $d \times d$*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

est symplectique si et seulement si

$$\begin{cases} A_{11}^T A_{21} = A_{21}^T A_{11} \\ A_{12}^T A_{22} = A_{22}^T A_{12} \\ A_{11}^T A_{22} - A_{21}^T A_{12} = I_d \end{cases} \tag{9.2.1}$$

En particulier, pour $2d = 2$, elle est symplectique si et seulement si $\det A = 1$.

Démonstration. On effectue le produit matriciel par blocs, il vient

$$A^T J A = \begin{pmatrix} A_{11}^T A_{21} - A_{21}^T A_{11} & A_{11}^T A_{22} - A_{21}^T A_{12} \\ A_{12}^T A_{21} - A_{22}^T A_{11} & A_{12}^T A_{22} - A_{22}^T A_{12} \end{pmatrix}.$$

La nullité des deux blocs diagonaux donne les deux premières relations et l'on conclut en remarquant que chaque bloc hors diagonal est l'opposé de la transposée de l'autre.

Dans le cas $d = 1$, les matrices A_{ij} sont des scalaires et les deux premières relations de (9.2.1) sont évidemment satisfaites. La troisième n'est autre dans ce cas que $\det A = 1$. ◊

Remarquons que les deux premières relations de (9.2.1) sont équivalentes à dire que les matrices $A_{11}^T A_{21}$ et $A_{12}^T A_{22}$ doivent être symétriques.

Définition 9.2.3 *Soit U un ouvert de $\mathbb{R}^{2d}$ et $f : U \to \mathbb{R}^{2d}$ de classe C^1. L'application f est dite* symplectique *si sa matrice jacobienne est symplectique en tout point de U.*

Proposition 9.2.4 *Soient $f : U \to \mathbb{R}^{2d}$ et $g : f(U) \to \mathbb{R}^{2d}$ deux applications symplectiques. L'application $g \circ f$ est symplectique.*

Démonstration. En effet, par différentiation des fonctions composées, il suffit de vérifier que le produit de deux matrices symplectiques A et B est symplectique. Or on a

$$(AB)^T JAB = B^T (A^T JA)B = B^T JB = J,$$

trivialement. ◊

Remarque 9.2.1 Remarquons que le déterminant d'une matrice symplectique n'est pas nul. En effet, $\det J = 1$, donc $(\det A)^2 = 1$. [9] Il s'ensuit qu'une matrice symplectique est inversible et que son inverse est symplectique. L'ensemble des matrices symplectiques forme donc un groupe pour la multiplication appelé *groupe symplectique*, $\mathrm{Sp}(2d, \mathbb{R})$. Cette remarque sur le déterminant implique d'ailleurs le *théorème de Liouville*, qui dit qu'une application symplectique conserve le volume dans $\mathbb{R}^{2d}$.

On considère ici des systèmes hamiltoniens dans $\mathbb{R}^{2d}$, c'est-à-dire que la fonction inconnue s'écrit $y(t) = (q(t), p(t))^T$ où q et p sont à valeurs dans $\mathbb{R}^d$. On suppose pour simplifier que les hamiltoniens considérés sont de la forme

$$H(q, p) = T(p) + V(q),$$

où T et V sont de classe C^2 de $\mathbb{R}^d$ dans $\mathbb{R}$. On dit dans ce cas que le hamiltonien est séparable. Le système hamiltonien

$$\begin{cases} \dot{y}(t) = J\nabla H(y(t)), \\ y(0) = y_0, \end{cases}$$

s'écrit donc

$$\begin{cases} \begin{pmatrix} \dot{q}(t) \\ \dot{p}(t) \end{pmatrix} = \begin{pmatrix} \nabla_p T(p(t)) \\ -\nabla_q V(q(t)) \end{pmatrix}, \\ \begin{pmatrix} q(0) \\ p(0) \end{pmatrix} = \begin{pmatrix} q_0 \\ p_0 \end{pmatrix} \in \mathbb{R}^{2d}, \end{cases} \tag{9.2.2}$$

où ∇_q et ∇_p désignent respectivement les gradients par rapport à q et p, c'est-à-dire les vecteurs des dérivées partielles par rapport à q_i d'une part et p_i de l'autre.

9. En fait, on montre qu'une matrice symplectique est telle que $\det A = 1$, c'est-à-dire que $\mathrm{Sp}(2d, \mathbb{R}) \subset \mathrm{SL}(2d, \mathbb{R})$, avec égalité si $d = 1$ et inclusion stricte sinon.

On rappelle l'exemple canonique des oscillations planes du pendule : les petites oscillations sont décrites par le hamiltonien $H(q,p) = \frac{p^2}{2} + \frac{q^2}{2}$ et les grands oscillations par le hamiltonien $H(q,p) = \frac{p^2}{2} + 1 - \cos(q)$ (ici q est la variable d'angle, p la vitesse angulaire, la constante k étant mise égale à 1).

On note φ_t le flot associé à (9.2.2). On suppose que $\varphi_t(y_0)$ est bien défini pour tout temps t et tout $y_0 \in \mathbb{R}^{2d}$, où $y_0 = (q_0^T \; p_0^T)^T$. Le lien entre systèmes hamiltoniens et tout ce qui est symplectique est le suivant.

Proposition 9.2.5 *Le flot d'un système hamiltonien est symplectique.*

Démonstration. Soit φ_t le flot d'une EDO $\dot{y}(t) = f(y(t))$. On montre que la matrice jacobienne du flot, $\nabla\varphi_t$, est solution de *l'EDO linéaire à coefficients variables* à valeurs matricielles dans $M_m(\mathbb{R})$

$$\frac{d}{dt}\big(\nabla\varphi_t\big) = \nabla f(y(t))\nabla\varphi_t,$$

avec la condition initiale $\nabla\varphi_0 = I$ (en effet, par définition, $\varphi_0 = id$). [10]

On cherche à montrer que $\nabla\varphi_t^T J\nabla\varphi_t = J$ pour tout t. C'est trivialement vrai pour $t = 0$. Dérivons le membre de gauche par rapport au temps. Il vient

$$\begin{aligned}\frac{d}{dt}\big(\nabla\varphi_t^T J\nabla\varphi_t\big) &= \frac{d}{dt}\big(\nabla\varphi_t\big)^T J\nabla\varphi_t + \nabla\varphi_t^T J\frac{d}{dt}\big(\nabla\varphi_t\big)\\ &= \big(\nabla f(y(t))\nabla\varphi_t\big)^T J\nabla\varphi_t + \nabla\varphi_t^T J\big(\nabla f(y(t))\nabla\varphi_t\big)\\ &= \nabla\varphi_t^T\big(\nabla f(y(t))^T J + J\nabla f(y(t))\big)\nabla\varphi_t\end{aligned}$$

Nous avons dans le cas hamiltonien, $\nabla f = \nabla(J\nabla H) = J\nabla^2 H$, où $\nabla^2 H$ est symétrique, puisqu'il s'agit de la hessienne de H. Par conséquent, $\nabla f^T J = \nabla^2 H J^T J = \nabla^2 H$ et $J\nabla f = J^2\nabla^2 H = -\nabla^2 H$. On voit donc que $\frac{d}{dt}\big(\nabla\varphi_t^T J\nabla\varphi_t\big) = 0$, d'où

$$\nabla\varphi_t^T J\nabla\varphi_t = \nabla\varphi_0^T J\nabla\varphi_0 = J$$

pour tout t. ◊

Par le théorème de Liouville, on en déduit qu'un flot hamiltonien conserve le volume dans $\mathbb{R}^{2d}$. Du point de vue numérique, il est important donc de préserver également ce volume, c'est-à-dire de définir des schémas numériques symplectiques. Dans cette optique, on propose les deux schémas numériques suivants :

Le schéma d'Euler symplectique défini par

$$\begin{cases} q_{n+1} = q_n + h\nabla_p T(p_n), \\ p_{n+1} = p_n - h\nabla_q V(q_{n+1}), \end{cases} \tag{9.2.3}$$

que l'on note

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \Phi_h \begin{pmatrix} q_n \\ p_n \end{pmatrix},$$

10. Voir une démonstration en annexe à titre culturel.

avec

$$\Phi_h \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} q + h\nabla_p T(p) \\ p - h\nabla_q V(q + h\nabla_p T(p)) \end{pmatrix}. \tag{9.2.4}$$

Le schéma de Störmer-Verlet [11] défini par

$$\begin{cases} p_{n+\frac{1}{2}} = p_n - \dfrac{h}{2}\nabla_q V(q_n), \\ q_{n+1} = q_n + h\nabla_p T(p_{n+\frac{1}{2}}), \\ p_{n+1} = p_{n+\frac{1}{2}} - \dfrac{h}{2}\nabla_q V(q_{n+1}), \end{cases} \tag{9.2.5}$$

soit

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \Psi_h \begin{pmatrix} q_n \\ p_n \end{pmatrix},$$

avec

$$\Psi_h \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} q + h\nabla_p T\big(p - \frac{h}{2}\nabla_q V(q)\big) \\ p - \frac{h}{2}\big[\nabla_q V(q) + \nabla_q V\big(q + h\nabla_p T\big(p - \frac{h}{2}\nabla_q V(q)\big)\big)\big] \end{pmatrix}. \tag{9.2.6}$$

Ces deux schémas sont explicites et à un pas. Il est à peu près évident qu'ils sont stables et consistants, donc convergents. Pour étudier leurs propriétés, on se place pour simplifier dans le cas $d = 1$. On écrira dans ce cas $\nabla_q V(q) = V'(q)$ et $\nabla_p T(p) = T'(p)$, puisque V et T sont alors des fonctions d'une seule variable.

Proposition 9.2.6 *Les applications Φ_h et Ψ_h sont symplectiques.*

Démonstration. On part de (9.2.4) pour calculer la matrice jacobienne de Φ_h,

$$\nabla\Phi_h = \begin{pmatrix} 1 & hT''(p) \\ -hV''(q + hT'(p)) & 1 - h^2V''(q + hT'(p))T''(p) \end{pmatrix}.$$

Or on a vu à la proposition 9.2.2 que dans le cas $d = 1$, une matrice est symplectique si et seulement si son déterminant vaut 1, ce qui est bien évidemment le cas de la matrice ci-dessus.

Pour montrer que Ψ_h est symplectique, au lieu de partir brutalement de la formule (9.2.6), qui est un peu longue, on la décompose sous la forme $\Psi_h = \Psi_h^1 \circ \Psi_h^0$ avec

$$\Psi_h^0 \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} q + \frac{h}{2}T'\big(p - \frac{h}{2}V'(q)\big) \\ p - \frac{h}{2}V'(q) \end{pmatrix}$$

et

$$\Psi_h^1 = \Phi_{\frac{h}{2}}.$$

Tout d'abord, Ψ_h^1 est symplectique d'après ce qui précède en changeant h en $\frac{h}{2}$. On vérifie que Ψ_h^0 est symplectique de la même manière que pour Φ_h. On utilise enfin la proposition 9.2.4 pour conclure. ◇

Corollaire 9.2.7 *Pour les deux schémas, l'application $(q_0, p_0) \mapsto (q_n, p_n)$ est symplectique pour tout n.*

11. Fredrik Carl Mülertz Størmer, 1874–1957 ; Loup Verlet, 1931–.

Démonstration. En effet, $(q_n, p_n) = \Phi_h^n(q_0, p_0)$ pour le schéma d'Euler symplectique et l'on conclut par la proposition 9.2.4, idem pour le schéma de Störmer-Verlet. ◊

On en déduit que les schémas d'Euler symplectique et de Störmer-Verlet conservent exactement les volumes, comme le système hamiltonien lui-même, au cours des itérations, modulo les erreurs d'arrondi qui peuvent finir par s'accumuler. [12] Les applications Φ_h^n et Ψ_h^n sont les flots numériques des deux schémas, voir Figures 9.10 et 9.11.

On l'a déjà vu numériquement, mais remarquons quand même que le schéma d'Euler classique n'est pas symplectique. Il correspond en effet à l'application

$$\Theta_h \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} q + hT'(p) \\ p - hV'(q) \end{pmatrix},$$

pour laquelle on a

$$\nabla\Theta_h = \begin{pmatrix} 1 & hT''(p) \\ -hV''(q) & 1 \end{pmatrix},$$

si bien que

$$\det \nabla\Theta_h = 1 + h^2 T''(p)V''(q) \neq 1$$

dès que ni T ni V ne sont affines. En fait, si T et V sont strictement convexes, comme c'est le cas pour les petites oscillations du pendule, on a $\det \nabla\Theta_h > 1$ ce qui correspond à une augmentation stricte des volumes au cours des itérations.

Pour ce qui concerne la conservation de l'hamiltonien, on a vu sur des exemples numériques que celui-ci ne l'est pas de façon exacte, mais est conservé " en moyenne " par les schémas symplectiques. On peut montrer que cela est dû à l'existence d'un hamiltonien approché qui est lui conservé par les schémas numériques.

On va maintenant étudier l'ordre des deux schémas symplectiques.

Proposition 9.2.8 *Le schéma d'Euler symplectique est d'ordre* 1 *et le schéma de Störmer-Verlet est d'ordre* 2*.*

Démonstration. Pour le schéma d'Euler symplectique, on a l'erreur de consistance

$$\begin{aligned} \varepsilon_h &= \begin{pmatrix} q(t_{n+1}) - q(t_n) - hT'(p(t_n)) \\ p(t_{n+1}) - p(t_n) + hV'\big(q(t_n) + hT'(p(t_n))\big) \end{pmatrix} \\ &= h \begin{pmatrix} O(h) \\ -V'(q(t_n)) + V'\big(q(t_n) + hT'(p(t_n))\big) + O(h) \end{pmatrix} = O(h^2), \end{aligned}$$

et le schéma est d'ordre 1.

Pour le schéma de Störmer-Verlet, on sépare les deux composantes de l'erreur de consistance, et l'on commence par la première, plus facile à traiter. En effet, on a

$$\begin{aligned} q(t_{n+1}) - q(t_n) - hT'\Big(p(t_n) - \frac{h}{2}V'(q(t_n))\Big) &= h\dot{q}(t_n) + \frac{h^2}{2}\ddot{q}(t_n) + O(h^3) \\ &\quad - hT'(p(t_n)) + \frac{h^2}{2}T''(p(t_n))V'(q(t_n)) + O(h^3) \\ &= O(h^3), \end{aligned}$$

12. Dans l'exemple du calcul astronomique cité plus haut, il y a 200 itérations par an, soit 5 10^{10} itérations vers le passé et 5 10^{10} itérations vers le futur.

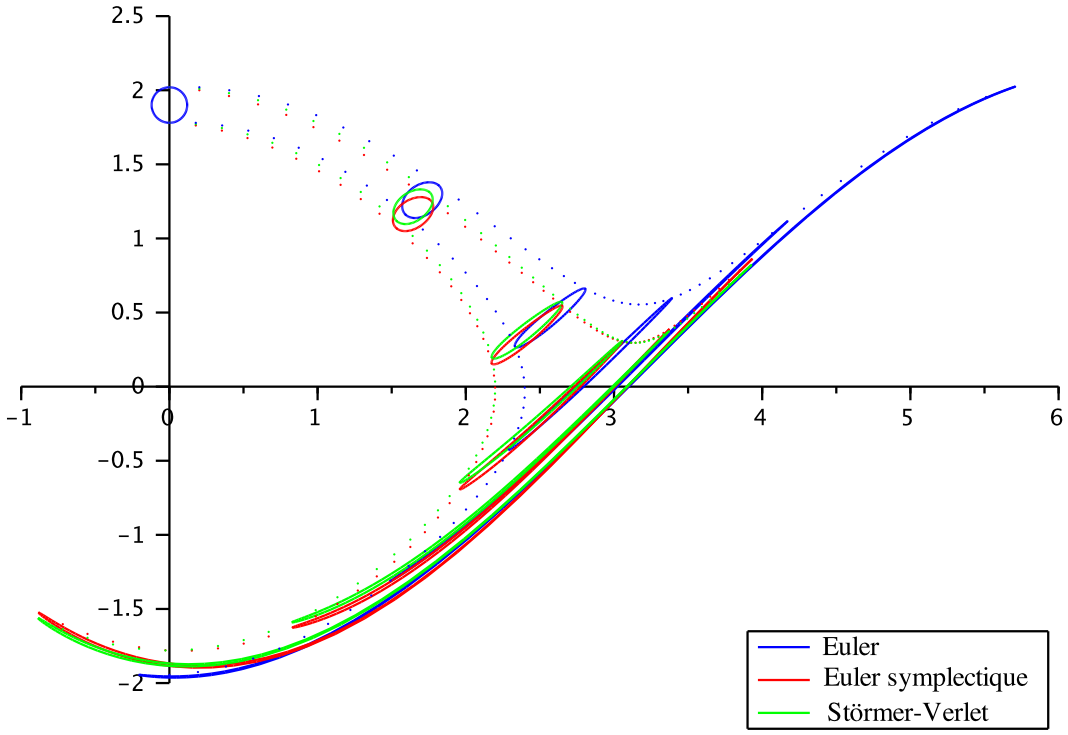


FIGURE 9.10 – Le flot approché des grandes oscillations du pendule. On a tracé les images par les flots numériques des schémas d'Euler, d'Euler symplectique et de Störmer-Verlet avec un pas de temps $h = 0.1$, du disque centré en $(0, 1.9)$ et de rayon 0.12, aux temps $t = 1$, 2, 3, 4 et 5. On a également tracé en pointillés les trajectoires numériques issues des points $(0, 1.9 \pm 0.12)$. On constate une différence rapidement croissante du schéma d'Euler par rapport aux deux schémas symplectiques. Ne pas hésiter à zoomer fortement sur la figure pour y voir plus clair.

puisque $\dot{q}(t_n) = T'(p(t_n))$ et $\ddot{q}(t_n) = T''(p(t_n))\dot{p}(t_n) = -T''(p(t_n))V'(q(t_n))$.

Pour la deuxième composante de l'erreur de consistance, on commence par remarquer que, posant $\tilde{p}_{n+\frac{1}{2}} = p(t_n) - \frac{h}{2}V'(q(t_n))$, on a

$$T'\left(\tilde{p}_{n+\frac{1}{2}}\right) = T'(p(t_n)) + O(h).$$

On obtient donc

$$\begin{aligned} p(t_{n+1}) - \tilde{p}_{n+\frac{1}{2}} + \frac{h}{2}V'\left(q(t_n) + hT'\left(\tilde{p}_{n+\frac{1}{2}}\right)\right) &= h\dot{p}(t_n) + \frac{h^2}{2}\ddot{p}(t_n) + O(h^3) \\ &\quad + hV'(q(t_n)) + \frac{h^2}{2}V''(q(t_n))T'(p(t_n)) + O(h^3) \\ &= O(h^3), \end{aligned}$$

puisque $\dot{p}(t_n) = -V'(q(t_n))$ et $\ddot{p}(t_n) = -V''(q(t_n))\dot{q}(t_n) = -V''(q(t_n))T'(p(t_n))$.

Le schéma de Störmer-Verlet est par conséquent d'ordre 2. ◊

On a déjà vu les performances du schéma Euler symplectique sur l'exemple du pendule, dans les Figures 4.4, 4.5 et 4.6.

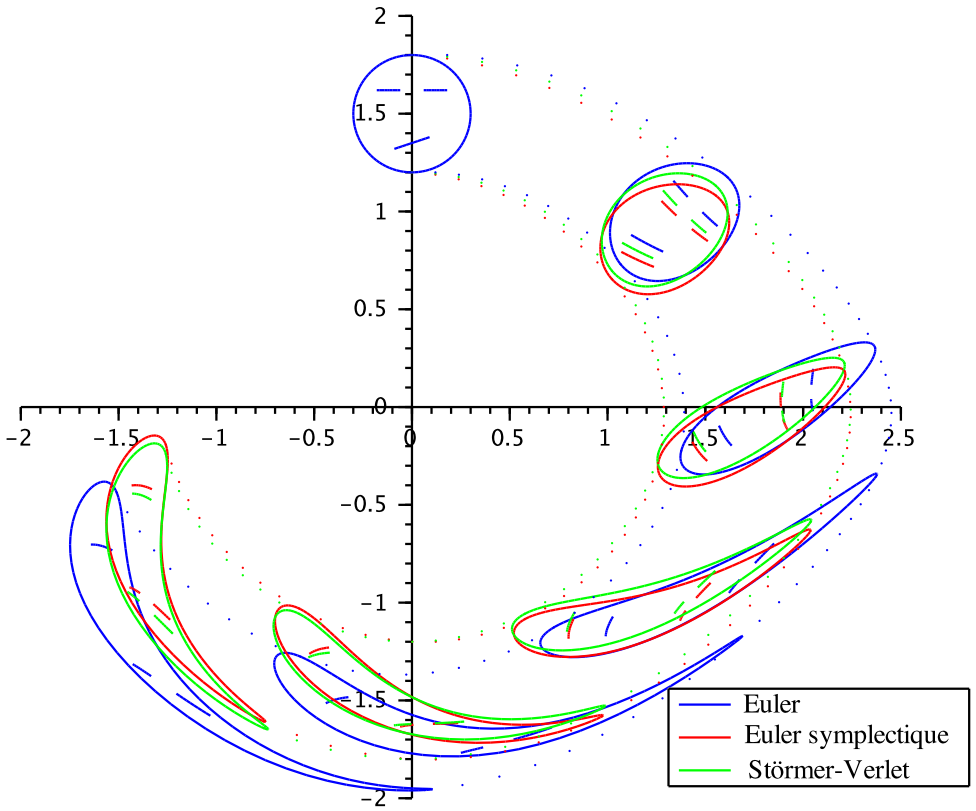


FIGURE 9.11 – Le même calcul avec le disque centré en $(0, 1.5)$ et de rayon 0.3 moins quelques traits. On voit assez clairement l'augmentation du volume causée par le schéma d'Euler.

Annexe : différentielle du flot

On établit ici la formule donnant la différentielle du flot. [13] On suppose la fonction second membre f de classe C^1. Procédant par condition nécessaire, il est facile de se convaincre que si le flot est différentiable, alors sa différentielle est nécessairement solution du problème de Cauchy

$$\frac{d}{dt}(\nabla\varphi_t) = \nabla f(\varphi_t)\nabla\varphi_t, \quad \nabla\varphi_0 = I.$$

Condition suffisante : on prend donc la solution du problème de Cauchy linéaire à valeurs matricielles

$$\begin{cases} \dfrac{dA}{dt}(t) = \nabla f(\varphi_t(y_0))A(t), \\ A(0) = I. \end{cases}$$

La fonction $t \mapsto A(t)$ existe et est unique sans problème. On va vérifier qu'elle nous donne bien la différentielle recherchée. Pour cela, on considère

$$z_h(t) = \varphi_t(y_0 + h) - \varphi_t(y_0) - A(t)h,$$

13. En confondant allègrement différentielle et matrice jacobienne, ce qui est mal dans l'absolu, mais on ne va pas être aussi pointilleux quand cela n'en vaut pas vraiment la peine.

et l'on va montrer que cette quantité tend vers 0 plus vite que $\|h\|$ quand $\|h\| \to 0$.

Rappelons tout d'abord quelques points concernant les EDO linéaires à coefficients variables, *cf.* paragraphe C.2.3. Si $y'(t) = B(t)y(t)$ et $\|B(t)\| \leq C$, alors $\|y(t)\| \leq e^{CT}\|y_0\|$. Par ailleurs, si $y'(t) = B(t)y(t) + b(t)$, alors on a $y(t) = W(t)\big(y_0 + \int_0^t W(s)^{-1}b(s)\,ds\big)$, où W désigne la matrice wronskienne, solution de $W'(t) = B(t)W(t)$, $W(0) = I$. On en déduit que $\|W(t)\| \leq e^{CT}$ et aussi que $\|W(t)^{-1}\| \leq e^{CT}$ en inversant le sens du temps.

Écrivons l'EDO satisfaite par la fonction z_h. On a

$$\begin{aligned} z_h'(t) &= f(\varphi_t(y_0 + h)) - f(\varphi_t(y_0)) - \nabla f(\varphi_t(y_0))A(t)h \\ &= \int_0^1 \nabla f\big(s\varphi_t(y_0 + h) + (1 - s)\varphi_t(y_0)\big)(z_h(t) + A(t)h)\,ds - \nabla f(\varphi_t(y_0))A(t)h \\ &= B_h(t)z_h(t) + b_h(t), \end{aligned}$$

avec

$$B_h(t) = \int_0^1 \nabla f\big(s\varphi_t(y_0 + h) + (1 - s)\varphi_t(y_0)\big)\,ds$$

et

$$b_h(t) = \left(\int_0^1 \big[\nabla f\big(s\varphi_t(y_0 + h) + (1 - s)\varphi_t(y_0)\big) - \nabla f(\varphi_t(y_0))\big]\,ds\right)A(t)h.$$

On voit donc que $\|B_h(t)\| \leq C$ pour h dans la boule unité par exemple et $\frac{\|b_h(t)\|}{\|h\|} \to 0$ quand $\|h\| \to 0$ uniformément par rapport à t. Comme par ailleurs $z_h(0) = 0$, il vient $\frac{\|z_h(t)\|}{\|h\|} \to 0$ quand $\|h\| \to 0$, d'où le résultat. ◊

Annexe A

Rappels de topologie

A.1 Espaces métriques

On connaît les espaces $\mathbb{R}^n$ munis de leurs diverses normes usuelles, toutes équivalentes entre elles. Il s'agit de cas particuliers d'une notion topologique bien plus générale, celle des espaces métriques. Ce n'est pas la notion la plus générale en topologie, mais elle nous suffira ici.

Définition A.1.1 *Soit E un ensemble et $d : E \times E \longrightarrow \mathbb{R}_+$ une application telle que*
i) $\forall (x, y) \in E^2, d(x, y) = d(y, x)$,
ii) $d(x, y) = 0$ si et seulement si $x = y$,
iii) $\forall (x, y, z) \in E^3, d(x, y) \leq d(x, z) + d(z, y)$,
est appelée une distance *sur E. Le couple (E, d) est appelé un* espace métrique.

Les trois propriétés i), ii) et iii), en particulier la troisième appelée *l'inégalité triangulaire*, sont des abstractions des propriétés de la distance physique de notre expérience de tous les jours, qui est la distance euclidienne dans $\mathbb{R}^3$. L'inégalité triangulaire est simplement l'expression du fait qu'il est plus court d'aller directement de x à y que d'y aller en passant par z.

Quelques exemples :

1. La distance dite usuelle sur $\mathbb{R}$, celle que l'on utilise sans y penser spécialement, est simplement définie par $d(x, y) = |x - y|$.

2. Sur $\mathbb{R}^n$, $n \geq 1$, on connaît plusieurs normes : $\|x\|_1 = \sum_{i=1}^n |x_i|$, $\|x\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$, $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$ (parmi tant d'autres). Chacune de ces normes donne naissance à une distance différente : $d_1(x, y) = \|x - y\|_1$, $d_2(x, y) = \|x - y\|_2$ qui est la distance euclidienne, $d_\infty(x, y) = \|x - y\|_\infty$, lesquelles donnent autant de structures d'espace métrique différentes sur $\mathbb{R}^n$: $(\mathbb{R}^n, d_1)$, $(\mathbb{R}^n, d_2)$, $(\mathbb{R}^n, d_\infty)$. Ces distances sont équivalentes entre elles, c'est-à-dire qu'il existe des constantes $0 < \alpha_{ij} \leq \beta_{ij}$ telles que pour tous x et y dans $\mathbb{R}^n$, $\alpha_{ij} d_i(x, y) \leq d_j(x, y) \leq \beta_{ij} d_i(x, y)$ avec i et j quelconques dans $\{1, 2, +\infty\}$. Ceci provient de l'équivalence correspondante des normes (qui se lit ci-dessus avec $y = 0$). Toutes les notions topologiques qui suivront (continuité, suites convergentes, etc.) seront en conséquence identiques. Toutes les normes sur un même espace vectoriel sur $\mathbb{R}$ de dimension finie sont équivalentes.

3. Plus généralement, tout espace vectoriel normé $(E, \|\cdot\|_E)$ est automatiquement doté d'une structure d'espace métrique compatible avec sa norme en posant $d(x,y) = \|x-y\|_E$. C'est une distance qui est invariante par translation : $d(x+z, y+z) = d(x,y)$. Ainsi, l'espace $C^0([0,1])$ des fonctions continues de l'intervalle $[0,1]$ à valeurs dans $\mathbb{R}$ muni de la norme naturelle pour cet espace, $\|g\|_{C^0} = \max_{t\in[0,1]} |g(t)|$, est un espace métrique pour la distance $d(f,g) = \max_{t\in[0,1]} |f(t) - g(t)|$. Il s'agit d'un espace vectoriel de dimension infinie. Il y a d'autres normes qui peuvent être utiles sur cet espace et qui ne sont pas équivalentes à la norme naturelle, comme $\|g\|_{L^1} = \int_0^1 |g(t)|\, dt$. La situation est donc plus compliquée qu'en dimension finie.
4. Si $g\colon \mathbb{R} \to \mathbb{R}$ est injective, alors $d(x,y) = |g(x) - g(y)|$ est une distance sur $\mathbb{R}$. Si $g(x) = x$ pour tout x, c'est la distance usuelle, sinon c'en est une autre. Si g n'est pas affine, cette distance ne provient pas d'une norme.
5. Si E est n'importe quel ensemble, $d(x,y) = 0$ si $x = y$, $d(x,y) = 1$ sinon, définit une distance sur E appelée *distance discrète*. En d'autres termes, n'importe quel ensemble peut être doté de cette structure d'espace métrique discret, mais cette structure n'est pas nécessairement très intéressante, sauf cas particulier.
6. Si (E,d) est un espace métrique, alors toute partie X de E peut être munie de la restriction de la distance à $X \times X$ et devenir ainsi automatiquement un autre espace métrique. On dit que c'est la distance *induite* par celle de E ou que (X,d) est un sous-espace métrique [1] de (E,d).

Attention au fait qu'un espace métrique est bien un *couple* (ensemble, distance sur cet ensemble). Un même ensemble muni de distances différentes correspond à plusieurs espaces métriques différents, *cf*. l'exemple de $\mathbb{R}^n$ avec différentes distances plus haut.

Dans un espace métrique (E,d), on peut définir des notions topologiques comme les ouverts — ce sont les réunions quelconques de boules ouvertes $B(x,r) = \{y \in E; d(y,x) < r\}$, les fermés — ce sont les complémentaires des ouverts, les applications continues, etc. Des distances différentes peuvent parfaitement engendrer les mêmes ouverts. C'est le cas dans $\mathbb{R}^n$ muni de ses distances usuelles. On dit alors qu'elles engendrent la même topologie. Mais le contraire peut parfaitement se produire également, comme dans l'exemple 3 plus haut où les deux distances engendrent des topologies différentes.

Voyons de plus près cette notion d'application continue g d'un espace métrique (E,d) à valeurs dans un autre espace métrique (F,δ). Soit $a \in E$, on dit que *g est continue en a* si

$$\forall \varepsilon > 0, \exists \eta > 0; \forall x \in E, d(x,a) \leq \eta \Rightarrow \delta(g(x), g(a)) \leq \varepsilon.$$

En d'autres termes, on peut être assuré que $g(x)$ est proche de $g(a)$ au sens de la distance δ si l'on prend x suffisamment proche de a au sens de la distance d. C'est très intuitif. Dans le cas où $E = F = \mathbb{R}$ et $d = \delta$ est la distance usuelle, on retrouve exacte la continuité d'une fonction telle que définie en L1.

Bien sûr, on dit que g est continue sur E si g est continue en tout point a de E. C'est équivalent à la propriété que l'image réciproque par g de tout ouvert de F est un ouvert de E. La continuité est donc en fait une notion topologique, et pas seulement métrique. [2] En

1. On ne distingue pas dans la notation d et sa restriction à $X \times X$.

2. Il en va de même pour la continuité en un point $a \in E$ qui s'exprime en termes purement topologiques, sans métrique, en disant que pour tout ouvert V de F contenant $g(a)$, il existe un ouvert de E contenant a dont l'image est incluse dans V. L'interprétation intuitive est exactement la même.

particulier sur $\mathbb{R}^n$ *muni de ses distances usuelles*, la continuité ou non d'une application à valeurs dans $\mathbb{R}^m$ muni aussi de ses distances usuelles, ne dépend pas du choix particulier de distance retenu. En conséquence, on prend la distance la plus pratique pour ce que l'on a à faire sur l'instant. Tout ce qui précède s'applique sans modification quand X est une partie de E munie de la distance induite et $g\colon X \to F$.

Dans un espace métrique, on peut également définir la *convergence* des suites comme on le fait dans $\mathbb{R}$ muni de sa distance usuelle : une suite x_n de E, c'est-à-dire une application de $\mathbb{N}$ dans E, converge vers un élément $x \in E$ si

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N}, \forall n \geq n_0, d(x_n, x) \leq \varepsilon.$$

On appelle bien sûr x la *limite* de la suite x_n au sens de l'espace métrique (E, d). On note dans ce cas $x_n \to x$ dans E quand $n \to +\infty$, ou bien $x = \lim_{n\to+\infty} x_n$ au sens de E, ou $x = \lim_{n\to+\infty} x_n$ sans rien préciser quand la distance est sous-entendue. [3] Le contenu intuitif est très clair aussi : les valeurs prises par la suite u_n se rapprochent autant qu'on le souhaite de la limite x au sens de la distance d à condition que l'on prenne n assez grand. [4]

Cette limite est unique si elle existe. En effet, si l'on prend deux limites x et $\bar{x}$ de la même suite, on obtient par l'inégalité triangulaire que pour tout $\varepsilon > 0$, $d(x, \bar{x}) \leq 2\varepsilon$, ce qui implique $d(x, \bar{x}) = 0$, ce qui implique $x = \bar{x}$. [5] Dans le cas de l'espace $C^0([0, 1])$ muni de la distance indiquée un peu plus haut, on reconnaît dans la convergence au sens de cette distance simplement la convergence uniforme d'une suite de fonctions continues sur $[0, 1]$, vers une fonction continue sur $[0, 1]$. [6]

Il n'est pas très difficile de voir qu'une application $g\colon E \to F$ est continue en $a \in E$ si et seulement si, pour toute suite $x_n \in E$ qui tend vers a dans E au sens de la distance d, on a que $g(x_n)$ tend vers $f(a)$ dans F au sens de la distance δ, c'est-à-dire que $\delta\big(g(x_n), g(a)\big) \to 0$ pour toute suite x_n telle que $d(x_n, a) \to 0$, ce qui ramène à des convergences dans $\mathbb{R}_+$.

Attention, *a priori* une suite dans un même ensemble peut très bien converger pour une distance et pas pour une autre. Encore pire, elle peut très bien converger vers une limite pour une distance et vers une autre limite différente pour une autre distance ! Tout cela est fondamentalement distance-dépendant.

Dans la pratique bien sûr, quand on a un problème spécifique à résoudre, il y a essentiellement toujours un choix de distance naturel qui s'impose. On ne s'amuse pas à fabriquer exprès des contre-exemples rien que pour le plaisir, mais il faut savoir que de tels contre-exemples existent.

Dans un espace métrique (E, d), on dispose également de la notion de *suite de Cauchy*, [7]

3. Attention quand même, il est loin d'être rare que l'on ait besoin d'avoir plusieurs distances non équivalentes sous la main en même temps sur un même ensemble.

4. La convergence d'une suite est également une notion qui est de nature purement topologique et s'exprime aussi uniquement en termes d'ouverts. Elle n'a pas besoin de distance, mais on se limite ici aux espaces métriques.

5. Plus généralement, un espace métrique est un espace *séparé*.

6. Notons que si l'on oublie le côté espace vectoriel normé, on sait quand même depuis longtemps que si une suite de fonctions continues converge uniformément vers une fonction, celle-ci est nécessairement continue. On peut remettre ce résultat élémentaire dans le cadre des espaces métriques en considérant l'espace (beaucoup) plus grand $B([0, 1])$ des fonctions bornées sur $[0, 1]$, muni de la même distance avec le max remplacé par un sup. Il dit alors que $C^0([0, 1])$ est un sous-espace vectoriel fermé de $B([0, 1])$.

7. Là, attention, ce n'est plus une notion exprimable en termes d'ouverts. Elle n'est pas topologique. La distance est bien utile. La raison en est simple : il est facile de construire deux distances sur un même ensemble qui engendrent les mêmes ouverts, mais qui n'ont pas les mêmes suites de Cauchy.

exactement comme dans $\mathbb{R}$ muni de sa distance usuelle :

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N}, \forall n, m \geq n_0, d(x_n, x_m) \leq \varepsilon.$$

Évidemment, toute suite convergente est de Cauchy, mais la convergence ou non des suites de Cauchy est une propriété qu'un espace métrique possède ou ne possède pas.

Définition A.1.2 *On dit qu'un espace métrique (E, d) est* complet *si toute suite de Cauchy y est convergente.* [8]

Les espaces $\mathbb{R}^n$ munis de leurs distances usuelles sont complets. Par contre, $\mathbb{Q}$ muni de la distance usuelle n'est pas complet, la suite des approximations de Héron de $\sqrt{2}$ étant de Cauchy et non convergente (dans $\mathbb{Q}$, puisque $\sqrt{2} \notin \mathbb{Q}$). L'espace $C^0([0,1])$ muni de la distance de la convergence uniforme est complet. Par contre, il n'est pas complet si on le munit de la distance $d(f,g) = \int_0^1 |f(t) - g(t)|\, dt$ associée à la norme L^1 de l'exemple 3. Plus généralement, tout espace vectoriel normé qui est complet pour la distance associée à sa norme est appelé un *espace de Banach*.

A.2 Normes d'application linéaire et norme subordonnée

Dans cette section, on considère des normes sur $M_n(\mathbb{R})$. Comme toutes les normes sont équivalentes en dimension finie, on pourrait prendre, par exemple $|||A||| = \max_{i,j} |a_{ij}|$. Pour ce cours, il est plus agréable pour travailler d'en choisir une qui soit adaptée à la norme que l'on a déjà adoptée sur $\mathbb{R}^n$. Pour cela, on utilise la notion de *norme matricielle subordonnée*. Cette notion vaut pour n'importe quelle norme sur $\mathbb{R}^n$. On va voir qu'il s'agit en fait du pendant matriciel de la norme d'application linéaire.

Définition A.2.1 *Soit $\|\cdot\|$ une norme quelconque sur $\mathbb{R}^n$. L'application $A \mapsto |||A||| = \sup_{\|x\| \leq 1} \|Ax\|$ est une norme sur $M_n(\mathbb{R})$ appelée* norme matricielle subordonnée *à la norme sur $\mathbb{R}^n$.*

Proposition A.2.2 *Pour tout $x \in \mathbb{R}^n$, on a*

$$\|Ax\| \leq |||A|||\, \|x\|.$$

De plus, pour tous $A, B \in M_n(\mathbb{R})$, on a

$$|||AB||| \leq |||A|||\, |||B|||.$$

Enfin, dans le cas de la norme $\|x\| = \max_i |x_i|$, on a

$$|||A||| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

8. Naturellement, la complétude n'est pas non plus une notion topologique. Il faut quelque chose en plus que juste les ouverts, une distance pour ce qui nous concerne et sans aller trop loin.

Démonstration. Pour tout $\lambda \in \mathbb{R}$, $|||\lambda A||| = \sup_{\|x\|\leq 1} \|\lambda Ax\| = |\lambda| \sup_{\|x\|\leq 1} \|Ax\| = |\lambda|\,|||A|||$, d'où la positivité homogène. De même, $|||A+B||| = \sup_{\|x\|\leq 1} \|(A+B)x\| = \sup_{\|x\|\leq 1} \|Ax + Bx\| \leq \sup_{\|x\|\leq 1}(\|Ax\| + \|Bx\|) \leq \sup_{\|x\|\leq 1} \|Ax\| + \sup_{\|x\|\leq 1} \|Bx\| = |||A||| + |||B|||$, d'où l'inégalité triangulaire. Enfin, si $\sup_{\|x\|\leq 1} \|Ax\| = 0$, c'est bien clairement que $A = 0$. On a affaire à une norme sur l'espace $M_n(\mathbb{R})$.

Pour tout $x \in \mathbb{R}^n$, il existe $u \in \mathbb{R}^n$, $\|u\| = 1$, tel que $x = \|x\|u$. En effet, si $x = 0$, n'importe quel u fait l'affaire et si $x \neq 0$, on prend $u = \frac{x}{\|x\|}$. Donc

$$\|Ax\| = \|x\|\,\|Au\| \leq |||A|||\,\|x\|.$$

Ensuite, pour tout $\|x\| \leq 1$, on a

$$\|(AB)x\| = \|A(Bx)\| \leq |||A|||\,\|Bx\| \leq |||A|||\,|||B|||,$$

d'où la deuxième inégalité en passant au sup à gauche.

Enfin dans le cas $\|x\| = \max_i |x_i| \leq 1$, on voit que

$$\|Ax\| = \max_{1\leq i\leq n}\Big|\sum_{j=1}^{n} a_{ij}x_j\Big| \leq \max_{1\leq i\leq n}\sum_{j=1}^{n}|a_{ij}|\,|x_j| \leq \max_{1\leq i\leq n}\sum_{j=1}^{n}|a_{ij}|,$$

avec égalité en prenant un indice i_0 où le max du terme de droite est atteint et en posant $x_j = \text{signe}\, a_{i_0,j}$ si $a_{i_0,j} \neq 0$, $x_j = 0$ sinon. ◊

Remarquons que pour toute norme matricielle subordonnée, on a $|||I||| = 1$. En fait, la norme matricielle subordonnée d'une matrice n'est rien d'autre que la norme d'application linéaire de l'application linéaire qu'elle définit sur $\mathbb{R}^n$ muni de sa base canonique. Il existe aussi des *normes matricielles*, c'est-à-dire des normes sur $M_n(\mathbb{R})$ qui vérifient l'inégalité de sous-multiplicativité $|||AB||| \leq |||A|||\,|||B|||$, mais qui ne sont pas nécessairement subordonnées à une norme sur $\mathbb{R}^n$.

A.3 L'ensemble des matrices inversibles est ouvert

Dans cette section, nous démontrons le lemme 5.2.2 qui assure que l'ensemble des matrices inversibles $GL_n(\mathbb{R})$ est un ouvert de $M_n(\mathbb{R})$.

Démonstration. Il suffit de montrer que pour toute matrice inversible $A \in GL_n(\mathbb{R})$, il existe une boule ouverte $B(A,r)$ incluse dans $GL_n(\mathbb{R})$. On traite d'abord le cas $A = I$. Pour tout $R \in M_n(\mathbb{R})$ et $N \in \mathbb{N}$, on a l'identité remarquable [9]

$$I - R^{N+1} = (I-R)(I + R + R^2 + \cdots + R^N).$$

D'après la proposition précédente, on a pour tout entier naturel i

$$|||R^i||| \leq |||R|||^i.$$

Supposons que $|||R||| < 1$, ce qui est équivalent à dire que $I - R \in B(I,1)$. On en déduit deux choses. D'une part, $R^{N+1} \to 0$ quand $N \to +\infty$, donc $I - R^{N+1} \to I$. D'autre part, la

9. Valable dans tout anneau unitaire.

suite $S_N = \sum_{i=0}^{N} R^i$ est une suite de Cauchy dans $M_n(\mathbb{R})$. En effet, par l'inégalité triangulaire et l'estimation ci-dessus

$$|||S_{N+m} - S_N||| \leq \sum_{i=N+1}^{N+m} |||R^i||| \leq \sum_{i=N+1}^{N+m} |||R|||^i = |||R|||^{N+1} \frac{1 - |||R|||^m}{1 - |||R|||} \leq \frac{|||R|||^{N+1}}{1 - |||R|||} \to 0$$

quand $N \to +\infty$. Or l'espace des matrices $M_n(\mathbb{R})$ est complet pour n'importe quelle norme, donc la suite S_N converge vers une matrice $S \in M_n(\mathbb{R})$. On peut alors passer à la limite quand $N \to +\infty$ dans l'identité remarquable ci-dessus. On obtient ainsi

$$I = (I - R)S.$$

En effet,

$$|||(I - R)S_N - (I - R)S||| \leq |||I - R||| \, |||S_N - S||| \to 0 \text{ quand } N \to +\infty.$$

Ceci montre que $I - R$ est inversible (avec $(I - R)^{-1} = S = \sum_{i=0}^{\infty} R^i$). On vient donc de montrer que $B(I, 1) \subset GL_n(\mathbb{R})$.

Revenons au cas général, avec $A \in GL_n(\mathbb{R})$. Pour tout $C \in M_n(\mathbb{R})$, on pose $R = A - C$. Comme A est inversible, en multipliant à gauche par A^{-1}, ceci se réécrit $A^{-1}C = I - A^{-1}R$. D'après l'étape précédente, si $|||A^{-1}R||| < 1$, alors on est assuré que $I - A^{-1}R = A^{-1}C$ est inversible. Ceci implique que $C = A(A^{-1}C)$ est également inversible. Il suffit donc que $|||R||| < \frac{1}{|||A^{-1}|||}$, pour que C soit inversible. En effet, dans ce cas, $|||A^{-1}R||| \leq |||R||| \, |||A^{-1}||| < 1$. En d'autres termes, on a montré que $B\big(A, \frac{1}{|||A^{-1}|||}\big) \subset GL_n(\mathbb{R})$, d'où le résultat avec $r = \frac{1}{|||A^{-1}|||}$. ◊

Remarque A.3.1 La preuve ci-dessus fonctionne pour n'importe quelle norme matricielle subordonnée, c'est-à-dire en fait pour n'importe quel choix de norme dans $\mathbb{R}^n$. Cela marche même pour toute norme matricielle. Donc finalement, la réunion de toutes les boules ouvertes de centre I et de rayon 1 pour toutes les normes matricielles possibles et imaginables est incluse dans $GL_n(\mathbb{R})$, et de même plus généralement pour $B\big(A, \frac{1}{|||A^{-1}|||}\big)$ dès que A est inversible. ◊

Annexe B

Rappels de calcul différentiel

Faisons d'abord quelques rappels de calcul différentiel. D'accord, c'est un prérequis, mais une petite révision expresse ne peut pas faire de mal. On se place systématiquement dans des espaces vectoriels E sur $\mathbb{R}$, avec des notions analogues sur $\mathbb{C}$ ou sur d'autres corps de nombres plus exotiques, munis d'une norme, c'est-à-dire d'une application de E dans $\mathbb{R}_+$ qui est positivement homogène, satisfait l'inégalité triangulaire et ne s'annule que sur le vecteur nul. Avec cette notion de norme viennent comme on l'a vu des notions de distance, de boules, d'ouverts, de fermés et de continuité. Notons que, pour toute norme, toutes les boules pour la distance associée sont convexes.

Une application g d'un ouvert U d'un espace vectoriel normé E à valeurs dans un autre espace vectoriel normé F est différentiable en un point $x \in U$ s'il existe une application linéaire continue de E dans F, notée df_x telle que l'on puisse écrire

$$g(x+h) = g(x) + dg_x h + \|h\|_E \varepsilon(h),$$

pour tout $h \in E$ tel que $x + h \in U$, où ε est une application de E dans F telle que

$$\|\varepsilon(h)\|_F \to 0 \text{ quand } \|h\|_E \to 0.$$

L'application linéaire dg_x est appelée la *différentielle* (de Fréchet) de g au point x. Le fait que cette application linéaire soit continue se traduit par l'existence d'une constante C telle que $\|dg_x h\|_F \leq C\|h\|_E$ pour tout $h \in E$. On voit que l'accroissement $g(x+h) - g(x)$ se comporte principalement comme le terme linéaire $dg_x h$, le terme suivant $\|h\|_E \varepsilon(h)$ se comportant comme un reste négligeable devant le terme linéaire quand $\|h\|_E$ est petit (sauf si le terme linéaire en question est nul...). Il est bien clair que si g est différentiable en x, alors elle est continue en x.

Quand g est différentiable en tout point $x \in U$, on dit qu'elle est différentiable sur U. Quand l'application $U \to \mathscr{L}(E;F)$, $x \mapsto dg_x$, est elle-même continue [1], on dit que g est continûment différentiable ou de classe C^1. Quand $E = F = \mathbb{R}$, alors dg_x est l'application linéaire de $\mathbb{R}$ dans $\mathbb{R}$ qui consiste à multiplier par la dérivée de g au point x : $dg_x h = g'(x)h$. Ne pas confondre la continuité de la différentielle en x, $h \mapsto dg_x h$, qui est automatique ici en dimension finie, et la continuité de $x \mapsto dg_x$ de $\mathbb{R}$ dans $\mathscr{L}(\mathbb{R};\mathbb{R})$, c'est-à-dire la classe C^1, qui n'est ici autre que la continuité de la fonction g' de $\mathbb{R}$ dans $\mathbb{R}$. En effet,

$$\|dg_x - dg_y\|_{\mathscr{L}(\mathbb{R};\mathbb{R})} = \sup\big\{|dg_x h - dg_y h|;\, |h| \leq 1\big\} = |g'(x) - g'(y)|.$$

1. L'espace $\mathscr{L}(E;F)$ étant lui-même un espace vectoriel normé, ceci a bien un sens. La norme naturelle sur cet espace est $\|g\|_{\mathscr{L}(E;F)} = \sup\{\|g(x)\|_F;\, \|x\|_E \leq 1\}$.

À toutes fins utiles, on rappelle qu'une dérivée partielle n'a rien de bien méchant. C'est juste une dérivée ordinaire par rapport à une des variables quand on fixe toutes les autres.

En dimension finie (de l'espace de départ), toute application linéaire est automatiquement continue, donc il est inutile de se focaliser sur la continuité de dg_x, qui est offerte gratuitement dans ce cas. Ce n'est qu'en dimension infinie qu'il faut payer un supplément pour cela. De plus, toujours en dimension finie, après un choix de base dans chacun des deux espaces vectoriels de départ et d'arrivée, la matrice qui représente la différentielle d'une application dans ces bases est appelée sa *matrice jacobienne*. Ses coefficients sont les dérivées partielles des différentes composantes.

Plus explicitement, soit $g\colon U \to F$ une application de U ouvert d'un espace vectoriel normé E de dimension k à valeurs dans un espace vectoriel normé F de dimension m. On suppose g différentiable au point $x_0 \in U$. Sa différentielle en x_0 est une application linéaire dg_{x_0} de E dans F. Si l'on choisit une base $(u_j)_{j=1,\dots,k}$ de E et une base $(v_i)_{i=1,\dots,m}$ de F, et que l'on note (x_j) les coordonnées cartésiennes associées dans E et (y_i) les coordonnées cartésiennes associées dans F, alors l'application g est représentée par m applications coordonnées g_i de l'ouvert de $\mathbb{R}^k$ contenant les coordonnées des points de U, à valeurs dans $\mathbb{R}$, de telle sorte que

$$g(x) = \sum_{i=1}^{m} g_i(x_1, x_2, \dots, x_k) v_i, \text{ où } x = \sum_{j=1}^{k} x_j u_j.$$

La différentielle dg_{x_0} de g en x_0 est alors représentée dans ces bases par la matrice jacobienne $\nabla g(x_0)$, matrice $m \times k$ dont les coefficients sont donnés par $(\nabla g(x_0))_{ij} = \frac{\partial g_i}{\partial x_j}(x_0)$, $i = 1, \dots, m$, $j = 1, \dots, k$. Cette représentation a lieu au sens usuel de l'algèbre linéaire, c'est-à-dire que pour tout vecteur $h = \sum_{j=1}^{k} h_j u_j$ de E, on a

$$dg_{x_0}h = \sum_{i=1}^{m} (dg_{x_0}h)_i v_i \quad \text{avec} \quad (dg_{x_0}h)_i = \sum_{j=1}^{k} (\nabla g(x_0))_{ij} h_j = \sum_{j=1}^{k} \frac{\partial f_i}{\partial x_j}(x_0) h_j.$$

On reconnaît un simple produit matrice-vecteur. C'est tout-à-fait normal, car si A est la matrice $m \times k$ qui représente l'application linéaire dg_{x_0} dans ces bases, on a $A_{ij} = (dg_{x_0}u_j)_i$ avec

$$g(x_0 + tu_j) = g(x_0) + t df_{x_0}u_j + |t| \|u_j\|_E \varepsilon(|t|),$$

d'où

$$\begin{aligned}(dg_{x_0}u_j)_i &= \left(\frac{g(x_0 + tu_j) - g(x_0)}{t}\right)_i - \left(\|u_j\|_E \varepsilon(|t|)\right)_i \\ &= \frac{g_i(x_0 + tu_j) - g_i(x_0)}{t} - \|u_j\|_E \varepsilon_i(|t|) \to \frac{\partial g_i}{\partial x_j}(x_0)\end{aligned}$$

quand $t \to 0$ par définition de ce qu'est une dérivée partielle. On voit bien sûr que le fait que g soit différentiable en x_0 implique que toutes ces dérivées partielles existent en x_0.

La composée de deux applications différentiables est différentiable. Si $g\colon U \to F$ est différentiable en $x_0 \in U \subset E$ et $g\colon V \to G$ est différentiable en $g(x_0) \in V \subset F$, U et V ouverts de leur espace respectif tels que $g(U) \subset V$, alors $g \circ f\colon U \to G$ est différentiable en x_0 et sa différentielle est la composée des différentielles de f et g, $d(g \circ f)_{x_0} = dg_{f(x_0)} \circ df_{x_0}$.

Avec des choix de bases dans les trois espaces vectoriels, comme la matrice de la composée de deux applications linéaires est le produit de leurs matrices (dans le même ordre), on en déduit pour les matrices jacobiennes $\nabla(g \circ f)(x_0) = \nabla g(f(x_0))\nabla f(x_0)$.

En explicitant tout cela avec des dérivées partielles, on obtient la très importante formule de dérivation des fonctions composées de plusieurs variables, qu'il faut absolument savoir appliquer quelles que soient les circonstances, même les plus adverses,

$$\frac{\partial(g \circ f)_l}{\partial x_j}(x_0) = \sum_{i=1}^{m} \frac{\partial g_l}{\partial y_i}(f(x_0))\frac{\partial f_i}{\partial x_j}(x_0),$$

pour $j = 1, \ldots, k$ et $l = 1, \ldots, n$ où n est la dimension de G.[2] C'est une application immédiate de la formule générale donnant les coefficients d'un produit matriciel en fonction des coefficients des matrices dont on effectue le produit. C'est de l'algèbre linéaire en fait (dont on ne saurait trop rappeler combien elle est fondamentale).

On a bien sûr la même chose avec la différentiabilité en tout point et avec la classe C^1 pour les fonctions composées.

Quand $\dim E = 1$, on rappelle que le théorème des accroissements finis, et plus généralement la formule de Taylor avec reste de Taylor-Lagrange, sont faux dès que $\dim F > 1$. On les remplace par des inégalités du même nom. L'inégalité des accroissements finis, avec $g\colon [a, b] \to F$,

$$\|g(b) - g(a)\|_F \leq \sup_{t\in[a,b]} \|g'(t)\|_F (b - a),$$

et l'inégalité de Taylor-Lagrange,

$$\left\|g(b) - \sum_{i=0}^{n} \frac{(b-a)^i}{i!} g^{(i)}(a)\right\|_F \leq \sup_{t\in[a,b]} \|g^{(n+1)}(t)\|_F \frac{(b-a)^{n+1}}{(n+1)!},$$

sous les hypothèses adéquates sur f. Les formules avec reste intégral restent par contre vraies

$$g(b) - g(a) = \int_a^b g'(t)\,dt,$$

et

$$g(b) - \sum_{i=0}^{n} \frac{(b-a)^i}{i!} g^{(i)}(a) = \int_a^b \frac{(t-a)^n}{n!} g^{(n+1)}(t)\,dt,$$

avec une notion adéquate d'intégrale à valeurs vectorielles, c'est-à-dire en dimension finie, en intégrant composante par composante.

Quand $\dim E > 1$, on se ramène à la dimension 1 en se plaçant sur des segments, à condition que ces segments restent entièrement dans U. Notons que l'inégalité des accroissements finis nous donne un moyen pratique de vérifier la condition de contraction stricte nécessaire pour pouvoir appliquer le théorème 4.2.2 de point fixe de Banach dans le cas de $\mathbb{R}^n$, que l'on munit de la norme que l'on préfère.

2. En fait, on a seulement besoin de se rappeler du cas $n = 1$, manifestement.

Proposition B.0.1 *Soit $g\colon \mathbb{R}^n \to \mathbb{R}^n$ une application de classe C^1 telle qu'il existe une boule $B(a,r)$ telle que*

$$\sup_{x\in\bar{B}(a,r)} \|dg_x\|_{\mathscr{L}(\mathbb{R}^n;\mathbb{R}^n)} = k < 1,$$

et

$$kr + \|f(a) - a\| \leq r,$$

alors $g(\bar{B}(a,r)) \subset \bar{B}(a,r)$, f y est strictement contractante.

Ici $\bar{B}(a,r) = \{x \in E; \|x-a\| \leq r\}$ désigne la boule fermée de centre a et de rayon r. On rappelle que

$$\|dg_x\|_{\mathscr{L}(\mathbb{R}^n;\mathbb{R}^n)} = \sup_{\substack{u\in\mathbb{R}^n \\ \|u\|\leq 1}} \|dg_x u\|.$$

Démonstration. Pour tous $x, y \in \bar{B}(a,r)$, on pose $h\colon [0,1] \to \mathbb{R}^n$, $h(t) = g(x + t(y-x))$ de telle sorte que $h(0) = g(x)$ et $h(1) = g(y)$. Par dérivation des fonctions composées, h est de classe C^1, avec

$$h'(t) = dg_{x+t(y-x)}(y-x).$$

Les propriétés de norme d'application linéaire impliquent que

$$\|h'(t)\| \leq \|dg_{x+t(y-x)}\|_{\mathscr{L}(\mathbb{R}^n;\mathbb{R}^n)}\|y-x\|.$$

Le segment $t \mapsto x + t(y-x)$ est entièrement contenu dans la boule qui est convexe, on en déduit donc que

$$\|h'(t)\| \leq k\|y-x\|.$$

L'inégalité des accroissements finis implique alors

$$\|g(y) - g(x)\| = \|h(1) - h(0)\| \leq \sup_{t\in[0,1]} \|h'(t)\|(1-0) \leq k\|y-x\|,$$

d'où la contraction stricte.

On utilise alors la deuxième hypothèse. Pour tout $x \in \bar{B}(a,r)$, c'est-à-dire $\|x-a\| \leq r$,

$$\begin{aligned}\|g(x) - a\| = \|g(x) - g(a) + g(a) - a\| &\leq \|g(x) - g(a)\| + \|g(a) - a\| \\ &\leq k\|x-a\| + \|g(a) - a\| \leq kr + \|g(a) - a\| \leq r,\end{aligned}$$

ce qui montre que $g(x) \in \bar{B}(a,r)$. ◊

Corollaire B.0.2 *Si g satisfait les hypothèses de la proposition précédente, alors g admet un point fixe unique dans la boule $\bar{B}(a,r)$.*

Démonstration. En effet, $\mathbb{R}^n$ est complet pour toute distance induite par une norme et une boule fermée est un fermé pour la topologie métrique associée à la norme. Elle est donc complète pour la distance en question. Ceci découle du fait que toute suite de Cauchy dans $\bar{B}(a,r)$ est aussi de Cauchy dans $\mathbb{R}^n$, puisqu'il s'agit de la même distance. Elle a donc une limite dans $\mathbb{R}^n$, mais comme $\bar{B}(a,r)$ est fermé, cette limite est dans $\bar{B}(a,r)$. Le théorème de point fixe de Banach s'applique donc dans l'espace métrique complet $\bar{B}(a,r)$ avec distance induite par la norme. ◊

Remarque B.0.1 1. Un examen rapide des arguments ci-dessus montre que la dimension finie n'y joue aucun rôle. En fait, la proposition est vrai dans exactement les mêmes termes dans un espace vectoriel normé E quelconque, et le corollaire est vrai dans un espace de Banach E quelconque.

2. Dans $\mathbb{R}^n$, toutes les normes sont équivalentes, donc le choix de la norme n'influe pas sur le caractère C^1 ou pas. Par contre, il influe sur la forme des boules ainsi que sur les valeurs numériques comme $\|dg_x\|_{\mathcal{L}(\mathbb{R}^n;\mathbb{R}^n)}$. Suivant ce choix de norme, la quantité $\|dg_x\|_{\mathcal{L}(\mathbb{R}^n;\mathbb{R}^n)}$ se calcule plus ou moins facilement, mais on peut très souvent au moins l'estimer. Le fait qu'on demande qu'elle soit majorée sur une boule par une constante strictement inférieure à 1 est à rapprocher de la notion de point fixe attractif vue en dimension $n = 1$. Ainsi, si a est un point fixe de g, alors la deuxième condition est automatiquement satisfaite. ◊

Annexe C

Rappels sur les équations différentielles ordinaires

C.1 Présentation générale

C.1.1 Systèmes différentiels d'ordre 1

Donnons maintenant quelques définitions de façon un peu plus formelle.

Définition C.1.1 *Soient m et n deux entiers non nuls. On se donne une application f de $\bar{I} \times (\mathbb{R}^m)^n$ dans $\mathbb{R}^m$. On appelle EDO d'ordre n de fonction second membre f, l'équation exprimant la dérivée $n^{ème}$, $y^{(n)}$, d'une fonction y définie sur l'intervalle $\bar{I}$ à valeurs dans $\mathbb{R}^m$, en fonction de ses dérivées d'ordre inférieur $y^{(i)}$, $i = 0, \ldots, n-1$, de la forme*

$$\forall t \in I, \quad y^{(n)}(t) = f(t, y(t), y'(t), \ldots, y^{(n-1)}(t)). \tag{C.1.1}$$

On appelle nombre de degrés de liberté *le produit mn.*

Une solution *du système* (C.1.1) *est une fonction de $\bar{I}$ à valeurs dans $\mathbb{R}^m$, n fois dérivable sur I et telle que l'égalité* (C.1.1) *soit satisfaite pour tout $t \in I$. L'image de $\bar{I}$ dans $(\mathbb{R}^m)^n$ par l'application $t \mapsto (y(t), y'(t), \ldots, y^{(n-1)}(t))$ est une* trajectoire *ou une* orbite *ou encore une* courbe de phase. *Le graphe d'une solution $\{t, y(t); t \in \bar{I}\} \subset \mathbb{R} \times \mathbb{R}^m$ est une* courbe intégrale.

Au lieu d'EDO d'ordre 1, d'ordre 2, on dit aussi du premier ordre, du second ordre, etc. L'égalité (C.1.1) est encore une égalité vectorielle entre vecteurs de $\mathbb{R}^m$. Par exemple, pour une équation d'ordre 2 à valeurs dans $\mathbb{R}^2$, le second membre est une application de $\bar{I} \times \mathbb{R}^2 \times \mathbb{R}^2$ à valeurs dans $\mathbb{R}^2$.

Notons immédiatement que la forme (2.1.1) ne concerne pas que les seules EDO du premier ordre, mais englobe également des équations d'ordre plus élevé. En fait, toute EDO d'ordre n peut se réécrire de façon canonique comme une EDO du premier ordre, et la généralité supplémentaire contenue dans la définition C.1.1 n'est qu'apparente.

Proposition C.1.2 *Soit y est une solution de* (C.1.1). *La fonction vectorielle $Y : \bar{I} \to (\mathbb{R}^m)^n$ définie par*

$$Y(t) = (y(t), y'(t), \ldots, y^{n-1}(t))$$

est alors solution du système d'équations différentielles du premier ordre

$$\forall t \in I, \quad \frac{dY(t)}{dt} = F(t, Y(t)), \tag{C.1.2}$$

où F est la fonction définie par

$$F\colon \bar{I}\times(\mathbb{R}^m)^n \longrightarrow (\mathbb{R}^m)^n$$
$$(t, Y_1, Y_2, \dots, Y_n) \longmapsto \big(Y_2, Y_3, \dots, Y_n, f(t, Y_1, Y_2, \dots, Y_n)\big),$$

les Y_i désignant des éléments génériques de $\mathbb{R}^m$.

Réciproquement, toute solution Y du système (C.1.2) *donne naissance à une solution de* (C.1.1) *en posant $y(t) = Y_1(t)$.*

Démonstration. Soit y est une solution de (C.1.1). On définit Y comme indiqué plus haut, c'est-à-dire que l'on pose $Y_i(t) = y^{(i-1)}(t)$ pour $i = 1, \dots, n$. Chacune des fonctions vectorielles Y_i, $i = 1, \dots, n$, constituant Y est donc une fonction de $\bar{I}$ dans $\mathbb{R}^m$, dérivable sur I. Par définition, on a $Y_i'(t) = \big(y^{(i-1)}\big)'(t) = y^{(i)}(t) = Y_{i+1}(t)$ pour tout $i = 1, \dots, n-1$. Pour la dernière fonction, on a

$$Y_n'(t) = \big(y^{(n-1)}\big)'(t) = y^{(n)}(t) = f\big(t, y(t), y'(t), \dots, y^{n-1}(t)\big) = f\big(t, Y_1(t), Y_2(t), \dots, Y_n(t)\big).$$

car y est solution de (C.1.1). On voit donc que Y est une solution du système (C.1.2).

Réciproquement, donnons-nous une solution Y de (C.1.2) et posons $y(t) = Y_1(t)$. Par définition, Y est dérivable sur I, donc chaque Y_i est dérivable. En particulier, $y = Y_1$ est dérivable. Montrons par récurrence que y est n fois dérivable sur I et que $y^{(i-1)} = Y_i$ pour tout $i = 1, \dots, n$. La propriété étant déjà établie pour $i = 1$, supposons la vraie pour $i < n$. On a vu que Y_i est dérivable, ce qui implique que $y^{(i-1)}$ est dérivable, ou encore que y est i fois dérivable avec $y^{(i)} = \big(y^{(i-1)}\big)' = Y_i' = Y_{i+1}$ en utilisant le système (C.1.2) pour $i < n$. La récurrence est donc achevée.

Pour conclure, on note que

$$y^{(n)}(t) = Y_n'(t) = f\big(t, Y_1(t), Y_2(t), \dots, Y_n(t)\big) = f\big(t, y(t), y'(t), \dots, y^{n-1}(t)\big),$$

d'après le système (C.1.2) pour $i = n$ et d'après la récurrence qui précède. La fonction y est par conséquent une solution de (C.1.1). ◊

Il faut faire attention que dans la proposition précédente, si l'on souhaite écrire les choses en composantes, on a besoin de mn fonctions scalaires pour définir la fonction vectorielle Y, c'est-à-dire le nombre de degrés de liberté. Un cas particulier important est celui des équations différentielles d'ordre n scalaires, *i.e.*, avec $m = 1$, qui se réécrivent canoniquement comme un système d'ordre 1 à n équations et n inconnues scalaire, si le besoin s'en fait sentir.

C.1.2 Le cas des équations différentielles autonomes

Définition C.1.3 *On appelle* équation différentielle autonome *une EDO dont le second membre ne dépend pas explicitement du temps.*

$$y^{(n)}(t) = f(y(t), y'(t), \dots, y^{(n-1)}(t)). \tag{C.1.3}$$

Définition C.1.4 *On appelle* espace des phases *l'espace $\mathbb{R}^{mn}$ où se trouvent les trajectoires des solutions de l'équation différentielle* (C.1.3) *ramenée à un système du premier ordre $Y'(t) = F(Y(t))$ avec $Y(t) = (y(t), y'(t), \dots, y^{(n-1)}(t))$.*

Le champ de vecteurs de vitesses des phases est l'image de l'espace des phases par la fonction second membre F.

Sans même résoudre l'équation différentielle, on peut comprendre beaucoup de choses sur ses solutions en traçant le champ de vitesses des phases dans l'espace des phases, ce qui est possible si $mn \leq 3$. Une étude des EDO utilisant pleinement ce point de vue est le livre d'Arnold [1] [1].

Définition C.1.5 *On définit les* points d'équilibre *ou* points fixes *ou* points stationnaires *d'un système différentiel autonome $y'(t) = f(y(t))$, comme les points $y_e \in \mathbb{R}^m$ tels que $f(y_e) = 0$.*

Par l'unicité de Cauchy-Lipschitz, si une solution se trouve à un instant t en un point d'équilibre, elle y reste éternellement, et d'ailleurs, elle y était déjà auparavant. Une fois déterminé un point d'équilibre, on s'intéresse souvent au comportement de la solution dans son voisinage. Partant d'une condition initiale proche de y_e, va-t-elle s'en rapprocher ou s'en éloigner ? C'est la notion de stabilité d'un point fixe, qui se décline en plusieurs variantes.

Définition C.1.6 *Le point d'équilibre y_e de $y'(t) = f(y(t))$ est*

— *stable si, pour tout $\varepsilon > 0$, il existe $r > 0$, tel que*

$$\|y(0) - y_e\| < r \Rightarrow \forall t > 0, \|y(t) - y_e\| < \varepsilon.$$

— *instable sinon.*

— *asymptotiquement stable, s'il est stable et si r peut être choisi tel que*

$$\|y(0) - y_e\| < r \Rightarrow \lim_{t\to\infty} \|y(t) - y_e\| = 0.$$

— *marginalement stable s'il n'est pas asymptotiquement stable et que la solution est bornée.*

Si le système est asymptotiquement stable quelle que soit la donnée initiale $y(0)$, alors le point d'équilibre est dit être globalement asymptotiquement (ou exponentiellement) stable.
Remarquons au passage que toute équation non autonome peut être rendue autonome en augmentant la dimension de l'espace d'une unité. En effet, si $y(t) = \big(y_1(t), y_2(t), \ldots, y_m(t)\big)$ est solution du problème de Cauchy $y'(t) = f(t, y(t))$, $y(0) = y_0$, alors posant

$$Y(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_m(t) \\ y_{m+1}(t) \end{pmatrix} \text{ et } F(Y) = \begin{pmatrix} f_1(y_{m+1}, y) \\ f_2(y_{m+1}, y) \\ \vdots \\ f_m(y_{m+1}, y) \\ 1 \end{pmatrix},$$

on voit que Y est solution du problème de Cauchy autonome $Y'(t) = F(Y(t))$, $Y(0) = (y_0, 0)$, et réciproquement. En ce sens, on dispose donc aussi d'un espace des phases pour une équation non autonome du premier ordre à valeurs dans $\mathbb{R}^m$, qui est $\mathbb{R}^{m+1}$, le champ de vecteurs ayant pour dernière composante 1, et chaque tranche $y_{m+1} = t$ contenant le champ de vecteurs dans $\mathbb{R}^m$ à l'instant t.

C.2 Équations différentielles linéaires

Dans cette section, on notera $M_m(\mathbb{R})$ l'espace des matrices carrées $m \times m$ à coefficients réels et $M_m(\mathbb{C})$ celui des matrices carrées $m \times m$ à coefficients complexes.

1. Vladimir Igorevich Arnold, 1937–2010.

C.2.1 Définitions et propriétés générales

Définition C.2.1 *On appelle équation différentielle linéaire sur l'intervalle I, toute équation différentielle de la forme*

$$\forall t \in I, \quad y'(t) = A(t)y(t) + b(t), \tag{C.2.1}$$

où

$$t \mapsto A(t) = (a_{ij}(t))_{1\leq i,j\leq m} \in M_m(\mathbb{R}), \quad t \mapsto b(t) = \begin{pmatrix} b_1(t) \\ \vdots \\ b_m(t) \end{pmatrix} \in \mathbb{R}^m$$

sont des fonctions continues sur $\bar{I}$ données, respectivement à valeurs dans $M_m(\mathbb{R})$ et $\mathbb{R}^m$.

La linéarité de l'équation vient du fait que la partie du second membre qui dépend de y est linéaire par rapport à y : $f(t, y) = A(t)y + b(t)$. On voit que c'est plutôt affine que linéaire, mais peu importe. On parle aussi bien sûr de système différentiel linéaire quand $m > 1$. Le problème de Cauchy prend naturellement la forme

$$\begin{cases} y'(t) = A(t)y(t) + b(t), \\ y(0) = y_0. \end{cases} \tag{C.2.2}$$

Dans le cas scalaire, $m = 1$, on sait depuis la première année d'université au moins, résoudre les EDO linéaires par la *méthode de variation de la constante*. Les matrices $A(t)$ sont des matrices 1×1 que l'on assimile à leur unique coefficient, les scalaires $a_{11}(t) = a(t)$. Dans ce contexte, la fonction b est aussi à valeurs scalaires, $b_1(t) = b(t)$ avec la même assimilation inoffensive.

Théorème C.2.2 (Variation de la constante) *Étant données deux fonctions continues a et b de $\bar{I}$ dans $\mathbb{R}$, il existe une unique solution du problème de Cauchy scalaire*

$$\begin{cases} y'(t) = a(t)y(t) + b(t), \\ y(0) = y_0, \end{cases}$$

laquelle est donnée par

$$\forall t \in \bar{I}, \quad y(t) = y_0 e^{\int_0^t a(s)ds} + \int_0^t b(u) e^{\int_u^t a(s)\,ds}\, du. \tag{C.2.3}$$

En particulier, si la fonction a est constante et $b = 0$, on retrouve bien $y(t) = y_0 e^{at}$.

Démonstration. Rappelons la méthode, qui est non seulement une méthode de calcul, mais aussi une démonstration de l'existence et l'unicité pour le problème de Cauchy dans ce cas très simple.

Étape 1 : Cas $b = 0$. On regarde l'équation $y'(t) = a(t)y(t)$. C'est une équation à variable séparées, mais on ne va pas procéder à la physicienne pour éviter les critiques justifiées des mathématiciens. Soit $\mathcal{A}(t) = \int_0^t a(s)\,ds$ une primitive de a sur $\bar{I}$, il en existe puisque a est continue. [2] On réécrit l'équation sous la forme $y'(t) - a(t)y(t) = 0$ que l'on multiplie par le *facteur intégrant* $e^{-\mathcal{A}(t)}$. On obtient de la sorte

$$0 = e^{-\mathcal{A}(t)}\big(y'(t) - a(t)y(t)\big) = \big(e^{-\mathcal{A}(t)}y(t)\big)',$$

et la fonction $t \mapsto e^{-\mathcal{A}(t)}y(t)$ est donc constante sur I, égale à un certain $K \in \mathbb{R}$. Par conséquent,

$$y(t) = Ke^{\mathcal{A}(t)},$$

pour tout t dans ce cas $b = 0$.

Étape 2 : On n'est plus à variables séparées, pas de salut à chercher du côté de la physique. L'idée est de chercher une solution de l'équation complète avec $b \neq 0$ en faisant varier la constante K, c'est-à-dire en injectant la forme $y(t) = K(t)e^{\mathcal{A}(t)}$ dans l'équation complète et en considérant K non plus comme une constante, mais comme une nouvelle fonction inconnue [3]. Comme l'exponentielle ne s'annule jamais, c'est un changement

2. On prend ici celle qui s'annule en 0, mais c'est juste pour fixer les idées, ce n'est en rien nécessaire.
3. D'où le nom de la méthode.

de fonction inconnue complètement légitime. Une autre façon de le dire, mais dont il est moins facile de se rappeler, est de poser simplement $K(t) = e^{-\mathscr{A}(t)}y(t)$.

On écrit donc [4]

$$y'(t) = K'(t)e^{\mathscr{A}(t)} + K(t)a(t)e^{\mathscr{A}(t)} = a(t)y(t) + b(t) = a(t)K(t)e^{\mathscr{A}(t)} + b(t),$$

les termes du milieu se simplifient (s'ils ne se simplifient pas, c'est que l'on s'est trompé dans les calculs) et il vient donc

$$K'(t)e^{\mathscr{A}(t)} = b(t) \quad \text{d'où} \quad K'(t) = b(t)e^{-\mathscr{A}(t)},$$

d'où en intégrant, ce qui ne pose pas de problème puisque b est continue,

$$K(t) = K(0) + \int_0^t b(u)e^{-\mathscr{A}(u)}\,du.$$

Multipliant par $e^{\mathscr{A}(t)}$, il vient alors

$$y(t) = K(0)e^{\mathscr{A}(t)} + e^{\mathscr{A}(t)}\int_0^t b(u)e^{-\mathscr{A}(u)}\,du = K(0)e^{\mathscr{A}(t)} + \int_0^t b(u)e^{\mathscr{A}(t)-\mathscr{A}(u)}\,du,$$

et utilisant la condition initiale $y(0) = y_0$ et le fait que $\mathscr{A}(t) - \mathscr{A}(u) = \int_u^t a(s)\,ds$, on retrouve bien l'expression (C.2.3).

On a jusqu'ici établi *l'unicité* de la solution du problème de Cauchy : toute solution éventuelle ne peut être que de la forme (C.2.3). Pour établir *l'existence*, il suffit de vérifier que cette expression satisfait bien l'EDO d'une part et la condition initiale d'autre part, ce qui n'est qu'un calcul de routine. En effet, pour la condition initiale, on a

$$y(0) = y_0 e^{\int_0^0 a(s)ds} + \int_0^0 b(u)e^{\int_u^0 a(s)\,ds}\,du = y_0,$$

car $\int_0^0(\text{n'importe quoi})\,ds = 0$ et $e^0 = 1$. Pour l'EDO elle-même, il est plus simple de prendre la forme $y(t) = y_0 e^{\int_0^t a(s)ds} + e^{\int_0^t a(s)ds}\int_0^t b(u)e^{-\int_0^u a(s)\,ds}\,du$ et il vient alors

$$\begin{aligned} y'(t) &= y_0 a(t)e^{\int_0^t a(s)ds} + a(t)e^{\int_0^t a(s)ds}\int_0^t b(u)e^{-\int_0^u a(s)\,ds}\,du + e^{\int_0^t a(s)ds}b(t)e^{-\int_0^t a(s)\,ds} \\ &= a(t)y(t) + b(t), \end{aligned}$$

par la formule de Leibniz de dérivation d'un produit, la formule de dérivation des fonctions composées et le fait que quand on dérive une intégrale par rapport à sa borne supérieure, on obtient la valeur de l'intégrande en cette même borne supérieure. ♢

Naturellement, la variation de la constante n'aboutit à une résolution analytique complète du problème de Cauchy que si les deux calculs de primitives intermédiaires sont possibles analytiquement. Elle montre néanmoins, comme on l'a déjà dit parce que c'est en fait le plus important, *l'existence* et *l'unicité* du problème de Cauchy dans ce cas très particulier, même quand les calculs analytiques de primitives sont impossibles, sous la seule hypothèse que les fonctions a et b soient continues.

Voyons maintenant ce que l'on peut dire dans le cas vectoriel avec m pas nécessairement égal à 1. Pas grand-chose pour l'instant.

Définition C.2.3 *Étant donnée une équation différentielle linéaire* (C.2.1)*, on appelle équation différentielle linéaire sans second membre ou homogène associée, l'équation différentielle*

$$\forall t \in I, \quad y'(t) = A(t)y(t). \tag{C.2.4}$$

4. Si l'on est allergique au caractère peut-être un peu parachuté de la méthode de variation de la constante (parachuté mais mnémotechnique !), on peut aussi écrire $K' = -ae^{-\mathscr{A}}y + e^{-\mathscr{A}}y' = -ae^{-\mathscr{A}}y + e^{-\mathscr{A}}(ay + b)$.

Le vocabulaire " sans second membre ", bien que traditionnel, est plutôt mal choisi puisqu'il y a une fonction second membre au sens antérieur, $f(t,y) = A(t)y$. On préférera le qualificatif homogène.

Commençons par énoncer deux propriétés générales des systèmes différentiels linéaires qui sont immédiates mais néanmoins très importantes pour la suite.

Proposition C.2.4 *Si y_1 et y_2 sont deux solutions de l'équation différentielle* (C.2.1)*, leur différence $y_2 - y_1$ est solution de l'équation différentielle homogène associée* (C.2.4).

Proposition C.2.5 *L'ensemble des solutions de l'équation différentielle homogène* (C.2.4) *est un espace vectoriel sur $\mathbb{R}$.*

On en déduit que

Proposition C.2.6 *L'ensemble des solutions de l'équation différentielle* (C.2.1) *forme un sous-espace affine*[5] *de l'espace vectoriel des applications de I dans $\mathbb{R}^m$.*

Démonstration. Soit S l'ensemble des solutions de l'équation différentielle (C.2.1). Pour tout $\lambda \in \mathbb{R}$ et $(y_1, y_2) \in S^2$, on a

$$\begin{aligned}\frac{d}{dt}\big(\lambda y_1(t) + (1-\lambda)y_2(t)\big) &= \lambda\big(A(t)y_1(t) + b(t)\big) + (1-\lambda)\big(A(t)y_2(t) + b(t)\big)\\ &= A(t)\big(\lambda y_1(t) + (1-\lambda)y_2(t)\big) + b(t).\end{aligned}$$

Par conséquent, $\lambda y_1 + (1-\lambda)y_2 \in S$. ◊

Les propositions C.2.4 à C.2.6 ne sont qu'une traduction en langage savant du *principe de superposition* des physiciens et de l'adage populaire " la solution générale est la somme d'une solution particulière et de la solution générale de l'équation homogène ". Dans la pratique, on appliquera l'adage populaire.

Intéressons-nous maintenant aux équations différentielles les plus simples parmi les plus simples.

C.2.2 Systèmes différentiels linéaires à coefficients constants

On s'intéresse dans ce paragraphe au cas où la matrice A apparaissant au second membre de l'EDO est indépendante de t. C'est à cette matrice que fait allusion l'expression " à coefficients constants ". La partie en $b(t)$ va rester variable en fonction du temps. Il va en fait être beaucoup confortable de travailler dans $\mathbb{C}^m$ que dans $\mathbb{R}^m$, donc avec des matrices complexes, des vecteurs complexes et des EDO à valeurs complexes, ce qui ne pose évidemment pas de problème de principe. L'objection : " Oui, mais je m'intéresse à une EDO à valeurs réelles, et je trouve des solutions à valeurs complexes ! C'est grave, docteur ? ", ne tient pas vraiment.

En effet, supposons que l'on ait une EDO linéaire avec une matrice A réelle et une fonction b à valeurs réelles, mais que l'on ait d'une façon ou d'une autre trouvé une solution $Y : \bar{I} \to \mathbb{C}^m$. En décomposant cette solution en partie réelle et partie imaginaire $Y(t) = \Re(Y(t)) + i\Im(Y(t))$, les deux fonctions à valeurs dans $\mathbb{R}^m$ ainsi obtenues sont solutions de la même EDO à valeurs réelles et l'on retombe sur ses pieds. Dans certaines applications, comme en électricité par exemple, il est même franchement avantageux de ne travailler qu'avec les solutions complexes.

On munit donc $\mathbb{C}^m$ de la norme hermitienne standard, qui se réduit sur $\mathbb{R}^m$ à la norme euclidienne standard, $\|x\| = \big(\sum_{i=1}^m |x_i|^2\big)^{1/2}$ avec $|x_i|^2 = x_i\bar{x}_i$. On rappelle que l'espace $M_m(\mathbb{C})$ est alors muni de la norme matricielle subordonnée définie par

$$\|A\| = \sup_{x\in\mathbb{C}^m\setminus\{0\}} \frac{\|Ax\|}{\|x\|}.$$

On en déduit que

$$\|Ax\| \le \|A\|\|x\|, \tag{C.2.5}$$

pour tout vecteur x. Il s'agit d'une norme matricielle au sens où pour tout couple de matrices A et B, on a $\|AB\| \le \|A\|\|B\|$. On en déduit immédiatement par récurrence que $\|A^k\| \le \|A\|^k$ pour tout entier positif k.

On aura besoin de la notion d'exponentielle de matrices.

5. On dit qu'une partie S d'un espace vectoriel est un sous-espace affine si elle est stable par combinaison barycentrique, autrement dit si elle vérifie la condition $\forall u, v \in S, \forall \lambda \in \mathbb{R}, \lambda u + (1-\lambda)v \in S$. Un sous-espace affine est ce que l'on obtient en translatant un sous-espace vectoriel par un vecteur constant.

Proposition C.2.7 *Pour toute matrice $A \in M_m(\mathbb{C})$, la série $\sum_{k=0}^{\infty} \frac{1}{k!} A^k$ est convergente dans $M_m(\mathbb{C})$.*

Démonstration. Regardons la série des normes associée $\sum_{k=0}^{\infty} \frac{1}{k!} \|A^k\|$. C'est une série à termes positifs majorée, donc convergente, puisque

$$\sum_{k=0}^{n} \frac{1}{k!} \|A^k\| \leq \sum_{k=0}^{n} \frac{1}{k!} \|A\|^k \leq \sum_{k=0}^{\infty} \frac{1}{k!} \|A\|^k = e^{\|A\|} < +\infty,$$

pour tout $n \in \mathbb{N}$. La série matricielle à laquelle on a affaire est donc une série normalement convergente. Comme l'espace des matrices est évidemment complet pour la norme précédente car de dimension finie, on en déduit que la série est convergente dans $M_m(\mathbb{C})$. ◊

On est donc fondé à poser la définition suivante.

Définition C.2.8 *Pour tout matrice $A \in M_m(\mathbb{C})$, on appelle* exponentielle de *A la somme de la série*

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k \in M_m(\mathbb{C}). \tag{C.2.6}$$

La démonstration de la proposition C.2.7 et l'inégalité triangulaire nous donnent gratuitement l'estimation $\left\|e^A\right\| \leq e^{\|A\|}$. On note aussi parfois l'exponentielle $\exp(A)$. Dans le cas $m = 1$, A s'identifie à un nombre complexe a et l'on retrouve la définition classique de l'exponentielle complexe comme somme d'une série numérique. Notons que la définition s'applique a fortiori aux matrices réelles : si $A \in M_m(\mathbb{R})$ alors $e^A \in M_m(\mathbb{R})$ puisque la série ne comprend que des termes réels.

On rappelle quelques propriétés utiles de l'exponentielle de matrice.

Théorème C.2.9 *Si A et $B \in M_m(\mathbb{C})$ commutent alors*

$$e^{A+B} = e^A e^B.$$

Attention, il est facile de trouver deux matrices A et B qui ne commutent pas et telles que $e^{A+B} \neq e^A e^B$! [6] La formule de Baker ?Campbell ?Hausdorff [7] permet de relier e^{A+B} à e^A, e^B puis une infinité d'autre termes faisant intervenir des commutateurs de commutateurs... Le commutateur de A et B est donné par $[A, B] = AB - BA$.

Notons aussi que si A et B commutent, alors e^A et e^B commutent aussi, comme conséquence immédiate du théorème C.2.9.

Le théorème C.2.9 a plusieurs conséquences utiles.

Corollaire C.2.10 *Soit $A \in M_m(\mathbb{C})$. On a*
i) e^A est inversible et $(e^A)^{-1} = e^{-A}$.
ii) L'application $t \mapsto e^{tA}$ est dérivable de $\mathbb{R}$ dans $M_n(\mathbb{C})$ et $(e^{tA})' = Ae^{tA} = e^{tA}A$.

Rappelons à toutes fins utiles qu'une fonction g continue de $[0, T]$ à valeurs dans $\mathbb{R}^m$ ou $\mathbb{C}^m$ s'intègre sur tout sous-intervalle $[0, t]$, au sens de Riemann [8] par exemple, tout simplement composante par composante. Son intégrale n'est autre que le vecteur

$$\int_0^t g(s)\, ds = \begin{pmatrix} \int_0^t g_1(s)\, ds \\ \vdots \\ \int_0^t g_m(s)\, ds \end{pmatrix}.$$

Au vu de cette expression, il doit être bien clair que $t \mapsto \int_0^t g(s)\, ds$ est une fonction dérivable et que $\frac{d}{dt}\left(\int_0^t g(s)\, ds\right) = g(t)$, c'est-à-dire que c'est la primitive de g qui s'annule en 0.

6. C'est anecdotique, mais on peut aussi, si on cherche bien, trouver quelques matrices A et B qui ne commutent pas, mais pour lesquelles on a quand même $e^{A+B} = e^A e^B$...

7. Henry Frederick Baker, 1866–1956 ; John Edward Campbell, 1862–1924 ; Felix Hausdorff, 1868–1942.

8. Georg Friedrich Bernhard Riemann, 1826–1866.

Naturellement, la linéarité de l'intégrale à valeurs scalaires persiste pour les intégrales à valeurs vectorielles. Si $B = (b_{ij})$ est une matrice $p \times m$ constante, alors on a $B\big(\int_0^t g(s)\,ds\big) = \int_0^t Bg(s)\,ds$. Il suffit en effet d'écrire les composantes pour $i = 1$ à p

$$\left(B\left(\int_0^t g(s)\,ds\right)\right)_i = \sum_{j=1}^m B_{ij}\left(\int_0^t g(s)\,ds\right)_j$$
$$= \sum_{j=1}^m B_{ij}\int_0^t g_j(s)\,ds = \int_0^t \left(\sum_{j=1}^m B_{ij}g_j(s)\right)ds = \int_0^t \big(Bg(s)\big)_i\,ds.$$

Cette linéarité permet d'ailleurs de voir que l'objet intégrale à valeur vectorielle défini plus haut ne dépend pas de la base choisie pour le calculer. On a donc tout aussi facilement des intégrales à valeurs dans un espace vectoriel de dimension finie quelconque.

Une propriété cruciale de l'intégrale que l'on utilisera souvent, est une généralisation de l'inégalité triangulaire

$$\left\|\int_0^t g(s)\,ds\right\| \le \int_0^t \|g(s)\|\,ds,$$

valable pour n'importe quelle norme sur $\mathbb{R}^m$ ou $\mathbb{C}^m$. Attention, à gauche on a la norme d'une intégrale vectorielle, alors qu'à droite, on a l'intégrale scalaire de la norme. Une démonstration rapide de cette inégalité, démonstration qui par contre ne marche que pour une norme euclidienne ou hermitienne, consiste à dire qu'il n'y a rien à montrer si $\int_0^t g(s)\,ds = 0$ et de poser sinon $u = \int_0^t g(s)\,ds / \big\|\int_0^t g(s)\,ds\big\|$, lequel est un vecteur unitaire tel que

$$\left\|\int_0^t g(s)\,ds\right\| = \left(\int_0^t g(s)\,ds\middle|u\right) = \int_0^t (g(s)|u)\,ds \le \int_0^t \|g(s)\|\,ds,$$

par linéarité de l'intégrale et l'inégalité de Cauchy-Schwarz [9] pour terminer.

Par ailleurs, si $t \mapsto A(t)$ est une fonction à valeurs matricielle dérivable et $t \mapsto z(t)$ est une fonction à valeurs vectorielles dérivable, alors le produit matrice-vecteur $t \mapsto A(t)z(t)$ est dérivable à valeurs vectorielles et la formule de Leibniz reste valable $\big(A(t)z(t)\big)' = A'(t)z(t) + A(t)z'(t)$ en faisant attention à l'ordre des facteurs. [10] En effet, $(A(t)z(t))_i = \sum_{j=1}^m a_{ij}(t)z_j(t)$ pour tout i, donc

$$(A(t)z(t))_i' = \sum_{j=1}^m \big(a_{ij}(t)z_j(t)\big)' = \sum_{j=1}^m \big(a_{ij}'(t)z_j(t) + a_{ij}(t)z_j'(t)\big) = (A'(t)z(t))_i + (A(t)z'(t))_i.$$

L'application de l'exponentielle de matrice aux EDO linéaires à coefficients constants se lit dans le résultat suivant, qui est l'analogue en dimension m de la variation de la constante en dimension 1 (dans le cas linéaire à coefficient constant).

Proposition C.2.11 *Soit $A \in M_m(\mathbb{C})$ quelconque et $b\colon \mathbb{R} \to \mathbb{C}^m$ continue. Le problème de Cauchy*

$$\begin{cases} y'(t) = Ay(t) + b(t), \\ y(0) = y_0, \end{cases}$$

admet une solution unique, laquelle s'écrit à l'aide de la formule de Duhamel

$$y(t) = e^{tA}y_0 + \int_0^t e^{(t-s)A}b(s)\,ds, \tag{C.2.7}$$

pour tout $t \in \mathbb{R}$.

Démonstration. On procède par condition nécessaire et condition suffisante. Condition nécessaire : si y est une solution, posons $z(t) = e^{-tA}y(t)$. On a donc

$$z'(t) = (e^{-tA})'y(t) + e^{-tA}y'(t) = -e^{-tA}Ay(t) + e^{-tA}(Ay(t) + b(t)) = e^{-tA}b(t),$$

9. Hermann Amandus Schwarz, 1843–1921.

10. On a bien sûr le même résultat pour la dérivée d'un produit de matrices : $(AB)' = A'B + AB'$.

d'après le corollaire C.2.10 ii). Comme $z(0) = e^0 y(0) = Iy_0 = y_0$, on en déduit que

$$z(t) = y_0 + \int_0^t e^{-sA} b(s)\, ds.$$

Par le corollaire C.2.10 i), on a $y(t) = e^{tA} z(t)$, d'où la formule (C.2.7) (on peut entrer et sortir à volonté e^{tA} de l'intégrale par linéarité, comme on vient de le voir plus haut, et bien sûr tA et $-sA$ commutent). On a ainsi montré l'unicité : s'il existe une solution, elle est forcément donnée par cette formule.

Montrons maintenant l'existence, c'est-à-dire la condition suffisante. Il faut montrer que la fonction y définie par la formule (C.2.7) est bien solution du problème de Cauchy de départ. Elle est manifestement dérivable, telle que $y(0) = y_0$. Pour calculer sa dérivée, on note que

$$y(t) = e^{tA} y_0 + e^{tA} \left(\int_0^t e^{-sA} b(s)\, ds \right),$$

il vient donc

$$y'(t) = Ae^{tA} y_0 + Ae^{tA} \left(\int_0^t e^{-sA} b(s)\, ds \right) + e^{tA} e^{-tA} b(t) = Ay(t) + b(t),$$

encore d'après le corollaire C.2.10 i) et ii). On a donc bien obtenu ainsi une solution du problème de Cauchy. ◊

La même formule reste naturellement a fortiori valable pour A, y_0 et b réels.

Corollaire C.2.12 *L'espace vectoriel des solutions de l'équation différentielle homogène $y'(t) = Ay(t)$ est de dimension m sur $\mathbb{C}$.*

Démonstration. Soit S l'espace vectoriel des solutions. Considérons l'application $\mathbb{C}^m \to S$, $y_0 \mapsto (t \mapsto e^{tA} y_0)$. C'est une application qui est trivialement linéaire. Elle est injective, car $e^{tA} y_0 = 0$ pour tout t implique que $y_0 = 0$ en prenant $t = 0$. La proposition C.2.11 dans le cas $b = 0$ implique par ailleurs qu'elle est surjective puisque toute solution s'écrit ainsi. Il s'agit donc d'un isomorphisme, et l'on en déduit que $\dim S = \dim \mathbb{C}^m = m$. ◊

Quand on travaille sur $\mathbb{R}$, on a le même résultat, mais la dimension m ci-dessus est alors à comprendre comme dimension de $\mathbb{R}$-espace vectoriel.

La question se pose maintenant de comment calculer l'exponentielle d'une matrice. On rappelle d'abord à ce propos la *décomposition de Dunford* [11] des matrices.

Théorème C.2.13 *Pour toute matrice $A \in M_m(\mathbb{C})$, il existe un unique couple de matrices (D, N) de $M_m(\mathbb{C})$ avec D diagonalisable, N nilpotente, D et N commutent et*

$$A = D + N. \tag{C.2.8}$$

On rappelle que si D est diagonalisable, il existe Δ diagonale et une matrice de passage [12] P, donc inversible, telles que

$$D = P^{-1} \Delta P.$$

Attention, on parle ici de diagonalisabilité sur $\mathbb{C}$ même quand les matrices sont réelles.

On rappelle également qu'une matrice N est nilpotente s'il existe un entier $p > 0$ tel que $N^p = 0$. Le plus petit entier p qui a cette propriété s'appelle l'*indice de nilpotence* de N. Notons, par l'unicité de la décomposition de Dunford, que A est elle-même diagonalisable si et seulement si $N = 0$.

11. Nelson Dunford, 1906–1986.

12. Dans ce paragraphe, la matrice P^{-1} est la matrice dont les colonnes sont les composantes d'une base de vecteurs propres. Il semble que quelques dérives de l'enseignement aient inculqué dans certains esprits la formule $P\Delta P^{-1}$, pourtant pleine de disharmonie... Dans le contexte présent, cela n'a aucune importance, il s'agit d'un jeu d'écriture et on préférera $P^{-1}\Delta P$ qui est beaucoup plus élégant, il faut bien l'admettre.

Proposition C.2.14 *Pour toute matrice $A \in M_m(\mathbb{C})$, soit (D, N) sa décomposition de Dunford, $\lambda_i \in \mathbb{C}$, $i = 1, \dots, m$, les valeurs propres [13] de D, P la matrice de passage et p l'indice de nilpotence de N. Alors on a*

$$e^A = e^D e^N = e^N e^D,$$

avec

$$e^D = P^{-1}\mathrm{diag}\big(e^{\lambda_i}\big)P$$

et $e^N = \sum_{k=0}^{p-1} \frac{1}{k!} N^k$ est un polynôme de degré $p-1$ en N.

On a noté ici $\mathrm{diag}(\mu_i)$ la matrice diagonale dont les coefficients diagonaux sont $\mu_1, \mu_2, \dots, \mu_m$.

Démonstration. Comme D et N commutent, la première formule résulte immédiatement de la décomposition de Dunford (C.2.8) et du théorème C.2.9.

Pour la partie diagonalisable, il est évident que $\big(\mathrm{diag}(\lambda_i)\big)^k = \mathrm{diag}(\lambda_i^k)$ pour tout $k \in \mathbb{N}$. Par conséquent, comme une somme de matrices diagonales est aussi évidemment diagonale, en passant à la limite dans les sommes partielles

$$\sum_{k=0}^{n} \frac{1}{k!}\mathrm{diag}(\lambda_i^k) = \mathrm{diag}\left(\sum_{k=0}^{n} \frac{\lambda_i^k}{k!}\right)$$

quand $n \to +\infty$, on voit que

$$e^\Delta = \sum_{k=0}^{\infty} \frac{1}{k!}\mathrm{diag}(\lambda_i^k) = \mathrm{diag}\left(\sum_{k=0}^{\infty} \frac{\lambda_i^k}{k!}\right) = \mathrm{diag}\big(e^{\lambda_i}\big).$$

Comme par ailleurs, $D = P^{-1}\Delta P$, il vient que pour tout entier $k \in \mathbb{N}$, $D^k = P^{-1}\Delta^k P$. Par conséquent, en factorisant à gauche par P^{-1} et à droite par P,

$$e^D = \sum_{k=0}^{\infty} \frac{1}{k!} P^{-1}\Delta^k P = P^{-1}\left(\sum_{k=0}^{\infty} \frac{1}{k!}\Delta^k\right)P = P^{-1}e^\Delta P.$$

Pour la partie nilpotente, on note que

$$e^N = \sum_{k=0}^{\infty} \frac{1}{k!} N^k = \sum_{k=0}^{p-1} \frac{1}{k!} N^k,$$

puisque toutes les puissances supérieures à p s'annulent. ◊

Notons que si A est une matrice réelle, sa décomposition de Dunford est formée de matrices D et N réelles, mais les valeurs propres qui interviennent sont bien souvent complexes et non toutes réelles. C'est le cas si D, par définition diagonalisable sur $\mathbb{C}$, est non diagonalisable sur $\mathbb{R}$. Les exponentielles des valeurs propres sont alors également complexes. Comme on sait que e^D est réelle, il s'ensuit que la matrice de passage P, tout aussi complexe, se débrouille pour que le résultat final soit réel (donc par exemple avec des $\sin(\mu_i t)$, $\cos(\mu_i t)$, $\mu_i = \Im\lambda_i \in \mathbb{R}$).

Dans l'application de ce résultat aux EDO, on retient donc que si (D, N) est la décomposition de Dunford de A, alors (tD, tN) est celle de tA pour tout $t \in \mathbb{R}$. Par conséquent, on obtient

$$y(t) = e^{tA}y_0 = \left(\sum_{k=0}^{p-1} \frac{t^k}{k!} N^k\right)P^{-1}\mathrm{diag}\big(e^{\lambda_i t}\big)Py_0,$$

pour la solution du problème de Cauchy homogène, puisque les valeurs propres de tD sont les $\lambda_i t$ et que la matrice de passage ne dépend pas [14] de t. On voit apparaître une partie polynomiale par rapport à t et une

13. On fait figurer plusieurs fois les éventuelles valeurs propres multiples de D. Notons que ce sont aussi les valeurs propres de A.

14. C'est-à-dire que l'on peut la choisir indépendante de t : dans la réduction des matrices diagonalisables, la matrice de passage n'est jamais unique. En effet, elle correspond au choix d'une base formée de vecteurs propres. Or on peut toujours multiplier les vecteurs propres par des scalaires non nuls, ou permuter les espaces propres, etc. Or ici, les espaces propres ne dépendent pas de t (sauf pour $t = 0$, mais peu importe). Donc on peut choisir le même P une fois pour toutes, pour tout t.

autre partie dont les coefficients sont des combinaisons linéaires des $e^{\lambda_i t}$. La présence de termes polynomiaux (de degré supérieur à 1) ne se produit que quand la matrice A n'est pas diagonalisable.

La formule exponentielle est élégante, mais il ne faut pas trop se laisser aveugler : en pratique, si l'on peut raisonnablement considérer qu'il est faisable de calculer exactement la décomposition de Dunford avec les valeurs propres pour une matrice 2×2 à la main, puisqu'il s'agit essentiellement de trouver les racines d'un polynôme du second degré, puis les matrices de passage correspondantes, pour $m = 3$ et $m = 4$, la tâche devient nettement plus ardue, et impossible en général pour $m \geq 5$. Il ne faut donc pas trop compter sur l'exponentielle calculée exactement via les valeurs propres pour obtenir des informations quantitatives sur les solutions pour $m = 1800$ par exemple...

Au lieu de la décomposition de Dunford, on aurait pu utiliser la *forme de Jordan* [15] de la matrice A, qui est plus précise. Mais comme la considération de cette forme de Jordan n'apporte pas grand-chose sur le plan pratique, nous la laisserons ici de côté. Appliquons ce qui précède aux équations différentielles scalaires d'ordre quelconque linéaires à coefficients constants.

Théorème C.2.15 *Soit l'équation différentielle scalaire d'ordre n à coefficients constants homogène*

$$y^{(n)}(t) + a_1 y^{(n-1)}(t) + \cdots + a_{n-1} y'(t) + a_n y(t) = 0, \tag{C.2.9}$$

posée sur $\mathbb{R}$. On désigne par P le polynôme caractéristique de l'équation

$$P(X) = X^n + a_1 X^{n-1} + \cdots + a_{n-1} X + a_n.$$

Les $p \leq n$ racines complexes distinctes de P sont notées λ_i pour $i = 1, \ldots, p$, et leur multiplicité respective est notée α_i. L'espace des solutions de (C.2.9) *est l'espace des fonctions de la forme*

$$y(t) = \sum_{i=1}^{p} Q_i(t) e^{\lambda_i t}$$

où Q_i désigne un polynôme de degré inférieur à $\alpha_i - 1$ à coefficients dans $\mathbb{C}$.

Démonstration. On se ramène à un système différentiel d'ordre 1, comme dans la proposition C.1.2. La nouvelle fonction inconnue est $Y(t) = \big(y(t), y'(t), \ldots, y^{(n-1)}(t)\big)$, solution du système $Y'(t) = AY(t)$ avec

$$A = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & 1 \\ -a_n & -a_{n-1} & \ldots & -a_2 & -a_1 \end{pmatrix}.$$

La matrice A est appelée la *matrice compagnon* du polynôme P. Son polynôme caractéristique est exactement P, [16] comme on le vérifie en développant le déterminant par rapport à la première colonne. Les valeurs propres de A sont donc les λ_i avec multiplicité algébrique α_i. Soit S l'espace vectoriel des solutions de $Y' = AY$, de dimension n par le corollaire C.2.12. Notons E l'espace des fonctions de la forme $y(t) = \sum_{i=1}^{p} Q_i(t) e^{\lambda_i t}$ avec Q_i de degré inférieur à $\alpha_i - 1$. Comme $\sum_{i=1}^{p} \alpha_i = n$, cet espace est aussi de dimension n (simple exercice d'algèbre linéaire, à faire quand même !). [17] On a vu précédemment que l'application $Y \mapsto Y_1$ envoie S dans E. Elle est évidemment linéaire. Elle est de plus injective. En effet, si $Y_1 = y = 0$, alors $Y_i = y^{(i-1)} = 0$ pour tout i. Comme $\dim S = \dim E$, il s'ensuit qu'elle est surjective. ◊

On a bien sûr exactement la même description de la solution générale de (C.2.9) sur n'importe quel intervalle de $\mathbb{R}$. Pour retrouver dans le cas réel les solutions à valeurs réelles, on prend les parties réelles et imaginaires, comme indiqué plus haut.

Donnons sans démonstration le résultat suivant pour le cas non homogène.

15. Marie Ennemond Camille Jordan, 1838–1922. Attention, c'est un garçon.
16. Bon, peut-être au signe près.
17. On rappelle que la dimension de l'espace des polynômes de degré inférieur à $\alpha_i - 1$ est α_i.

Théorème C.2.16 *Soit Q un polynôme de degré q et $\lambda \in \mathbb{C}$ une constante. L'équation différentielle scalaire d'ordre n à coefficients constants*

$$y^{(n)}(t) + a_1 y^{(n-1)}(t) + \cdots + a_n y(t) = Q(t)e^{\lambda t}$$

admet une solution particulière de la forme $R(t)e^{\lambda t}$ où $R(t)$ est un polynôme de degré q si λ n'est pas une racine de $P(x)$ et un polynôme de degré $q + \alpha_i$ si $\lambda = \lambda_i$.

C.2.3 Systèmes différentiels linéaires à coefficients variables

Nous allons repasser ici au cas réel, dans la mesure où les questions de vecteurs propres et valeurs propres ne vont pas jouer de rôle particulier.

Il n'y a pas de difficulté supplémentaire à traiter le cas complexe, naturellement. Dans ce qui suit, A désigne donc une fonction continue de $[0,T]$ à valeurs dans $M_m(\mathbb{R})$ et b une fonction continue de $[0,T]$ à valeurs dans $\mathbb{R}^m$. Une tentation légitime serait d'essayer des formules du type $y(t) = e^{\int_0^t A(s)\,ds} y_0$, inspirées par le cas à coefficients constants, et par le cas général en dimension un. Malheureusement, c'est complètement faux la plupart du temps... En effet, si l'on a bien $\frac{d}{dt}\left(\int_0^t A(s)\,ds\right) = A(t)$, on n'a pas du tout que $\frac{d}{dt}\left(e^{\int_0^t A(s)\,ds}\right) = A(t)e^{\int_0^t A(s)\,ds}$ en général. [18]

Céder à cette tentation étant donc voué à l'échec, il faut procéder autrement. Commençons par obtenir une forme intégrale équivalente du problème, qui vaudra également pour le cas non linéaire général (2.1.3).

Proposition C.2.17 *Soit y une solution du problème de Cauchy* (C.2.2)*, continue sur $[0,T]$. Alors on a pour tout $t \in [0,T]$*

$$y(t) = y_0 + \int_0^t \big(A(s)y(s) + b(s)\big)\,ds. \tag{C.2.10}$$

Réciproquement, soit y une fonction continue sur $[0,T]$ à valeurs dans $\mathbb{R}^m$ et satisfaisant l'équation intégrale (C.2.10)*. Alors, y est dérivable sur $]0,T[$ et solution du problème de Cauchy* (C.2.2)*.*

Démonstration. Soit y une solution du problème de Cauchy (C.2.2), continue sur $[0,T]$. Comme $y'(t) = A(t)y(t) + b(t)$, on voit que y' est continue sur $]0,T[$ avec un prolongement continu en 0 et en T. Par conséquent, on en déduit que $y(t) - y(0) = \int_0^t y'(s)\,ds = \int_0^t \big(A(s)y(s) + b(s)\big)\,ds$ pour tout $t \in [0,T]$. Comme $y(0) = y_0$ par la donnée initiale du problème de Cauchy, on obtient bien (C.2.10).

Réciproquement, soit y une fonction continue vérifiant $y(t) = y_0 + \int_0^t \big(A(s)y(s) + b(s)\big)ds$ pour tout t. Alors y est automatiquement dérivable avec une dérivée continue sur $]0,T[$. En effet, l'intégrande est continue. On peut alors dériver cette égalité par rapport à t dans I, ce qui donne $y'(t) = A(t)y(t) + b(t)$, et donc y est bien solution de l'équation différentielle. Par ailleurs, faisant $t = 0$, on obtient bien $y(0) = y_0 + \int_0^0 \big(A(s)y(s) + b(s)\big)ds = y_0$. ◇

Les deux formulations, problème de Cauchy et équation intégrale, sont donc équivalentes. Nous allons maintenant établir l'existence et l'unicité de la solution du problème de Cauchy dans le cas d'un système linéaire à coefficients variables. Contrairement au cas des coefficients constants, nous ne pouvons pas nous reposer sur une formule explicite à base d'exponentielles, il va donc falloir construire cette solution à partir de rien...

Théorème C.2.18 *On suppose A et b continues sur $\bar{I} = [0,T]$. Le problème de Cauchy* (C.2.2) *admet une solution unique y.*

Démonstration. On va utiliser la forme intégrale de la proposition C.2.17. On définit sur $\bar{I}$ par récurrence une suite de fonctions vectorielles

$$\forall t \in \bar{I}, \quad y_0(t) = y_0, \tag{C.2.11}$$

$$y_{n+1}(t) = y_0 + \int_0^t \big(A(s)y_n(s) + b(s)\big)\,ds. \tag{C.2.12}$$

18. Essayer de s'en convaincre en tentant de le démontrer : on doit buter très vite sur des problèmes de non commutativité.

Le premier terme de la suite est donc la fonction constante égale à la condition initiale, puis on applique l'itération (C.2.12) pour définir chacun des termes suivants. Cette récurrence est manifestement bien définie et l'on a $y_n \in C^0(\bar{I};\mathbb{R}^m)$ pour tout n.

La démonstration consiste à prouver que la suite y_n converge uniformément sur $\bar{I}$ vers une fonction solution de (C.2.10), puis que cette solution est unique. Il est ensuite facile d'étendre le résultat à un intervalle I quelconque. Cette méthode s'étend au cas non linéaire (voir paragraphe C.3.1) et est connue dans la littérature, pour la partie existence, sous le nom de *méthode de Picard* [19].

On va majorer assez finement la différence de deux termes successifs de la suite de Picard. La continuité de la fonction $A(t)$ entraîne celle de la fonction scalaire $t \mapsto \|A(t)\|$. L'intervalle $\bar{I}$ étant compact, cette fonction est bornée sur $\bar{I}$, c'est-à-dire qu'il existe un réel α tel que pour tout $t \in \bar{I}$, $\|A(t)\| \leq \alpha$. D'après la propriété de la norme matricielle (C.2.5), on a $\|A(s)z\| \leq \|A(s)\|\,\|z\| \leq \alpha\|z\|$ pour tout $z \in \mathbb{R}^m$ et tout $s \in \bar{I}$. De même, la fonction b étant continue sur le compact $\bar{I}$, elle est, elle aussi, bornée en norme par une constante β.

Comme $y_1(t) - y_0(t) = \int_0^t \big(A(s)y_0 + b(s)\big)\,ds$, il vient par l'inégalité triangulaire

$$\forall t \in [0,T], \quad \|y_1(t) - y_0(t)\| \leq \int_0^t \big(\|A(s)y_0\| + \|b(s)\|\big)\,ds \leq \int_0^t \big(\alpha\|y_0\| + \beta\big)\,ds = \big(\alpha\|y_0\| + \beta\big)t.$$

Pour tout $n \geq 1$, on a de plus la relation

$$y_{n+1}(t) - y_n(t) = \int_0^t A(s)\big(y_n(s) - y_{n-1}(s)\big)\,ds.$$

en soustrayant l'égalité (C.2.12) pour $n-1$ de cette même égalité pour n. Par conséquent, comme pour tout $s \in \bar{I}$,

$$\big\|A(s)\big(y_n(s) - y_{n-1}(s)\big)\big\| \leq \|A(s)\|\|y_n(s) - y_{n-1}(s)\| \leq \alpha\|y_n(s) - y_{n-1}(s)\|,$$

on obtient en intégrant et utilisant l'inégalité triangulaire

$$\forall t \in [0,T], \quad \|y_{n+1}(t) - y_n(t)\| \leq \alpha \int_0^t \|y_n(s) - y_{n-1}(s)\|\,ds,$$

pour tout $n \geq 1$.

Des inégalités précédentes, on va déduire la majoration

$$\forall t \in [0,T], \quad \|y_{n+1}(t) - y_n(t)\| \leq \frac{\alpha\|y_0\| + \beta}{\alpha}\frac{(\alpha t)^{n+1}}{(n+1)!}. \tag{C.2.13}$$

On procède par récurrence. Tout d'abord, l'estimation (C.2.13) est vraie pour $n = 0$, on vient de le voir un peu plus haut. Supposons la vraie en n, on obtient alors pour tout $t \in [0,T]$,

$$\begin{aligned}
\|y_{n+2}(t) - y_{n+1}(t)\| &\leq \alpha \int_0^t \|y_{n+1}(s) - y_n(s)\|\,ds \\
&\leq \alpha \int_0^t \frac{\alpha\|y_0\| + \beta}{\alpha}\frac{(\alpha s)^{n+1}}{(n+1)!}\,ds \\
&= \alpha^{n+1}\frac{\alpha\|y_0\| + \beta}{(n+1)!}\int_0^t s^{n+1}\,ds \\
&= \alpha^{n+1}\frac{\alpha\|y_0\| + \beta}{(n+1)!}\frac{t^{n+2}}{n+2} = \frac{\alpha\|y_0\| + \beta}{\alpha}\frac{(\alpha t)^{n+2}}{(n+2)!}.
\end{aligned}$$

L'estimation (C.2.13) étant maintenant établie, il s'ensuit que pour tout $n \geq 0$,

$$\|y_{n+1} - y_n\|_{C^0([0,T];\mathbb{R}^m)} = \max_{t\in[0,T]} \|y_{n+1}(t) - y_n(t)\| \leq \max_{t\in[0,T]} \left(\frac{\alpha\|y_0\| + \beta}{\alpha}\frac{(\alpha t)^{n+1}}{(n+1)!}\right) = \frac{\alpha\|y_0\| + \beta}{\alpha}\frac{(\alpha T)^{n+1}}{(n+1)!}.$$

19. Charles Émile Picard, 1856–1941.

La série numérique à termes positifs $\sum_{n=1}^{\infty} \|y_n - y_{n-1}\|_{C^0([0,T];\mathbb{R}^m)}$ est dominée par la série de terme général $\frac{(\alpha T)^n}{n!}$, notoirement convergente. Elle donc est convergente, avec de plus

$$\sum_{n=1}^{\infty} \|y_n - y_{n-1}\|_{C^0([0,T];\mathbb{R}^m)} \leq \frac{\alpha\|y_0\| + \beta}{\alpha} \sum_{n=1}^{\infty} \frac{(\alpha T)^n}{n!} = \frac{\alpha\|y_0\| + \beta}{\alpha}\left(e^{\alpha T} - 1\right).$$

La série de fonctions $\sum_{n=1}^{\infty}(y_n - y_{n-1})$ est donc normalement convergente dans l'espace complet $C^0([0,T];\mathbb{R}^m)$, par conséquent elle converge dans ce même espace. Écrivant la somme télescopique [20]

$$y_n = y_0 + \sum_{k=1}^{n}(y_k - y_{k-1}),$$

on voit que la suite y_n converge uniformément sur $[0,T]$ vers la fonction continue

$$y = y_0 + \sum_{k=1}^{\infty}(y_k - y_{k-1}).$$

Comme

$$\|A(t)y_n(t) - A(t)y(t)\| = \|A(t)(y_n(t) - y(t))\| \leq \alpha\|y_n(t) - y(t)\|,$$

on en déduit en passant au max sur t d'abord à droite, puis à gauche, que Ay_n converge uniformément vers Ay sur $[0,T]$. On peut donc sans difficulté passer à la limite dans l'intégrale du membre de droite de (C.2.12) pour obtenir (C.2.10).

Prenons maintenant deux solutions y et $\tilde{y}$. Leur différence $z = y - \tilde{y}$ vérifie

$$z(t) = \int_0^t A(s)z(s)\,ds. \tag{C.2.14}$$

On montre exactement comme précédemment qu'alors on a, pour tout $n \geq 0$,

$$\|z(t)\| \leq M\frac{(\alpha t)^n}{n!}.$$

où $M = \max_{[0,T]} \|z(t)\|$. Or $\frac{(\alpha t)^n}{n!} \to 0$ quand $n \to +\infty$ pour tout t. Ceci implique que $z(t) = 0$ pour tout t, l'unicité de la solution de (C.2.17) est donc établie. ◊

Remarquons que l'on peut voir cette démonstration d'un peu plus haut comme une application du théorème de point fixe de Picard (ou de Banach [21] suivant le pays dans lequel on se trouve) sur les applications strictement contractantes dans un espace métrique complet. [22] Mais il est un peu dommage de se priver de l'approche itérative de Picard pour un peu plus d'abstraction sans vraiment gagner tant que ça en longueur de preuve...

Ce théorème, ainsi que le théorème de Cauchy-Lipschitz qui viendra bientôt le compléter, est une illustration particulièrement éclatante de la puissance de la notion de complétude. En effet, nous sommes ici devant un problème dont nous n'avions au départ pas la moindre idée s'il admettait ou non une solution. Nous construisons alors une suite de fonctions (dans un espace complet) qui, si elle converge, a de bonnes chances de nous donner l'existence d'une telle solution. Pour montrer qu'elle converge, c'est-à-dire pour montrer qu'il existe une limite à cette suite, et par conséquent résoudre le problème de départ, on montre simplement qu'elle est de Cauchy et la complétude de l'espace ambiant fait le reste !

La Figure C.1 montre un exemple pour l'edo $y'(t) = \begin{pmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{pmatrix} y(t)$ ($m = 2$) avec la donnée initiale $y(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

20. Cette astuce de passer par une somme télescopique qui se trouve être une série convergente est assez courante.

21. Stefan Banach, 1892–1945.

22. À condition de bien choisir l'espace métrique en question. Celui que nous avons pris ici est un peu trop naïf. Prendre $m = 1$, $A(s) = 1$ et $T > 1$ pour le voir. Nous retrouverons ce théorème de point fixe de Picard dans la seconde partie du cours sur les approximations numériques.

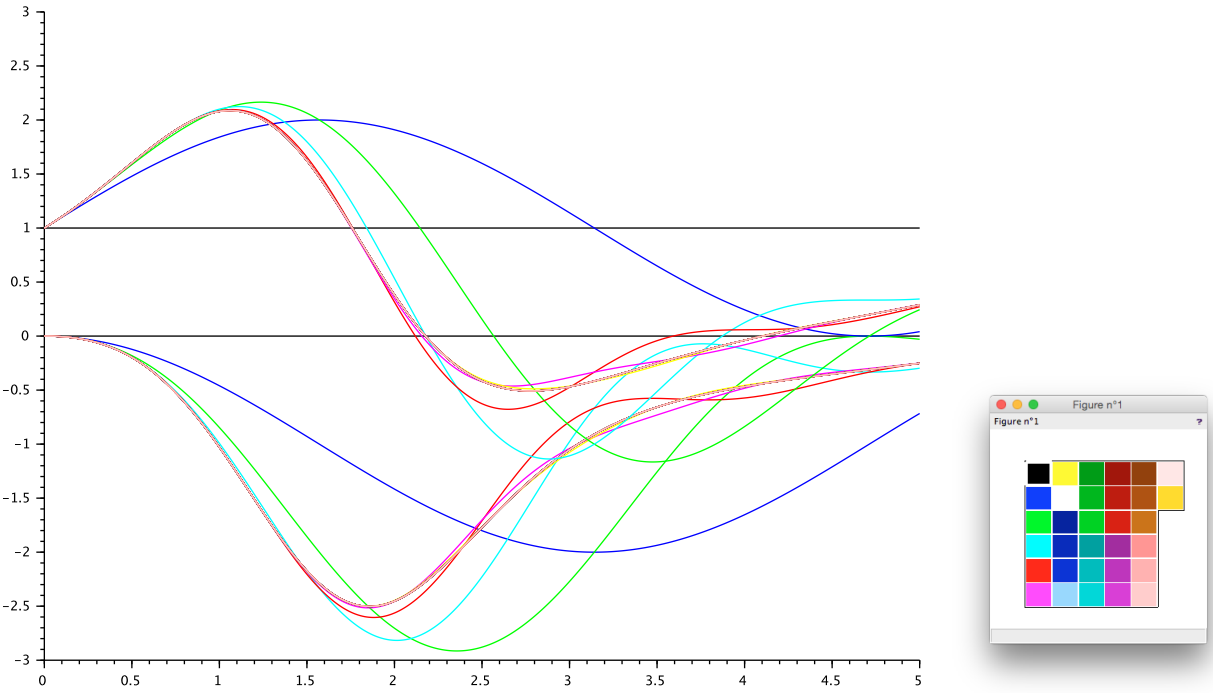


FIGURE C.1 – Un exemple d'itérations de Picard. Les courbes de même couleur correspondent aux deux composantes de la même itération, en commençant par noir, bleu foncé, etc. À droite, l'ordre des couleurs de la palette par défaut de scilab.

Corollaire C.2.19 *L'ensemble des solutions de l'équation homogène $y'(t) = A(t)y(t)$ est un espace vectoriel de dimension m.*

Démonstration. Soit S cet ensemble. On a déjà vu que c'est un espace vectoriel, *cf.* Proposition C.2.5. [23] L'application qui à $y \in S$ fait correspondre $y(0) \in \mathbb{R}^m$ est trivialement linéaire. Elle est surjective puisque tout $y_0 \in \mathbb{R}^m$ donne naissance à un $y \in S$ par l'existence du théorème C.2.18, et injective, car si $y(0) = 0$, alors $y(t) = 0$ pour tout t, par l'unicité du même théorème. C'est par conséquent un isomorphisme de S sur $\mathbb{R}^m$. Comme $\dim \mathbb{R}^m = m$, on en déduit que $\dim S = m$. ♢

Exemple C.2.1 Exponentielle de matrices, le retour.
Reprenons le cas d'une équation linéaire autonome sur $\mathbb{R}^m$, $y'(t) = Ay(t)$ avec $y(0) = y_0$, où $A \in M_m(\mathbb{R})$ ne dépend pas de t. On est bien dans le cadre d'application du théorème C.2.18 et on peut expliciter la méthode itérative utilisée dans la démonstration.

La première itération donne

$$y_1(t) = y_0 + \int_0^t Ay_0\, ds = Iy_0 + tAy_0 = [(tA)^0 + (tA)^1]y_0.$$

La seconde itération donne

$$y_2(t) = y_0 + \int_0^t Ay_1(s)\, ds = y_0 + \int_0^t A(y_0 + sAy_0)\, ds$$
$$= Iy_0 + tAy_0 + \frac{t^2}{2}A^2y_0 = \left[(tA)^0 + (tA)^1 + \frac{(tA)^2}{2}\right]y_0.$$

23. Considérée comme évidente, donc donnée sans preuve...

Une récurrence immédiate (mais ce n'est pas une raison de ne pas la faire en détail) conduit alors à

$$y_n(t) = \Big[(tA)^0 + (tA)^1 + \frac{(tA)^2}{2} + \cdots + \frac{(tA)^n}{n!}\Big]y_0 = \Big[\sum_{k=0}^{n} \frac{(tA)^k}{k!}\Big]y_0.$$

Comme on l'a vu dans le cas général, la suite $\big(y_n\big)_{n\in\mathbb{N}}$ converge vers la solution y du problème de Cauchy. On retrouve donc par ce biais itératif que l'on a bien $y(t) = e^{tA}y_0$ dans ce cas particulier. ◊

On a également un résultat sur la dépendance continue de la solution par rapport aux conditions initiales.

Proposition C.2.20 *Soient y_0 et $\tilde{y}_0$ deux conditions initiales pour le problème de Cauchy* (C.2.2)*, et y et $\tilde{y}$ les solutions correspondantes sur $[0,T]$. Alors on a*

$$\|y - \tilde{y}\|_{C^0([0,T];\mathbb{R}^m)} \leq e^{\alpha T}\|y_0 - \tilde{y}_0\|,$$

où $\alpha = \max_{[0,T]} \|A(t)\|$, et l'application $y_0 \mapsto y$ est donc continue de $\mathbb{R}^m$ dans $C^0([0,T];\mathbb{R}^m)$.

Démonstration. Soit y (resp. $\tilde{y}$) la solution de (C.2.2) correspondant à la donnée initiale $y_0 \in \mathbb{R}^m$ (resp. $\tilde{y}_0$). On a donc $y'(t) - \tilde{y}'(t) = A(t)(y(t) - \tilde{y}(t))$ pour tout t. Prenons le produit scalaire [24] de l'égalité précédente par $y(t) - \tilde{y}(t)$. Pour le membre de gauche, il vient

$$\big(y'(t) - \tilde{y}'(t)\big|y(t) - \tilde{y}(t)\big) = \frac{1}{2}\frac{d}{dt}\|y(t) - \tilde{y}(t)\|^2.$$

En effet, pour toute fonction $t \mapsto z(t)$ dérivable de I dans $\mathbb{R}^m$, on peut écrire à t fixé $z(t+h) = z(t)+hz'(t)+h\varepsilon(h)$ où $\varepsilon(h)$ est une fonction à valeurs vectorielles telle que $\|\varepsilon(h)\| \to 0$ quand $h \to 0$. Il vient donc

$$\begin{aligned}
\|z(t+h)\|^2 &= \big(z(t+h)\big|z(t+h)\big)\\
&= \big(z(t) + hz'(t) + h\varepsilon(h)\big|z(t) + hz'(t) + h\varepsilon(h)\big)\\
&= \big(z(t)\big|z(t)\big) + \big(hz'(t)\big|z(t)\big) + \big(z(t)\big)\big|hz'(t)\big) + h\tilde{\varepsilon}(h)\\
&= \|z(t)\|^2 + 2h\big(z'(t)\big|z(t)\big) + h\tilde{\varepsilon}(h),
\end{aligned}$$

où l'on a posé $\tilde{\varepsilon}(h) = \big(\varepsilon(h)\big|2z(t) + hz'(t) + h\varepsilon(h)\big) + h\|z'(t)\|^2$, si bien que $|\tilde{\varepsilon}(h)| \to 0$ quand $h \to 0$, d'où le résultat en repassant $\|z(t)\|^2$ au membre de gauche, en divisant par h puis en faisant tendre h vers 0. Pour le membre de droite, on trouve

$$\big(A(t)(y(t) - \tilde{y}(t))\big|y(t) - \tilde{y}(t)\big) \leq \|A(t)(y(t) - \tilde{y}(t))\|\|y(t) - \tilde{y}(t)\| \leq \alpha\|y(t) - \tilde{y}(t)\|^2,$$

par l'inégalité de Cauchy-Schwarz, la définition d'une norme matricielle et celle de la constante α. Posons $v(t) = \|y(t) - \tilde{y}(t)\|^2$. On a donc obtenu l'inéquation différentielle

$$v'(t) \leq 2\alpha v(t),$$

qui se résout très facilement à l'aide d'un facteur intégrant

$$0 \geq e^{-2\alpha t}(v'(t) - 2\alpha v(t)) = \big(e^{-2\alpha t}v(t)\big)'.$$

La fonction $t \mapsto e^{-2\alpha t}v(t)$ est donc décroissante, ce qui implique que $e^{-2\alpha t}v(t) \leq v(0)$ pour tout $t \in [0,T]$, soit

$$\|y(t) - \tilde{y}(t)\| \leq e^{\alpha t}\|y_0 - \tilde{y}_0\| \leq e^{\alpha T}\|y_0 - \tilde{y}_0\|.$$

On en déduit que

$$\|y - \tilde{y}\|_{C^0([0,T];\mathbb{R}^m)} = \max_{t\in[0,T]} \|y(t) - \tilde{y}(t)\| \leq e^{\alpha T}\|y_0 - \tilde{y}_0\|,$$

d'où la continuité annoncée. ◊

24. Dans le cas complexe, il faut prendre la partie réelle du produit scalaire hermitien sur $\mathbb{C}^m$, pour lequel $(u|v) = \overline{(v|u)}$.

Le résultat signifie entre autres que quand $\tilde{y}_0$ tend vers y_0 dans $\mathbb{R}^m$, alors les solutions correspondantes convergent uniformément sur $[0, T]$. On dit qu'il y a *dépendance continue par rapport aux données initiales*. Pour être continue, cette dépendance n'en peut pas moins être très sensible. En effet, il se peut que le facteur exponentiel $e^{\alpha T}$ soit effectif, c'est-à-dire pas seulement une majoration mais essentiellement atteint, et il peut-être numériquement énorme dès que αT est modérément grand, comme dans toute exponentielle qui se respecte.

Le théorème C.2.18 assure l'existence et l'unicité de la solution du problème de Cauchy. On a vu que des formules du genre $y(t) = e^{\int_0^t A(s)\,ds} y_0$ ne marchent pas. Est-il néanmoins possible d'écrire cette solution de façon plus ou moins explicite, comme dans le cas des coefficients constants ?

Prenons tout d'abord le cas homogène, $b = 0$. On va considérer ici le problème de Cauchy posé sur $\mathbb{R}$ entier, avec donnée initiale à l'instant t_0 au lieu de 0. Bien sûr, le résultat d'existence et d'unicité est inchangé.

Théorème C.2.21 *Il existe une application R de $\mathbb{R}^2$ dans $\mathrm{GL}_m(\mathbb{R})$, appelée la* résolvante du système différentiel, *telle que la solution du problème de Cauchy homogène*

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

est donnée par

$$y(t) = R(t, t_0)y_0.$$

Démonstration. Soit S l'espace vectoriel des solutions de l'EDO. L'application $y_0 \mapsto y$ est un isomorphisme de $\mathbb{R}^m$ dans S, car c'est l'application réciproque de l'isomorphisme du corollaire C.2.19. Fixons t. L'application qui à y fait correspondre $y(t)$ est trivialement linéaire de S dans $\mathbb{R}^m$ et aussi un isomorphisme. L'application $y_0 \mapsto y(t)$ est la composée de ces deux isomorphismes, c'est donc un automorphisme de $\mathbb{R}^m$. La résolvante est simplement la matrice — inversible — de cet automorphisme dans la base canonique. ◊

Dans le cas d'une équation à coefficients constants, on retrouve $R(t, t_0) = e^{(t-t_0)A}$. En utilisant encore l'unicité de la solution du problème de Cauchy, on obtient la propriété suivante pour tous t_0, t_1, t_2, $R(t_2, t_0) = R(t_2, t_1)R(t_1, t_0)$. Faisant $t_2 = t_0$ et en raison du fait évident que $R(t_0, t_0) = I$, on en déduit que $(R(t_1, t_0))^{-1} = R(t_0, t_1)$.

Définition C.2.22 *Soient $(y^i(t))_{i=1,\dots,m}$, les m solutions du problème de Cauchy homogène, obtenues en prenant comme conditions initiales à t_0 les m vecteurs d'une base quelconque de $\mathbb{R}^m$. On dit qu'elles forment un* système fondamental *et on définit leur* matrice wronskienne *par*

$$W(t) = \begin{pmatrix} y_1^1(t) & \cdots & y_1^m(t) \\ \vdots & & \vdots \\ y_m^1(t) & \cdots & y_m^m(t) \end{pmatrix}.$$

Proposition C.2.23 *La matrice wronskienne satisfait $W'(t) = A(t)W(t)$ et est liée à la résolvante par $R(t, t_0) = W(t)W(t_0)^{-1}$.*

Démonstration. La première relation est juste l'expression du produit matriciel AW à l'aide des produits matrice-vecteur de A avec les colonnes de W. Soit y_0 une donnée initiale en t_0. On la décompose sur la base $(y^i(t_0))_{i=1,\dots,m}$ sous la forme $y_0 = \sum_{i=1}^m \lambda_i y^i(t_0)$, c'est-à-dire $y_0 = W(t_0)\lambda$, où λ désigne le vecteur colonne des λ_i. Par unicité du problème de Cauchy, on vérifie comme précédemment que la solution y correspondant à y_0 est donnée par $y(t) = \sum_{i=1}^m \lambda_i y^i(t)$, c'est-à-dire $y(t) = W(t)\lambda$. Comme par ailleurs, $y(t) = R(t, t_0)y_0 = R(t, t_0)W(t_0)\lambda$, on obtient $W(t)\lambda = R(t, t_0)W(t_0)\lambda$ pour tout $\lambda \in \mathbb{R}^m$. Il s'ensuit que $W(t) = R(t, t_0)W(t_0)$. ◊

Les espoirs de formules explicites sont un peu déçus. En effet, en général, on ne peut pas calculer explicitement la matrice wronskienne d'une base, ni donc la résolvante. Il se trouve que son déterminant, que l'on appelle le *wronskien* [25], est solution de l'EDO scalaire $\frac{d}{dt}(\det W)(t) = \operatorname{tr} A(t) \det W(t)$. Cette équation à

25. Josef Hoëné-Wronski, 1776–1853.

variables séparées se résout avec le facteur intégrant qui s'impose et on connaît sa donnée initiale, qui est le déterminant de la base de départ. Il s'agit du théorème de Liouville.

Venons-en au cas non homogène. On reprend la méthode de la variation de la constante en posant $y(t) = W(t)\lambda(t)$, ce qui est loisible puisque $W(t)$ est toujours inversible. Il vient d'une part

$$\begin{aligned} y'(t) &= W'(t)\lambda(t) + W(t)\lambda'(t) \\ &= A(t)W(t)\lambda(t) + W(t)\lambda'(t) \\ &= A(t)y(t) + W(t)\lambda'(t), \end{aligned}$$

par la formule de Leibniz et l'expression de la dérivée de la matrice wronskienne. D'un autre côté, l'EDO nous dit que $y'(t) = A(t)y(t) + b(t)$. Comparant les deux expressions, on en déduit que $\lambda'(t) = W(t)^{-1}b(t)$ et de là, en intégrant entre t_0 et t

$$\lambda(t) = \lambda(t_0) + \int_{t_0}^{t} W(s)^{-1}b(s)\,ds = W(t_0)^{-1}y_0 + \int_{t_0}^{t} W(s)^{-1}b(s)\,ds.$$

Reportant l'expression de λ ainsi obtenue dans y, on a montré le

Théorème C.2.24 *La solution unique du problème de Cauchy*

$$\begin{cases} y'(t) = A(t)y(t) + b(t), \\ y(t_0) = y_0, \end{cases}$$

est donnée par

$$\begin{aligned} y(t) &= W(t)\Big(W(t_0)^{-1}y_0 + \int_{t_0}^{t} W(s)^{-1}b(s)\,ds\Big) \\ &= R(t,t_0)y_0 + \int_{t_0}^{t} R(t,s)b(s)\,ds. \end{aligned}$$

C'est la formule de Duhamel dans le cas des coefficients variables. Sauf que, en général, on ne peut pas calculer $R(t,s)$. Dans le cas très particulier où $A(t)$ et $A(s)$ commutent pour tous t et s, la forme générale se simplifie et on peut expliciter la solution en fonction de $A(t)$ et $b(t)$.

Proposition C.2.25 *Supposons que pour tous $t, s \in I$, $A(t)A(s) = A(s)A(t)$. Alors La solution unique du problème de Cauchy*

$$\begin{cases} y'(t) = A(t)y(t) + b(t), \\ y(t_0) = y_0, \end{cases}$$

est donnée par

$$y(t) = R(t,t_0)y_0 + \int_{t_0}^{t} R(t,s)b(s)\,ds.$$

avec

$$R(t,t_0) = e^{\int_{t_0}^{t} A(s)\,ds}.$$

Démonstration. Montrons d'abord que A commute avec ses primitives. En effet

$$\Big(\int_{t_0}^{t} A(s)\,ds\Big)A(t) = \int_{t_0}^{t} A(s)A(t)\,ds = \int_{t_0}^{t} A(t)A(s)\,ds = A(t)\Big(\int_{t_0}^{t} A(s)\,ds\Big)$$

(la multiplication à gauche ou à droite par $A(t)$ est linéaire et passe donc dans l'intégrale, écrire les coefficients pour le voir). Considérons maintenant une fonction $t \mapsto F(t)$ dérivable à valeurs dans $M_m(\mathbb{R})$ qui commute avec sa dérivée $F(t)F'(t) = F'(t)F(t)$. On en déduit par une récurrence immédiate (mais à faire quand même)

que pour tout $k \in \mathbb{N}$, $(F^k)'(t) = kF(t)^{k-1}F'(t) = kF'(t)F(t)^{k-1}$ sur I. [26] Reprenant la série entière qui définit l'exponentielle, on en déduit que

$$\frac{d}{dt}\left(e^{F(t)}\right) = e^{F(t)}F'(t) = F'(t)e^{F(t)},$$

relation qu'il suffit d'appliquer à $F(t) = \int_{t_0}^{t} A(s)\,ds$ pour conclure. ◊

On retrouve que dans le cas où A est constant, donc commute à tout temps, $R(t, t_0) = e^{(t-t_0)A}$.

C.3 Existence et unicité dans le cas général

Nous abordons maintenant le cas général où le second membre $f(t, y)$ du système différentiel n'est plus linéaire par rapport à y. Après un détour dans le passé glorieux, nous traitons au paragraphe C.3.1 le cas où f est *globalement lipschitzienne*, et où il y a existence et unicité d'une solution sur tout l'intervalle de temps I.

C.3.1 Résultats d'existence et d'unicité dans le cas général

Nous allons maintenant énoncer et démontrer le premier grand théorème concernant les EDO dans le cas général, le théorème de Cauchy-Lipschitz global. C'est *le résultat fondamental* d'existence et d'unicité pour une vaste classe de problèmes de Cauchy.

On introduit d'abord une propriété qui est en quelque sorte intermédiaire entre la continuité et la différentiabilité.

Définition C.3.1 *On dit qu'une fonction f de $[0, T] \times \mathbb{R}^m$ dans $\mathbb{R}^m$ est* globalement lipschitzienne relativement à la variable y, uniformément par rapport à t, *s'il existe une constante L telle que, pour tous y et $z \in \mathbb{R}^m$ et pour tout $t \in [0, T]$ on ait*

$$\|f(t, y) - f(t, z)\| \leq L\|y - z\|,$$

où $\|\cdot\|$ désigne (par exemple) la norme euclidienne sur $\mathbb{R}^m$. La plus petite constante L pour laquelle la majoration a lieu s'appelle la constante de Lipschitz *de la fonction f.*

Remarquons que pour tout t, l'application $y \mapsto f(t, y)$ est alors continue de $\mathbb{R}^m$ dans $\mathbb{R}^m$. Elle n'est par contre pas nécessairement différentiable, comme l'exemple $y \mapsto |y|$ de $\mathbb{R}$ dans $\mathbb{R}$ le montre.

On sait que toutes les normes sur $\mathbb{R}^m$ sont équivalentes, donc le choix de la norme euclidienne n'est pas fondamental, n'importe quelle autre norme ferait aussi bien l'affaire.

Cette condition de Lipschitz semble devoir jouer un rôle crucial par la suite, alors autant donner tout de suite des conditions suffisantes faciles à vérifier qui l'entraînent.

Proposition C.3.2 *Supposons que f possède des dérivées partielles par rapport à y_i, $i = 1, \ldots, m$, continues par rapport à y et bornées sur $\mathbb{R} \times \mathbb{R}^m$. Alors f est globalement lipschitzienne relativement à y, uniformément par rapport à t.*

Démonstration. On suppose donc que $\frac{\partial f}{\partial y_i}(t, y)$ existe pour tout i et définit une fonction continue par rapport à y et bornée au sens où il existe C tel que $\left\|\frac{\partial f}{\partial y_i}(t, y)\right\| \leq C$ pour tout $(t, y) \in \mathbb{R} \times \mathbb{R}^m$ et tout i.

Prenons deux points y et z de $\mathbb{R}^m$. Pour tout $t \in [0, T]$, l'application $g\colon [0, 1] \to \mathbb{R}^m$, $s \mapsto f(t, sy + (1-s)z)$ est de classe C^1 par dérivation des fonctions composées, avec

$$\begin{aligned} g'(s) &= \sum_{i=1}^{m} \frac{\partial f}{\partial y_i}(t, sy + (1-s)z)\frac{d}{ds}(sy + (1-s)z)_i \\ &= \sum_{i=1}^{m} \frac{\partial f}{\partial y_i}(t, sy + (1-s)z)(y_i - z_i). \end{aligned}$$

26. Ces formules sont bien sûr violemment fausses si $F(t)$ et $F'(t)$ ne commutent pas, ce qui est le cas général.

Comme g' est continue sur $[0,1]$, il s'ensuit que

$$\begin{aligned} f(t,y) - f(t,z) &= g(1) - g(0) \\ &= \int_0^1 g'(s)\,ds \\ &= \sum_{i=1}^m (y_i - z_i) \int_0^1 \frac{\partial f}{\partial y_i}(t, sy + (1-s)z)\,ds. \end{aligned}$$

Prenant la norme des deux membres, on en déduit par l'inégalité triangulaire

$$\|f(t,y) - f(t,z)\| \leq C \sum_{i=1}^m |y_i - z_i| \leq C\sqrt{m}\|y - z\|,$$

avec l'inégalité de Cauchy-Schwarz pour conclure. ◊

Remarque C.3.1 Il est en général assez facile de voir si une fonction donnée a des dérivées partielles et si ces dérivées partielles sont continues et bornées, d'où l'intérêt de ce qui précède. Dans le cas d'une équation autonome où f ne dépend pas de t, il suffit donc que f soit de classe C^1 et ait des dérivées partielles bornées sur $\mathbb{R}^m$. Dans le cas scalaire, $m = 1$, $y'(t) = f(y(t))$, il suffit donc qu'il existe C tel que $|f'(y)| \leq C$ pour tout $y \in \mathbb{R}$.

Attention, la condition de la proposition C.3.2 n'est qu'une condition suffisante. Il existe évidemment de nombreuses fonctions lipschitziennes qui ne sont pas C^1, ni même ne sont partout dérivables. Il est donc hautement préférable d'éviter de déclarer C^1 une fonction qui ne l'est manifestement pas. ◊

Notons que toute solution de toute EDO raisonnable a une régularité minimale automatique.

Proposition C.3.3 *Supposons que f soit continue par rapport à (t,y). Alors toute solution y de l'EDO $y'(t) = f(t,y(t))$ est de classe C^1.*

Démonstration. Soit y une telle solution. Par définition, elle est dérivable sur son intervalle de définition, donc continue. Par composition des fonctions continues, on en déduit que $t \mapsto f(t,y(t))$ est continue, c'est-à-dire que y' est continue, c'est-à-dire que y est C^1. ◊

Remarquons que si y est continue jusqu'aux bornes de son intervalle de définition, il en va de même pour y', au sens où y' admet un prolongement continu sur l'intervalle fermé. On reviendra plus loin sur ces questions de régularité de la solution d'une EDO. Nous énonçons maintenant le premier résultat fondamental d'existence et d'unicité, le théorème de Cauchy-Lipschitz global.

Théorème C.3.4 (Cauchy-Lipschitz global) *Soit $T > 0$ un réel fixé. Soit f une fonction continue sur $[0,T]\times\mathbb{R}^m$ à valeurs dans $\mathbb{R}^m$ et globalement lipschitzienne par rapport à la variable y, uniformément par rapport à t. Alors pour toute donnée initiale $y_0 \in \mathbb{R}^m$, il existe une unique solution y au problème de Cauchy* (2.1.3).

Avant d'entamer la preuve, un petit commentaire sur les hypothèses demandées sur f. En effet, celles-ci sont un tout petit peu redondantes, car on a la proposition suivante.

Proposition C.3.5 *Toute fonction $f : [0,T]\times\mathbb{R}^m \to \mathbb{R}^m$ qui est continue par rapport à t à y fixé et globalement lipschitzienne par rapport à y, uniformément par rapport à t, est en fait continue par rapport au couple (t,y).*

Démonstration. Soit $(t,y) \in [0,T]\times\mathbb{R}^m$ et prenons une suite $(t_n, y_n) \to (t,y)$ quand $n \to +\infty$ quelconque. On peut écrire

$$f(t_n,y_n) - f(t,y) = f(t_n,y_n) - f(t_n,y) + f(t_n,y) - f(t,y),$$

si bien que par l'inégalité triangulaire

$$\begin{aligned} \|f(t_n,y_n) - f(t,y)\| &\leq \|f(t_n,y_n) - f(t_n,y)\| + \|f(t_n,y) - f(t,y)\| \\ &\leq L\|y_n - y\| + \|f(t_n,y) - f(t,y)\| \to 0 \text{ quand } n \to +\infty, \end{aligned}$$

par la continuité par rapport à t à y fixé pour le second terme. ◊

Naturellement, une fonction continue par rapport au couple (t, y) est trivialement continue par rapport à t à y fixé sans autre hypothèse. Ceci dit, dans les applications pratiques, la continuité par rapport à (t, y) se voit aussi facilement que la continuité par rapport à t à y fixé, donc le fait que l'énoncé du théorème C.3.4 comporte des hypothèses légèrement redondantes n'est pas dramatique en soi. On rappelle quand même qu'une fonction de deux variables peut très bien être continue séparément par rapport à chaque variable, mais pas continue par rapport au couple de variables, comme par exemple $g\colon \mathbb{R}^2 \to \mathbb{R}$ définie par $g(0,0) = 0$ et $g(t,y) = \frac{ty}{t^2+y^2}$ pour $(t,y) \neq (0,0)$.

La preuve du théorème de Cauchy-Lipschitz global est pratiquement identique à celle déjà faite dans le cas linéaire au paragraphe C.2, *cf.* théorème C.2.18, page 170. Le sentiment de déjà-vu est donc normal, car les idées principales ont déjà été exposées. En vue de l'unicité, montrons d'abord un résultat d'intérêt général, le lemme de Grönwall [27] (ou une de ses versions les plus simples).

Proposition C.3.6 *Soient deux réels $\alpha \neq 0$ et β. Soit une fonction continue v de $[0,T]$ dans $\mathbb{R}$, dérivable sur $]0,T[$ et vérifiant*

$$v'(t) \leq \alpha v(t) + \beta$$

sur ce dernier intervalle. Alors pour tout $t \in [0,T]$,

$$v(t) + \frac{\beta}{\alpha} \leq \left(v(0) + \frac{\beta}{\alpha}\right) e^{\alpha t}.$$

Démonstration. On effectue le changement de fonction

$$z(t) = \left(v(t) + \frac{\beta}{\alpha}\right) e^{-\alpha t}.$$

La fonction z est dérivable sur $]0,T[$ et l'on a

$$z'(t) = -(\alpha v(t) + \beta) e^{-\alpha t} + v'(t) e^{-\alpha t} \leq 0.$$

Le théorème des accroissements finis implique que z est décroissante, en particulier que $z(t) \leq z(0)$, ce qui est exactement la conclusion du lemme de Grönwall. ◊

On a déjà rencontré le lemme de Grönwall en passant dans le cas $\beta = 0$ et c'est le plus souvent dans ce cas que nous l'utiliserons ici. Mais enfin, il ne coûte pas beaucoup plus cher quand $\beta \neq 0$, alors pourquoi s'en priver ?

Entamons maintenant la démonstration du théorème de Cauchy-Lipschitz global C.3.4. On notera de façon générique L la constante de Lipschitz de f par rapport à y.

Proposition C.3.7 *Sous les hypothèses du Théorème* C.3.4*, si y et $\tilde{y}$ sont deux solutions de l'équation différentielle $y'(t) = f(t, y(t))$, on a, pour tout $t \in [0,T]$,*

$$\|y(t) - \tilde{y}(t)\| \leq e^{Lt} \|y(0) - \tilde{y}(0)\|.$$

Démonstration. La preuve est strictement identique à celle de la proposition C.2.20, mais nous la recopions quand même ici.

Soient y et $\tilde{y}$ deux solutions de (2.1.3). On a donc $y'(t) - \tilde{y}'(t) = f(t, y(t)) - f(t, \tilde{y}(t))$ pour tout t. Prenons le produit scalaire de l'égalité précédente par $y(t) - \tilde{y}(t)$. On obtient

$$\frac{1}{2}\frac{d}{dt}\|y(t) - \tilde{y}(t)\|^2 = \big(f(t, y(t)) - f(t, \tilde{y}(t)) \big| y(t) - \tilde{y}(t)\big).$$

Par l'inégalité de Cauchy-Schwarz et le fait que f est globalement lipschitzienne par rapport à y, uniformément par rapport à t et de constante de Lipschitz L,

$$\big(f(t, y(t)) - f(t, \tilde{y}(t)) \big| y(t) - \tilde{y}(t)\big) \leq \|f(t, y(t)) - f(t, \tilde{y}(t))\| \|y(t) - \tilde{y}(t)\| \leq L\|y(t) - \tilde{y}(t)\|^2.$$

27. Thomas Hakon Grönwall, 1877–1932.

Posons $v(t) = \|y(t) - \tilde{y}(t)\|^2$, on a obtenu l'inéquation différentielle

$$v'(t) \leq 2Lv(t).$$

Le lemme de Grönwall avec $\alpha = 2L$ et $\beta = 0$ (ou le facteur intégrant qui s'impose) nous assure alors que

$$v(t) \leq v(0)e^{2Lt}.$$

Mais $v(0) = \|y(0) - \tilde{y}(0)\|^2$, d'où le résultat en prenant la racine carrée. ◊

Proposition C.3.8 (Cauchy-Lipschitz, unicité) *Sous les hypothèses du Théorème* C.3.4*, le problème de Cauchy* (2.1.3) *a au plus une solution.*

Démonstration. On applique la proposition C.3.7 à deux solutions y et $\tilde{y}$ du problème de Cauchy. Pour ces deux solutions, on a $y(0) = \tilde{y}(0) = y_0$, donc $\|y(0) - \tilde{y}(0)\| = 0$. ◊

La proposition C.3.7 a un intérêt propre, car c'est un résultat de continuité des solutions (encore éventuelles à ce stade) par rapport aux données initiales, comme dans le cas linéaire. En effet, l'estimation étant valable pour tout t, on en déduit en passant au maximum sur $[0, T]$ que

$$\|y - \tilde{y}\|_{C^0([0,T];\mathbb{R}^m)} \leq e^{LT}\|y(0) - \tilde{y}(0)\|.$$

Si l'on prend donc une suite de données initiales y_0^k telle que $y_0^k \to y_0$ quand $k \to +\infty$, alors la suite y^k des solutions correspondantes du problème de Cauchy converge uniformément sur $[0, T]$ vers la solution y correspondant à la donnée initiale y_0.

Ce résultat est de plus quantitatif. En effet, on voit que la différence des valeurs prises par deux solutions au temps t, avec des valeurs initiales distinctes, n'augmente pas plus vite en norme avec t, relativement à la différence initiale, que le facteur e^{Lt}. Une petite différence sur les valeurs initiales conduit à une différence au temps t qui est au plus exponentiellement amplifiée avec le temps. C'est ce que l'on appelle populairement l'*effet papillon* [28]. Attention, la preuve précédente ne montre pas que cet effet se produit, puisque ce n'est qu'une majoration. Il se trouve que l'effet a effectivement lieu pour certaines EDO, c'est-à-dire que la différence entre deux solutions de données initiales distinctes croît effectivement exponentiellement avec le temps, et pas pour d'autres. Et bien sûr, une exponentielle de ce type devient rapidement énorme, c'est-à-dire que l'évolution du système devient en pratique imprévisible car on ne connaît jamais exactement la condition initiale. C'est ce que l'on appelle le *chaos déterministe*. Il n'y a par ailleurs aucune contradiction entre cette imprévisibilité en pratique, si elle se produit, et la continuité par rapport aux données initiales précédemment évoquée, ni avec le déterminisme sous-jacent.

Une dernière remarque sur l'instant initial 0. Cette valeur spécifique n'a naturellement aucune importance. Il est bien clair que pour tout $t_0 \in [0, T]$, on a $\|y(t) - \tilde{y}(t)\| \leq \|y(t_0) - \tilde{y}(t_0)\|e^{L|t-t_0|}$ pour tout $t \in [0, T]$.

Prouvons maintenant l'*existence de solutions*, de la même manière que dans le cas linéaire. On part de la forme intégrale équivalente du problème.

Proposition C.3.9 *On suppose que la fonction second membre de l'EDO, f, est une fonction continue sur $\bar{I} \times \mathbb{R}^m$. Soit y une solution du problème de Cauchy* (2.1.3)*, continue sur $[0, T]$. Alors on a pour tout $t \in [0, T]$*

$$y(t) = y_0 + \int_0^t f\big(s, y(s)\big)\, ds. \tag{C.3.1}$$

Réciproquement, soit y une fonction continue sur $[0, T]$ à valeurs dans $\mathbb{R}^m$ et satisfaisant l'équation intégrale (C.3.1)*. Alors, y est dérivable sur I et solution du problème de Cauchy* (2.1.3)*.*

Démonstration. La démonstration est strictement identique au cas linéaire, voir proposition C.2.17. ◊

Les deux formulations, problème de Cauchy et équation intégrale, sont donc équivalentes.

28. Sauf que dans le cas du papillon, ce sont des EDP et non des EDO qui sont en cause. Mais la problématique est la même, c'est juste beaucoup plus compliqué.

Proposition C.3.10 (Cauchy-Lipschitz, existence) *Sous les hypothèses du théorème* C.3.4*, le problème de Cauchy* (2.1.3) *a au moins une solution.*

Démonstration. On utilise encore la méthode des approximations successives de Picard. [29] On définit donc une suite de fonctions continues $(y_n)_{n\in\mathbb{N}}$, en posant pour tout $t \in [0,T]$,

$$y_0(t) = y_0, \quad y_{n+1}(t) = y_0 + \int_0^t f(s, y_n(s))\, ds.$$

Il vient d'abord

$$y_1(t) - y_0(t) = \int_0^t f(s, y_0)\, ds,$$

puis pour tout $n \geq 1$

$$y_{n+1}(t) - y_n(t) = \int_0^t \big(f(s, y_n(s)) - f(s, y_{n-1}(s))\big)\, ds.$$

Comme $f(t,x)$ est continue par hypothèse, la fonction $s \mapsto \|f(s, y_0)\|$ est continue sur $[0,T]$ et donc bornée puisque l'intervalle $[0,T]$ est compact. En notant M un majorant de cette fonction, on obtient pour tout $t \in [0,T]$,

$$\|y_1(t) - y_0(t)\| \leq \int_0^t \|f(s, y_0)\|\, ds \leq Mt.$$

Ensuite, en utilisant le caractère lipschitzien de f, on voit que pour tout $t \in [0,T]$,

$$\|y_{n+1}(t) - y_n(t)\| \leq \int_0^t \big\|f(s, y_n(s)) - f(s, y_{n-1}(s))\big\|\, ds \leq L \int_0^t \|y_n(s) - y_{n-1}(s)\|\, ds.$$

On en déduit par récurrence sur n, exactement comme dans le cas linéaire, [30] que

$$\forall t \in [0,T], \quad \|y_{n+1}(t) - y_n(t)\| \leq \frac{M}{L} \frac{(Lt)^{n+1}}{(n+1)!}.$$

Il s'ensuit comme précédemment que la série $\sum_{n\geq 1}(y_n - y_{n-1})$ est normalement convergente dans l'espace complet $C^0([0,T];\mathbb{R}^m)$. Ceci implique de la même façon que la suite y_n converge uniformément sur $[0,T]$ vers une fonction continue $y\colon [0,T] \to \mathbb{R}^m$.

Comme $\|f(s, y_n(s)) - f(s, y(s))\| \leq L\|y_n(s) - y(s)\|$ car f est globalement lipschitzienne, on en déduit que la suite de fonctions continues $\big(s \mapsto f(s, y_n(s))\big)_n$ converge uniformément vers la fonction continue $s \mapsto f(s, y(s))$ sur $[0,T]$. On peut donc passer à la limite dans l'intégrale du membre de droite de la définition de y_{n+1} et obtenir, pour tout $t \in [0,T]$,

$$y(t) = y_0 + \int_0^t f(s, y(s))\, ds,$$

puisque le membre de gauche tend uniformément vers y. Par la proposition C.3.9, on en déduit que y est solution du problème de Cauchy. ◊

Le théorème de Cauchy-Lipschitz global est donc maintenant démontré par la conjonction des propositions C.3.8 et C.3.10. La solution est dite globale car elle existe sur la totalité de l'intervalle d'étude $[0,T]$.

Répétons ici que le fait de choisir l'instant initial à $t = 0$ n'a aucune importance, pas plus que le sens du temps d'ailleurs.

Corollaire C.3.11 *i) Si $f\colon [t_0,T] \times \mathbb{R}^m \to \mathbb{R}^m$ satisfait les hypothèses de Cauchy-Lipschitz global, alors le problème de Cauchy $y'(t) = f(t, y(t))$ sur $]t_0,T[$, $y(t_0) = y_0$, admet une solution unique pour tout $y_0 \in \mathbb{R}^m$.*

ii) Si $f\colon [-T,0] \times \mathbb{R}^m \to \mathbb{R}^m$, $T > 0$, satisfait les hypothèses de Cauchy-Lipschitz global, alors le problème de Cauchy rétrograde $y'(t) = f(t, y(t))$ sur $]-T,0[$, $y(0) = y_0$, admet une solution unique pour tout $y_0 \in \mathbb{R}^m$.

29. D'ailleurs, en anglais le théorème de Cauchy-Lipschitz est connu sous le nom de *Picard-Lindelöf theorem*. Ernst Leonard Lindelöf, 1870–1946.

30. L joue le rôle du α du cas linéaire et M celui de $\alpha\|y_0\| + \beta$. Ces rôles sont bien évidents.

Démonstration. i) Introduisons le changement de variable qui s'impose, $s = t - t_0$. Soit $g\colon [0, T-t_0]\times\mathbb{R}^m \to \mathbb{R}^m$ définie par $g(s, y) = f(s + t_0, y)$. Il est bien clair que g satisfait les hypothèses du théorème de Cauchy-Lipschitz global [31], donc le problème de Cauchy $z'(s) = g(s, z(s))$, $z(0) = y_0$, admet une solution unique, qui engendre la solution unique du problème de départ $y(t) = z(t - t_0)$.

ii) Le changement de variable qui s'impose ici est $s = -t$. Soit $g\colon [0, T] \times \mathbb{R}^m \to \mathbb{R}^m$ définie par $g(s, y) = -f(-s, y)$. Il est bien clair que g satisfait les hypothèses du théorème de Cauchy-Lipschitz global [32], donc le problème de Cauchy $z'(s) = g(s, z(s))$, $z(0) = y_0$, admet une solution unique, qui engendre la solution unique du problème de départ $y(t) = z(-t)$. ◊

Notons une conséquence importante du théorème de Cauchy-Lipschitz, qui est que deux courbes intégrales distinctes ne peuvent pas se croiser. En d'autres termes, si jamais deux courbes intégrales se croisent, alors elles sont en fait confondues et correspondent à la même solution.

Corollaire C.3.12 *Sous les hypothèses du théorème* C.3.4*, si deux solutions y et $\tilde{y}$ correspondant aux données initiales y_0 et $\tilde{y}_0$ sont telles qu'il existe $t_0 \in [0, T]$ tel que $y(t_0) = \tilde{y}(t_0)$, alors $y_0 = \tilde{y}_0$ et $y(t) = \tilde{y}(t)$ pour tout t.*

Démonstration. Si $t_0 = 0$, on conclut immédiatement par l'unicité de Cauchy-Lipschitz. Supposons $t_0 > 0$ et soit $z_0 = y(t_0) = \tilde{y}(t_0)$. On considère le problème de Cauchy

$$\begin{cases} z'(s) = -f(t_0 - s, z(s)), \text{ pour tout } s \in [0, t_0], \\ z(0) = z_0. \end{cases}$$

Ce problème relève évidemment du théorème de Cauchy-Lipschitz, il admet une solution et une seule sur $[0, t_0]$. On remarque que $z(s) = y(t_0 - s)$ et $\tilde{z}(s) = \tilde{y}(t_0 - s)$ en sont solution. En effet, $z'(s) = -y'(t_0 - s) = -f(t_0 - s, y(t_0 - s)) = -f(t_0 - s, z(s))$ et de même pour $\tilde{z}$. Il s'ensuit par l'unicité que $z = \tilde{z}$. En particulier pour $s = t_0$, il vient $y_0 = y(0) = z(t_0) = \tilde{z}(t_0) = \tilde{y}(0) = \tilde{y}_0$. Par unicité du problème de Cauchy initial, en repartant dans le sens normal du temps, on en déduit que $y = \tilde{y}$ sur $[0, T]$. ◊

On a simplement remonté le temps et décalé l'origine, c'est-à-dire les deux situations du corollaire immédiatement précédent en même temps... Dans le cas d'un système autonome, on voit même que deux orbites distinctes sont d'intersection vide, ce qui est plus fort que la non intersection des courbes intégrales. Pour une équation non autonome par contre, les orbites peuvent parfaitement se croiser.

C.4 existence locale d'une solution

Dans ce paragraphe, on traite de cas où l'existence des solutions n'est pas assurée sur tout l'intervalle d'étude. Cela peut naturellement se produire si les hypothèses du théorème de Cauchy-Lipschitz global ne sont pas satisfaites.

Introduisons quelques définitions. Pour fixer les idées, nous prendrons toujours l'instant initial à $t = 0$, mais il doit maintenant être bien clair que ceci n'a aucune sorte d'importance.

Définition C.4.1 *Soit f continue de $[0, T[\times \mathbb{R}^m$ dans $\mathbb{R}^m$, avec $T \in \mathbb{R}_+^* \cup \{+\infty\}$. On appelle solution locale du problème de Cauchy*

$$y'(t) = f(t, y(t)), \quad y(0) = y_0,$$

tout couple (I, y), où $I = [0, \tau[$, $\tau \leq T$, est un intervalle contenant 0 et y une fonction continue sur I à valeurs dans $\mathbb{R}^m$, dérivable sur $]0, \tau[$ et vérifiant l'EDO et la condition initiale.

Une solution locale, si elle existe, existe donc pour un certain laps de temps, mais pas nécessairement sur la totalité de l'intervalle en temps où la fonction second membre f est définie. Remarquons que l'on insiste sur la continuité en $t = 0$, indispensable pour que la notion même de condition initiale ait un sens, alors que l'on laisse beaucoup plus de mou à l'autre extrémité de l'intervalle I qui est ouvert en τ, tout comme l'intervalle en temps où f est définie qui est ouvert en T, T pouvant d'ailleurs très bien être égal à $+\infty$. À cet égard, c'est très

31. Si ce n'est pas clair, ne pas hésiter à le vérifier.
32. Si ce n'est pas clair, ...

différent de tout ce que l'on a raconté dans le contexte du théorème de Cauchy-Lipschitz global, où l'on était toujours sur un intervalle de temps *compact* $[0, T]$, $T < +\infty$, compacité qui a joué un rôle très important à plusieurs endroits dans les preuves.

L'intervalle I s'appelle le domaine de définition de la solution locale (I, y). On peut comparer les domaines de définition de différentes solutions locales avec les définitions suivantes.

Définition C.4.2

1. *On dit qu'une solution locale* (I, y) prolonge *la solution locale* (J, z) *si* $J \subset I$ *et* $y(t) = z(t)$ *pour tout* $t \in J$. *Si de plus* $J \neq I$, *on dit que* (I, y) *prolonge* strictement (J, z).
2. *On dit que la solution locale* (I, y) *est une* solution maximale *du problème de Cauchy s'il n'existe pas de solution locale qui la prolonge strictement.*
3. *On dit que* (I, y) *est une* solution globale *du problème de Cauchy si c'est une solution locale et si* $\tau = T$.

On retient qu'une solution locale en prolonge une autre si elles coïncident sur le domaine de définition de la deuxième, mais que la première continue à exister plus loin au sens large (vraiment strictement plus loin si le prolongement est strict). On retient également qu'une solution est maximale s'il est impossible de continuer plus loin. Une solution globale est évidemment maximale, mais l'inverse n'est pas vrai comme on en verra des exemples.

Attention à la petite équivoque du vocabulaire : dans le contexte présent, une solution globale est a priori seulement définie sur l'intervalle semi-ouvert $[0, T[$, alors que dans le contexte du théorème de Cauchy-Lipschitz global, elle était définie sur $[0, T]$ fermé en T. Ce n'est pas la première fois, ni la dernière fois, que le même mot a des significations (très légèrement) différentes suivant le contexte... On devrait ici parler de solution locale globale, ce qui est quand même légèrement bizarre. Du coup, on ne le fait pas.

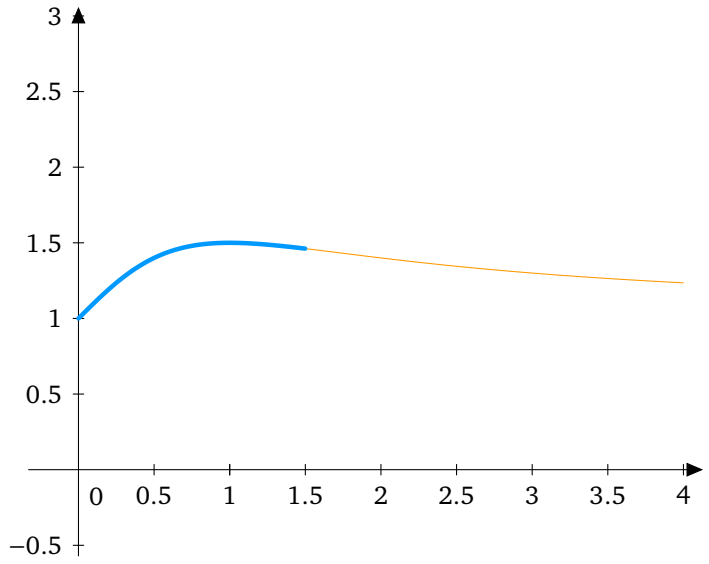


FIGURE C.2 – La solution orange sur $[0, 4]$ prolonge strictement la solution bleue sur $\left[0, \frac{3}{2}\right]$.

La notion d'unicité d'une solution locale est précisée par la définition suivante.

Définition C.4.3 *On dit que le problème de Cauchy sur* $[0, T[$,

$$y'(t) = f(t, y(t)), \quad y(0) = y_0,$$

admet une solution locale unique *si deux solutions locales quelconques coïncident sur l'intersection de leurs domaines de définition.*

Ce vocabulaire n'est pas très bon, puisque l'" unicité d'une solution locale " au sens de la définition précédente repose sur la coïncidence de plusieurs solutions locales sur des sous-intervalles... mais enfin, c'est le vocabulaire usuel, il faut vivre avec. On comprend le principe qui est derrière : s'il y a unicité locale, il n'y a pas de " branchement " des solutions locales, c'est-à-dire que si l'on prend deux solutions locales, l'une des deux prolonge forcément l'autre. A contrario, s'il n'y a pas unicité locale, de tels branchements peuvent se produire, voir figure C.5 plus loin pour un exemple de non unicité locale.

Proposition C.4.4 *Étant donnée une solution locale, il existe au moins une solution maximale qui la prolonge.*

Démonstration. Admis, voir [3], la difficulté étant dans la non unicité locale éventuelle. C'est en fait un résultat de nature ensembliste, sans grand intérêt pratique. ◊

Proposition C.4.5 *On suppose f continue sur $[0, T[\times \mathbb{R}^m$, $T \in \mathbb{R}_+^*$ ou $+\infty$. Si (I, y) est une solution maximale non globale, alors I est de la forme $[0, T_m[$, avec $T_m < T$, et y n'est pas bornée sur I.*

Démonstration. Admis, voir [3]. Nous montrerons ces deux résultats dans un cadre un peu plus restrictif plus loin. ◊

En conséquence de la Proposition C.4.5, une solution locale (I, y) telle que y est bornée sur $I = [0, \tau[$ avec $\tau < T$ n'est certainement pas maximale.

Exemple C.4.1 1. Le problème de Cauchy

$$y'(t) = -2ty(t)^2, \quad y(0) = 1$$

a pour fonction second membre $f(t, y) = -2ty^2$ qui est continue sur $[0, +\infty[\times \mathbb{R}$, mais pas globalement lipschitzienne. En intégrant l'équation par séparation des variables, on obtient la solution

$$y(t) = \frac{1}{1 + t^2} \text{ sur } [0, +\infty[.$$

La solution est globale puisque définie sur tout l'intervalle d'étude (elle est même définie sur $\mathbb{R}$ entier). On a ainsi illustré que la propriété d'être globalement lipschitzienne n'est pas nécessaire pour avoir une solution globale.
2. Le problème de Cauchy

$$y'(t) = 2ty(t)^2, \quad y(0) = 1$$

a pour fonction second membre $f(t, y) = 2ty^2$ qui est continue sur $[0, +\infty[\times \mathbb{R}$, mais pas globalement lipschitzienne. En intégrant l'équation par séparation des variables, on obtient la solution

$$y(t) = \frac{1}{1 - t^2} \text{ sur } [0, 1[.$$

On ne peut pas la prolonger au delà de 1 puisqu'elle tend vers $+\infty$ quand $t \to 1^-$. C'est donc une solution maximale et elle n'est pas globale. Notons que l'on peut aussi la prolonger pour $t < 0$ jusqu'à $t = -1$, mais pas plus loin.
3. Le problème de Cauchy

$$y'(t) = y(t)^2, \quad y(0) = 1$$

a pour fonction second membre $f(t, y) = y^2$ qui est continue sur $[0, +\infty[\times \mathbb{R}$, mais pas globalement lipschitzienne. En intégrant l'équation par séparation des variables, on obtient la solution

$$y(t) = \frac{1}{1 - t}.$$

Cette solution ne peut manifestement pas être prolongée au delà de $t = 1$. Le couple $([0, 1[, \frac{1}{1-t})$ est donc une solution locale maximale. Elle n'est pas globale. Par contre, pour $t < 0$, on peut la prolonger jusqu'à $-\infty$. ◊

Nous allons maintenant formuler une version du théorème de Cauchy-Lipschitz, plus générale que la version globale, et qui permet de prendre en compte les exemples précédents.

On considère donc maintenant le cas où $f(t,x)$ est seulement définie sur $[0,T[\times V$ où V est un ouvert connexe non vide de $\mathbb{R}^m$, et non pas sur $[0,T]\times\mathbb{R}^m$. Les diverses notions associées aux solutions locales ne sont pas modifiées. On généralise également la condition de lispchitzianité.

Définition C.4.6 *On dit que $f:[0,T[\times V\to\mathbb{R}^m$ est* localement lipschitzienne *par rapport à y, uniformément par rapport à t, si pour tout $y_0\in V$, il existe une boule fermée $\bar B$ contenue dans l'ouvert V centrée en y_0 et $\tau<T$, tels qu'il existe une constante L telle que, pour tout $y,z\in\bar B$ et $t\in[0,\tau]$,*

$$\|f(t,y)-f(t,z)\|\le L\|y-z\|.$$

En d'autres termes, f est lipschitzienne par rapport à y et uniformément par rapport à t sur $[0,\tau]\times\bar B$. Naturellement, la constante L dépend a priori de la boule $\bar B$ et du temps τ, même si on ne l'écrit pas explicitement. Nous allons montrer le théorème suivant.

Théorème C.4.7 (Cauchy-Lipschitz local) *Soit $f:[0,T[\times V\to\mathbb{R}^m$ continue et localement lipschitzienne par rapport à y, uniformément par rapport à t. Alors pour tout $y_0\in V$, le problème de Cauchy*

$$y'(t)=f(t,y(t)),\quad y(0)=y_0,$$

admet une unique solution locale.

On va en fait se ramener immédiatement au théorème global grâce au lemme de prolongement suivant.

Lemme C.4.8 *Soit $\bar B$ une boule fermée et $\tau<T$ un temps tels que la restriction de f à $[0,\tau]\times\bar B$ soit lipschitzienne par rapport à y uniformément par rapport à t. Alors cette restriction admet un prolongement à $[0,\tau]\times\mathbb{R}^m$ qui est continu et globalement lipschitzien par rapport à y uniformément par rapport à t.*

Admettons le lemme l'espace d'un instant.

Démonstration du théorème C.4.7. Prenons une boule $\bar B$ et un temps τ associés à y_0 et appelons $\tilde f$ le prolongement en question. Grâce au théorème global C.3.4, le problème de Cauchy $\tilde y'(t)=\tilde f(t,\tilde y(t))$, $\tilde y(0)=y_0$, admet une unique solution globale sur $[0,\tau]$. Comme y_0 est au centre de $\bar B$ et que $t\mapsto\tilde y(t)$ est une fonction continue, il s'ensuit qu'il existe $0<\tau_*\le\tau$ tel que $\tilde y(t)\in B$, où B désigne la boule ouverte dont $\bar B$ est l'adhérence, pour tout $t\in[0,\tau_*[$. Pour ces valeurs de t, on a donc $\tilde f(t,\tilde y(t))=f(t,\tilde y(t))$. Posant $y(t)=\tilde y(t)$ pour $t<\tau_*$, on a en fait construit une solution locale $([0,\tau_*[,y)$ de notre problème de Cauchy.

Montrons que celle-ci est unique au sens de la définition C.4.3. Soient $([0,\tau_1[,y_1)$ et $([0,\tau_2[,y_2)$ deux solutions locales avec $\tau_2\ge\tau_1$. Montrons que $y_{2|[0,\tau_1[}=y_1$. En effet, soit

$$\sigma=\sup\{s<\tau_1, y_1(t)=y_2(t)\text{ pour tout }0\le t\le s\}.$$

Supposons que $\sigma<\tau_1$. Par continuité, on a $y_1(\sigma)=y_2(\sigma)$. Toujours d'après la partie existence du raisonnement, le problème de Cauchy $z'(t)=f(t,z(t))$, $z(\sigma)=y_1(\sigma)$ admet une solution locale sur un intervalle [33] $[\sigma,\sigma+\alpha[$ avec $\alpha>0$, qui est unique puisqu'en fait, c'est encore le théorème global qui s'applique via un autre prolongement de f autour de $(\sigma,y_1(\sigma))$. Cela implique que $z(t)=y_1(t)=y_2(t)$ sur $[\sigma,\sigma+\alpha[$, contradiction avec le fait que σ soit la borne supérieure des intervalles contenant 0 où ceci a lieu. Par conséquent, $\sigma=\tau_1$, ce qu'il fallait démontrer. ◊

Au vu de cette démonstration, il est bien clair que bon nombre de propriétés globales, comme les corollaires C.3.7 et C.3.12, restent vraies localement dans ce contexte.

Il s'agit maintenant de montrer le lemme de prolongement, qui est en fait un résultat d'intérêt général.

33. On applique l'existence à partir de l'instant initial σ plutôt que 0, ce qui n'est clairement pas un problème.

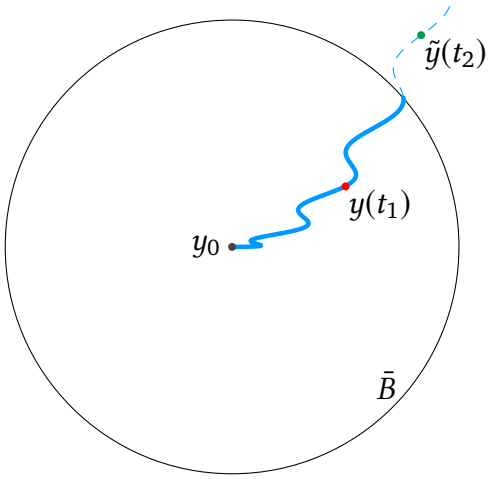


FIGURE C.3 – Tant que la solution du problème de Cauchy prolongé reste dans la boule où $\tilde{f}$ et f coïncident, c'est-à-dire pour $t_1 \leq \tau_*$, c'est une solution locale du problème de Cauchy de départ. Peu importe ce qu'elle fait ensuite pour $t_2 > \tau_*$, on la laisse vivre sa vie en pointillés.

Démonstration du lemme C.4.8. Il suffit de construire le prolongement composante par composante. Notons g une telle composante générique de la restriction de f à $[0,\tau]\times\bar{B}$. On a donc $g\colon [0,\tau]\times\bar{B}\to\mathbb{R}$ continue et telle que $|g(t,y)-g(t,z)|\leq L\|y-z\|$ pour tous t,y,z. On pose

$$\forall (t,y)\in[0,\tau]\times\mathbb{R}^m,\quad \tilde{g}(t,y)=\inf_{u\in\bar{B}}(g(t,u)+L\|y-u\|). \tag{C.4.1}$$

Montrons que $\tilde{g}$ convient. Tout d'abord, c'est bien un prolongement de g. En effet, si $y\in\bar{B}$, on a pour tout $u\in\bar{B}$,

$$g(t,y)\leq g(t,u)+|g(t,y)-g(t,u)|\leq g(t,u)+L\|y-u\|.$$

Comme $\tilde{g}(t,y)$ est le plus grand des minorants du terme de droite, il s'ensuit que $g(t,y)\leq\tilde{g}(t,y)$. D'un autre côté, comme $y\in\bar{B}$, on peut prendre $u=y$ au second membre de (C.4.1), ce qui montre que $\tilde{g}(t,y)\leq g(t,y)+L\|y-y\|=g(t,y)$. Par conséquent $\tilde{g}(t,y)=g(t,y)$ dans ce cas.

Prenons maintenant deux points y et z de $\mathbb{R}^m$. Pour tout t, l'application $u\mapsto g(t,u)+L\|z-u\|$ est continue. Elle atteint donc sa borne inférieure $\tilde{g}(t,z)$ sur le compact $\bar{B}$ en un point v. Il vient donc

$$\tilde{g}(t,y)-\tilde{g}(t,z)\leq g(t,v)+L\|y-v\|-g(t,v)-L\|z-v\|=L\|y-v\|-L\|z-v\|\leq L\|y-z\|,$$

par l'inégalité triangulaire. On obtient le fait que $\tilde{g}$ est globalement lipschitzienne, de constante L, uniformément par rapport à t, en inversant les rôles de y et z.

Il reste à voir que $\tilde{g}$ est continue. On sait déjà que $\tilde{g}$ est lipschitzienne par rapport à y uniformément par rapport à t. On a vu à la proposition C.3.5 qu'il suffit alors de montrer que l'application $t\mapsto\tilde{g}(t,y)$ est continue pour tout $y\in\mathbb{R}^m$ fixé. Donnons-nous $(t,y)\in[0,\tau]\times\mathbb{R}^m$. Comme plus haut, il existe $v\in\bar{B}$ tel que $\tilde{g}(t,y)=g(t,v)+L\|y-v\|$. Par conséquent, pour tout $t'\in[0,\tau]$,

$$\tilde{g}(t',y)-\tilde{g}(t,y)\leq g(t',v)+L\|y-v\|-g(t,v)-L\|y-v\|\leq|g(t',v)-g(t,v)|.$$

Or l'ensemble $[0,\tau]\times\bar{B}$ est compact, donc l'application g y est uniformément continue. Pour tout $\varepsilon>0$, il existe $\alpha>0$, indépendant de v, tel que si $|t'-t|\leq\alpha$, alors $|g(t',v)-g(t,v)|\leq\varepsilon$. On conclut en échangeant les rôles de t et t'. ◊

Le prolongement (C.4.1) s'appelle *prolongement de McShane-Whitney* [34] quand g ne dépend pas de t, voir figure C.4. [35] Notons que l'on peut donner une démonstration de la continuité du prolongement par rapport à

34. Edward James McShane, 1904 –1989 ; Hassler Whitney, 1907–1989.

35. À propos de cette figure, en dimension 1, le prolongement par des constantes égales aux valeurs aux extrémités est encore plus simple et marche également.

t utilisant des suites. En effet, soit une suite $t_n \to t$ et un point $y \in \mathbb{R}^m$. On note $v \in \bar{B}$ un point réalisant la borne inférieure pour (t, y) et v_n réalisant cette même borne pour (t_n, y). On a

$$\tilde{g}(t_n, y) - \tilde{g}(t, y) \leq g(t_n, v) + L\|y - v\| - g(t, v) - L\|y - v\| = g(t_n, v) - g(t, v),$$

donc $\limsup_{n\to+\infty} \tilde{g}(t_n, y) \leq \tilde{g}(t, y)$ (c'est-à-dire que $\tilde{g}$ est semi-continue supérieurement, ce qui est toujours vrai pour un inf de fonctions continues). D'un autre côté, on a également

$$\tilde{g}(t, y) - \tilde{g}(t_n, y) \leq g(t, v_n) + L\|y - v_n\| - g(t_n, v_n) - L\|y - v_n\| = g(t, v_n) - g(t_n, v_n).$$

Extrayons une sous-suite $n_p \to +\infty$ telle que $\tilde{g}(t_{n_p}, y) \to \liminf_{n\to+\infty} \tilde{g}(t_n, y)$ quand $p \to +\infty$. La suite v_{n_p} reste dans le compact $\bar{B}$, on peut en extraire une sous-suite $n_{p_q} \to +\infty$ telle que $v_{n_{p_q}} \to w$ quand $q \to +\infty$ pour un certain $w \in \bar{B}$. Comme $(t, v_{n_{p_q}}) \to (t, w)$ et $(t_{n_{p_q}}, v_{n_{p_q}}) \to (t, w)$ et que g est continue, on déduit de l'inégalité précédente restreinte à la suite extraite n_{p_q} que $\tilde{g}(t, y) \leq \liminf_{n\to+\infty} \tilde{g}(t_n, y)$. Par conséquent la limite supérieure et la limite inférieure coïncident avec $\tilde{g}(t, y)$.

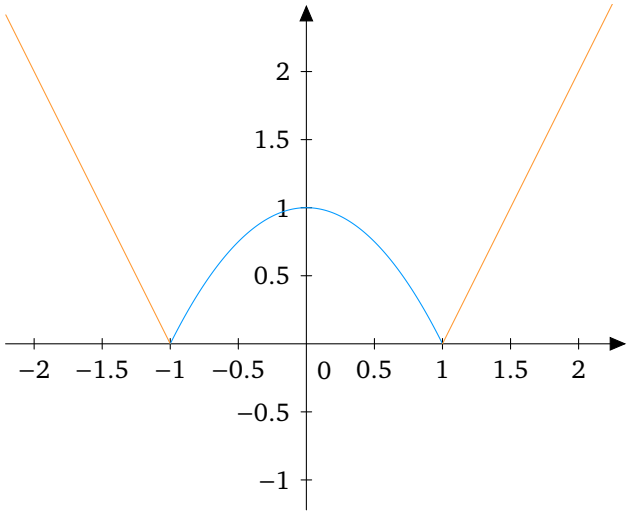


FIGURE C.4 – Le prolongement de McShane-Whitney de la fonction $y \mapsto 1 - y^2$ à l'extérieur de la boule de dimension un, $\bar{B} = [-1, 1]$.

Pour pouvoir appliquer le théorème de Cauchy-Lipschitz local, nous avons également besoin de moyens pratiques d'en vérifier les hypothèses On aura évidemment commencé par regarder si on ne peut pas appliquer le théorème de CL global. A ce sujet un critère facile à vérifier dans le cas autonome, est le caractère non borné des dérivées partielles de la fonction second membre, la contra posée de la propriété C.3.2 en quelque sorte :

Proposition C.4.9 *Supposons que $f(y)$ possède des dérivées partielles continues par rapport à y_i, $i = 1, \ldots, m$, et qu'au moins une de ces dérivées est non bornée sur $\mathbb{R}^m$. Alors f n'est pas globalement lipschitzienne.*

Démonstration. On suppose que $\frac{\partial f_j}{\partial y_i}(y)$ existe, définit une fonction continue par rapport à y et est non bornée sur $\mathbb{R}^m$. C'est-à-dire que pour tout $C > 0$ il existe $y \in \mathbb{R}^m$ tel que $\left\|\frac{\partial f_j}{\partial y_i}(y)\right\| > C$. On a donc pour $z = y + \varepsilon e_i$, avec e_i le vecteur unitaire dans la direction i et $\varepsilon \geq 0$, la formule de Taylor

$$f_j(z) - f_j(y) = \varepsilon \frac{\partial f_j}{\partial y_i}(y + \eta e_i), \quad \text{avec } \eta \in [0, \varepsilon].$$

On peut choisir ε suffisamment petit pour que par continuité de $\frac{\partial f_j}{\partial y_i}(y)$ on ait

$$\|f(z) - f(y)\| \geq |f_j(z) - f_j(y)| = \varepsilon \left|\frac{\partial f_j}{\partial y_i}(y + \eta e_i)\right| > C\varepsilon = C\|z - y\|.$$

◊

On dira qu'une fonction g définie sur $[0,T[\times V$ est *localement bornée* si pour tout $y_0 \in V$, il existe $0 < \tau < T$, une boule fermée $\bar{B}$ de centre y_0 incluse dans V, et un nombre M tel que $\|g(t,y)\| \leq M$ pour tous $(t,y) \in [0,\tau] \times \bar{B}$.

Proposition C.4.10 *Soit f définie sur $[0,T[\times V$, différentiable par rapport y pour tout t et dont les dérivées partielles par rapport à y_i, $i = 1,\dots,m$, sont localement bornées. Alors f est localement lipschitzienne par rapport à y, uniformément par rapport à t.*

Démonstration. Il s'agit essentiellement de la même démonstration que la proposition C.3.2. Soient τ, $\bar{B}$ et M tels que $\left\|\frac{\partial f}{\partial y_i}(t,y)\right\| \leq M$ sur $[0,\tau] \times \bar{B}$. Donnons-nous y et z dans $\bar{B}$. Comme la boule est convexe, on voit que pour tout $s \in [0,1]$, on a $sy + (1-s)z \in \bar{B}$. On peut donc définir $g\colon [0,1] \to \mathbb{R}^m$ par $g(s) = f(t, sy + (1-s)z)$. On voit que g est dérivable, par dérivation des fonctions composées, avec

$$g'(s) = \sum_{i=1}^{m} \frac{\partial f}{\partial y_i}\big(t, sy + (1-s)z\big)(y-z)_i.$$

Toujours par convexité de la boule, $\left\|\frac{\partial f}{\partial y_i}(t, sy + (1-s)z)\right\| \leq M$. Par conséquent,

$$\|g'(s)\| \leq M \sum_{i=1}^{m} |(y-z)_i| \leq M\sqrt{m}\|y-z\|,$$

par l'inégalité de Cauchy-Schwarz. L'inégalité des accroissements finis implique alors que

$$\|f(t,y) - f(t,z)\| = \|g(1) - g(0)\| \leq M\sqrt{m}\|y-z\|,$$

et f est localement lipschitzienne uniformément par rapport à t. ◊

Quand on lui ajoute l'hypothèse de continuité par rapport au couple (t,y), la proposition C.4.10 donne une condition suffisante pour pouvoir appliquer le théorème C.4.7 de Cauchy-Lipschitz local. Ce n'est pas du tout une condition nécessaire. On a une condition suffisante encore plus facile à vérifier.

Proposition C.4.11 *Soit $f\colon [0,T[\times V \to \mathbb{R}^m$ de classe C^1. Alors f continue et localement lipschitzienne par rapport à y, uniformément par rapport à t.*

Démonstration. La fonction f est C^1, donc continue. De même, les dérivées partielles de f par rapport à y sont des fonctions continues. Elles sont donc bornées sur tout compact de la forme $[0,\tau] \times \bar{B}$ inclus dans $[0,T[\times V$. Il suffit alors d'appliquer la proposition C.4.10. ◊

La proposition s'applique en particulier aux équations autonomes pour lesquelles f ne dépend pas de t. Il suffit donc dans ce cas que l'application f soit C^1 de V dans $\mathbb{R}^m$.

Nous pouvons maintenant assez facilement démontrer les propositions C.4.4 et C.4.5, avec en plus l'unicité, sous les hypothèses du théorème de Cauchy-Lipschitz local, bien que ces deux propositions soient valables dans un cadre plus vaste.

Proposition C.4.12 *Sous les hypothèses du théorème de Cauchy-Lipschitz local, tout problème de Cauchy admet une solution maximale unique y_m définie sur $[0,T_m[$ avec $T_m \leq T$. De plus, si $T_m < T$, alors il existe une suite $t_k \to T_m^-$ telle que $\|y_m(t_k)\| \to +\infty$ quand $k \to +\infty$.*

Démonstration. Soit $([0,\tau[, y)$ une solution locale fournie par le théorème de Cauchy-Lipschitz local C.4.7. Soit S l'ensemble des solutions locales qui prolongent y. Par l'unicité locale, dans tout couple d'éléments de S, l'un des éléments prolonge l'autre. Posons $T_m = \sup\{\sigma; ([0,\sigma[, z) \in S\}$. Pour $t < T_m$, on définit sans ambiguité $y_m(t) = z(t)$ pour n'importe quel z tel que $([0,\sigma[, z) \in S$ avec $\sigma \geq t$, puisque toutes ces valeurs coïncident par la remarque précédente. Clairement, $([0,T_m[, y_m)$ est encore une solution locale du problème de Cauchy. Elle

est maximale, car si elle ne l'était pas, l'existence d'un prolongement strict contredirait la définition de T_m comme borne supérieure. Elle est bien sûr unique, puisque deux solutions maximales appartenant à S, l'une prolonge nécessairement l'autre, donc elles sont égales.

Enfin, dans le cas où $T_m < T$, supposons que $\|y_m(t)\| \leq R$ pour tout $t < T_m$ et pour un certain $R < +\infty$. Soit $\bar{B}_R$ la boule fermée de centre 0 et de rayon R. Comme la fonction f est continue et que $[0, T_m] \times \bar{B}_R$ est compact, il s'ensuit que $\|y'_m(t)\|$ est borné sur $[0, T_m[$. Par l'inégalité des accroissements finis, on en déduit que y_m est uniformément continue sur $[0, T_m[$. Elle admet donc un unique prolongement continu à $[0, T_m]$ qui vaut un certain $\bar{y} \in \bar{B}_R$ en $t = T_m$. Le problème de Cauchy $z'(t) = f(t, z(t))$, $z(T_m) = \bar{y}$ admet alors une unique solution locale sur un intervalle $]T_m - \alpha, T_m + \alpha[$ pour un certain $\alpha > 0$. On connaît déjà une solution sur $]T_m - \alpha, T_m]$, le prolongement de y_m par continuité (regardé de manière rétrograde). Par conséquent, la fonction $\bar{y}_m(t) = y_m(t)$ pour $t < T_m$, $\bar{y}_m(t) = z(t)$ pour $T_m \leq t < T_m + \alpha$ est une solution locale du problème de Cauchy de départ, qui prolonge strictement y_m, contradiction. ◊

Rappelons la remarque utile suivante, déjà faite plus haut, si une solution locale non globale est bornée, alors elle n'est pas maximale. Par ailleurs, une solution globale peut être bornée ou non bornée, on ne peut rien en dire à ce sujet a priori.

On a un résultat d'existence plus général, puisqu'on enlève une des hypothèses, donné ici sans démonstration, voir [3], et dont l'importance pratique est moindre.

Théorème C.4.13 (Peano) *On suppose la fonction f continue au voisinage du point $(0, y_0)$. Alors le problème de Cauchy* (2.1.3) *admet une solution locale.*

Démonstration. La preuve de ce théorème se trouve par exemple dans [3]. ◊

Attention, l'unicité locale de la solution est par contre perdue, comme on le voit sur l'exemple suivant.

Exemple C.4.2 On considère le problème de Cauchy $y'(t) = \sqrt{|y(t)|}$, $y(0) = 0$. La fonction second membre $f(t, y) = \sqrt{|y|}$ est continue sur $\mathbb{R} \times \mathbb{R}$, mais pas localement lipschitzienne par rapport à y au voisinage de $y = 0$. Le théorème de Peano [36] s'applique, pas celui de Cauchy-Lipschitz. En résolvant l'équation à variables séparées, on s'aperçoit en fait que pour tout $c \geq 0$, la fonction

$$y(t) = 0 \text{ pour } t \leq c, y(t) = (t - c)^2/4, \text{ pour } t > c,$$

est solution du problème de Cauchy. Par ailleurs, la fonction nulle est aussi solution du problème de Cauchy. Il y a une infinité (non dénombrable) de solutions, la valeur de c où une solution décolle de 0, si elle en décolle, n'est pas déterminée par le problème de Cauchy, voir Figure C.5. ◊

Il existe un théorème d'existence encore plus général, le théorème de Carathéodory [37] qui relâche les conditions de continuité du second membre f par rapport à la variable t, mais il est préférable de le passer ici sous silence... Par ailleurs, le théorème de Cauchy-Lipschitz se généralise pour l'existence et l'unicité dans d'autres directions, par exemple sur des variétés (des objets géométriques qui généralisent en toute dimension les courbes et surfaces du plan et de l'espace) ou encore en dimension infinie. Le théorème de Peano est par contre faux en dimension infinie.

C.4.1 Existence globale à l'aide des fonctions de Liapounov

Quand f est seulement localement lipschitzienne, mais pas globalement lipschitzienne, [38] l'existence d'une solution globale n'est pas assurée par le théorème de Cauchy-Lipschitz. D'ailleurs on a vu des exemples où il n'en existe pas. Dans un certain nombre d'applications, on parvient tout de même à montrer l'existence de solutions globales s'il existe une fonction dite de Liapounov [39] associée à l'équation différentielle. On se place dans le cas où $V = \mathbb{R}^m$ pour simplifier.

36. Giuseppe Peano, 1858–1932.
37. Constantin Carathéodory, 1873–1950.
38. Qui est une condition quand même bien restrictive.
39. Alexandre Mikhaïlovitch Liapounov, 1857–1918.

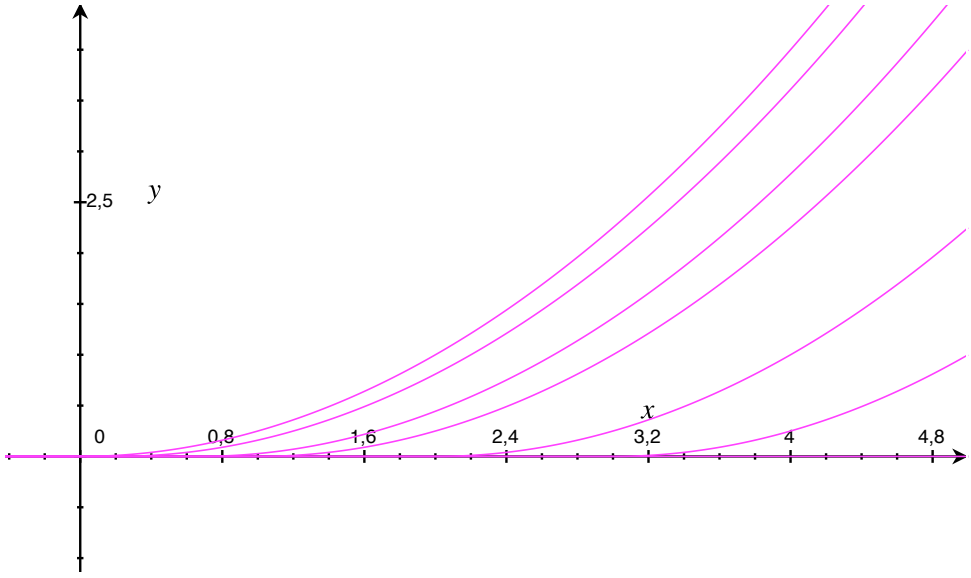


FIGURE C.5 – Quelques-unes parmi l'infinité (non dénombrable) des solutions du problème de Cauchy $y'(t) = \sqrt{|y(t)|}$, $y(0) = 0$, avec une infinité de branchements contredisant l'unicité locale.

Définition C.4.14 *Soit U une fonction de $\mathbb{R}^m$ dans $\mathbb{R}_+$, continûment différentiable. On dit que U est une fonction de Liapounov pour l'équation différentielle $y'(t) = f(t, y(t))$ si*

1. *$U(y) \longrightarrow +\infty$ quand $\|y\| \longrightarrow +\infty$,*
2. *il existe deux constantes $\alpha \geq 0$ et $\beta \geq 0$ telles que pour tout $t \in [0, T]$ et $y \in \mathbb{R}^m$*

$$dU(y)f(t, y) \leq \alpha U(y) + \beta,$$

où $dU(y)$ désigne la différentielle de la fonction U au point y (c'est une forme linéaire sur $\mathbb{R}^m$, représentée dans la base duale de la base canonique par le vecteur ligne des dérivées partielles de U).

Lorsque $\alpha = \beta = 0$ et que $dU(y)f(t, y) < 0$ quand $f(t, y) \neq 0$, on dit que U est une fonction de Liapounov *au sens strict* pour l'équation différentielle.

Il suffit souvent de prendre la fonction $U(y) = \|y\|^2$, qui vérifie clairement 1. La différentielle de U est donnée par $dU(y)z = 2(y|z)$ pour tout $z \in \mathbb{R}^m$, et l'on regarde alors s'il existe des constantes α et β positives telles que pour tout $t \in [0, T]$ et $y \in \mathbb{R}^m$, on ait

$$(y|f(t, y)) \leq \alpha\|y\|^2 + \beta.$$

C'est le cas par exemple de $f(t, y) = -y^3$, pour $m = 1$, second membre qui n'entre pas dans le cadre globalement lipschitzien. Par contre, ce n'est pas le cas pour $f(t, y) = y^3$.

Notons que la condition 2. est automatiquement satisfaite avec la fonction $U(y) = \|y\|^2$, s'il existe une constante C telle que, pour tout $t \in [0, T]$ et $y \in \mathbb{R}^m$

$$\|f(t, y)\| \leq C(1 + \|y\|).$$

Cette dernière condition est elle-même automatiquement satisfaite si f est globalement lipschitzienne en y, uniformément en t (mais on n'a pas besoin de fonctions de Liapounov dans ce cas...).

L'intérêt de l'existence d'une fonction de Liapounov pour les questions qui nous intéressent ici est le suivant.

Proposition C.4.15 *S'il existe une fonction de Liapounov U pour l'équation différentielle $y'(t) = f(t, y(t))$, où $f(t, y)$ est continue et localement lipschitzienne en y uniformément en t, alors pour toute donnée initiale $y_0 \in \mathbb{R}^m$, la solution du problème de Cauchy est globale.*

Démonstration. On sait d'après le théorème de Cauchy-Lipschitz local et d'après la proposition C.4.4 qu'il existe une solution locale maximale y du problème de Cauchy. Supposons qu'elle ne soit pas globale. Dans ce cas d'après le lemme C.4.5, la solution y n'est pas bornée sur son intervalle de définition $I = [0, T_m[$ avec $T_m < +\infty$. Cela signifie qu'il existe une suite t_n d'instants de $[0, T_m[$ tels que $t_n \to T_m$ et $\|y(t_n)\| \to +\infty$ quand $n \to +\infty$.

Considérons la fonction $g(t) = U(y(t))$. Elle est non bornée sur I par la condition 1. de la définition des fonctions de Liapounov, en effet $g(t_n) \to +\infty$. Par dérivation des fonctions composées, on a par ailleurs

$$g'(t) = dU(y(t))y'(t) = dU(y(t))f(t, y(t)) \leq \alpha U(y(t)) + \beta = \alpha g(t) + \beta,$$

en utilisant la condition 2. de la définition. Si $\alpha = 0$, on en déduit que $g(t) \leq U(y_0) + \beta T_m$ et si $\alpha \neq 0$ on utilise le lemme de Grönwall C.3.6 pour montrer que

$$g(t) \leq \left(U(y_0) + \frac{\beta}{\alpha}\right) e^{\alpha T_m} - \frac{\beta}{\alpha}.$$

Dans les deux cas, on obtient que g est majorée sur I, ce qui est une contradiction. ◊

Notons qu'une EDO pour laquelle certains problèmes de Cauchy n'admettent pas de solution globale ne peut en aucun cas posséder une fonction de Liapounov. Dans le cas d'une fonction de Liapounov au sens strict, on voit d'après le calcul qui précède que la fonction $t \mapsto U(y(t))$ est strictement décroissante tant que $y'(t)$ ne s'annule pas. Dans le cas d'une équation autonome où f ne dépend pas de t, cela implique donc que cette fonction est strictement décroissante sur tout trajectoire sauf sur celles qui correspondent aux points d'équilibre, *i.e.*, les points y tels que $f(y) = 0$, voir aussi définition C.1.5, où elle n'a pas vraiment d'autre choix que d'être constante.

Exemple C.4.3 Équations de type gradient. Soit U une fonction continûment différentiable de $\mathbb{R}^m$ dans $\mathbb{R}$. Le gradient de U au point y est le vecteur de $\mathbb{R}^m$

$$\nabla U(y) = \begin{pmatrix} \frac{\partial U}{\partial y_1}(y) \\ \vdots \\ \frac{\partial U}{\partial y_m}(y) \end{pmatrix}.$$

Il sert à représenter la différentielle $dU(y)$ de U au point y, qui est une forme linéaire et qui elle est représentée par la matrice ligne [40] (voir [2])

$$dU(y) = \begin{pmatrix} \frac{\partial U}{\partial y_1}(y) & \cdots & \frac{\partial U}{\partial y_m}(y) \end{pmatrix},$$

avec le produit scalaire canonique de $\mathbb{R}^m$

$$dU(y)z = (\nabla U(y)|z).$$

Définition C.4.16 *Une équation différentielle autonome est* de type gradient *s'il existe une fonction U de $\mathbb{R}^m$ dans $\mathbb{R}$, deux fois continûment différentiable, telle que pour tout $y \in \mathbb{R}^m$,*

$$f(t, y) = -\nabla U(y).$$

Dans ce cas, si U vérifie la condition 1 de la définition C.4.14, c'est une fonction de Liapounov au sens strict pour l'EDO, puisqu'alors, pour tout $y \in \mathbb{R}^m$,

$$dU(y)f(t, y) = -\|\nabla U(y)\|^2 \leq 0.$$

On obtient donc l'existence d'une solution globale y du problème de Cauchy. De plus, on voit que la fonction $t \mapsto U(y(t))$ est décroissante.

40. Sa matrice jacobienne en fait, notée ∇U tout au début. Une petite incohérence locale de notation, pas si grave que ça. On a tendance à utiliser ∇ pour signifier le vecteur gradient dans le cas d'une fonction scalaire et ∇ pour signifier matrice jacobienne pour une fonction à valeurs vectorielles, même si celle-ci est la transposée du gradient dans le cas scalaire.

L'exemple le plus simple est celui des systèmes linéaires autonomes sur $\mathbb{R}^m$, $y'(t) = Ay(t)$ où A est une matrice symétrique définie négative, ce qui correspond à poser

$$f(y) = -\nabla U(y) \quad \text{avec} \quad U(y) = -\frac{1}{2}(Ay|y).$$

La plupart des EDO autonomes, même linéaires, ne sont pas de type gradient : dans l'exemple **??**, l'équation linéarisée du pendule n'est pas de type gradient.

En revanche, on peut facilement construire des équations de type gradient non triviales, par exemple

$$m = 2, \quad f(t, y) = -\begin{pmatrix} 2y_1 e^{y_2} + y_2^2 e^{y_1} \\ 2y_2 e^{y_1} + y_1^2 e^{y_2} \end{pmatrix},$$

qui correspond à $U(y) = y_1^2 e^{y_2} + y_2^2 e^{y_1}$.

Exemple C.4.4 Équations de Hamilton [41]. Il s'agit au départ d'une reformulation extrêmement importante des lois de la mécanique classique.

Soit H une fonction deux fois continûment différentiable de $\mathbb{R}^{2m}$ dans $\mathbb{R}$. On note la variable d'espace $y = \begin{pmatrix} q \\ p \end{pmatrix}$, q désignant le vecteur des m premières coordonnées (dites de position en mécanique) et p celui des m suivantes (dites d'impulsion). On pose, pour tout $y \in \mathbb{R}^{2m}$,

$$f(y) = J\nabla H(y),$$

où J est l'opérateur linéaire de $\mathbb{R}^{2m}$ dans $\mathbb{R}^{2m}$, défini par

$$J\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} 0_m & I_m \\ -I_m & 0_m \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} p \\ -q \end{pmatrix},$$

où 0_m désigne la matrice nulle $m \times m$ et I_m l'identité $m \times m$. On dit que H est l'*hamiltonien* de l'équation différentielle $y'(t) = f(y(t))$, qui est elle qualifiée de *système hamiltonien*. Elle s'écrit donc en fonction des variables q et p sous la forme

$$\begin{cases} \dot{q} = \nabla_p H(q, p) \\ \dot{p} = -\nabla_q H(q, p), \end{cases}$$

où ∇_q et ∇_p désignent les gradients partiels par rapport à q et p.

L'opérateur J est antisymétrique : on a

$$(y|Jy)_{2m} = (q|p)_m - (q|p)_m = 0$$

pour tout $y = \begin{pmatrix} q \\ p \end{pmatrix} \in \mathbb{R}^{2m}$, où $(\cdot|\cdot)_k$ désigne le produit scalaire canonique sur $\mathbb{R}^k$.

Là encore si H vérifie la condition 1 de la définition C.4.14, elle constitue une fonction de Liapounov pour l'EDO avec $\alpha = \beta = 0$, puisque

$$dH(q, p)f(q, p) = \big(\nabla H(q, p)\big|J\nabla H(q, p)\big) = 0.$$

Dans ce cas, on a donc existence globale sur $[0, +\infty[$ des solutions du problème de Cauchy pour toute donnée initiale. De plus, on a la propriété de *conservation de l'hamiltonien* car pour toute solution $y(t) = (q(t), p(t))$ du système, on a évidemment

$$\frac{d}{dt}H(y(t)) = \big(\nabla H(q(t), p(t))\big|J\nabla H(q(t), p(t))\big) = 0,$$

donc pour tout $t \in [0, +\infty[$, $H(y(t)) = H(y_0)$. On dit que l'hamiltonien est une *intégrale première*. Une interprétation physique peut être qu'il représente l'énergie du système, qui est conservée au cours du mouvement.

41. Sir William Rowan Hamilton, 1805–1865.

Plus généralement, si $A\colon \mathbb{R}^{2m} \to \mathbb{R}$ est une *observable*, c'est-à-dire une fonction suffisamment régulière sur l'espace des phases [42], alors on a

$$\begin{aligned}\frac{d}{dt}A(y(t)) &= \sum_{i=1}^{m}\left(\frac{\partial A}{\partial q_i}(y(t))\frac{dq_i}{dt}(t) + \frac{\partial A}{\partial p_i}(y(t))\frac{dp_i}{dt}(t)\right)\\ &= \sum_{i=1}^{m}\left(\frac{\partial A}{\partial q_i}(y(t))\frac{\partial H}{\partial p_i}(y(t)) - \frac{\partial A}{\partial p_i}(y(t))\frac{\partial H}{\partial q_i}(y(t))\right)\\ &= \big\{A, H\big\}(y(t)),\end{aligned}$$

relation écrite parfois un peu rapidement à la physicienne $\frac{dA}{dt} = \{A, H\}$, où l'expression

$$\{A, B\} = \sum_{i=1}^{m}\left(\frac{\partial A}{\partial q_i}\frac{\partial B}{\partial p_i} - \frac{\partial A}{\partial p_i}\frac{\partial B}{\partial q_i}\right)$$

s'appelle le *crochet de Poisson* [43] de A et de B. Le crochet de Poisson est une notion d'une grande importance en mécanique classique, puis en mécanique quantique entre autres. Notons en particulier que $\{H, H\} = 0$, $\{q_i, q_j\} = \{p_i, p_j\} = 0$ et $\{q_i, p_j\} = \delta_{ij}$.

Comme exemple, prenons le cas des oscillations du pendule (voir plus loin exemple **??**). En posant $q = y_1$ et $p = y_2$ on a $\dot{p} = -k\sin(q) = -V'(q)$ avec $V(q) = -k\cos(q)$ et $\dot{q} = p = T'(p)$ avec $T(p) = p^2/2$. Du point de vue physique et à une constante multiplicative près, T est l'énergie cinétique, V l'énergie potentielle et le hamiltonien $H = T + V$ représente l'énergie totale qui est bien conservée en l'absence de frottements. On a bien

$$\begin{cases}\dot{q} = \dfrac{\partial H}{\partial p},\\ \dot{p} = -\dfrac{\partial H}{\partial q},\end{cases}$$

ce qui peut aussi s'écrire

$$\frac{d}{dt}\begin{pmatrix}q\\p\end{pmatrix} = \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\begin{pmatrix}\frac{\partial H}{\partial q}\\ \frac{\partial H}{\partial p}\end{pmatrix} = J\nabla H(q,p).$$

Le système du pendule est donc hamiltonien, tout comme celui des petites oscillations du pendule, d'ailleurs. ◊

42. Voir définition C.1.4 un peu plus loin.
43. Siméon-Denis Poisson, 1781–1840.

Bibliographie

[1] V. Arnold. *Equations Différentielles Ordinaires*, volume 5ème Édition. Librairie Du Globe. Épuisé, mais il y a une version anglaise.

[2] G. Christol, A. Cot et C.-M. Marle, 1996. *Calcul Différentiel*. Ellipses.

[3] M. Crouzeix et A.-L. Mignot, 1983. *Analyse Numérique des Equations Différentielles*. Masson.

[4] S. Delabrière et M. Postel, 2004. *Méthodes d'approximation. Equations différentielles. Applications Scilab*. Ellipses.

[5] Leonhard Euler, 1787. *Institutiones calculi differentialis. in typographeo Petri Galeatii.* Service de la documentation Université de Strasbourg - Digital old books, http://num-scd-ulp.u-strasbg.fr:8080/.

[6] A. Fienga, H. Manche, J. Laskar et M. Gastineau, January 2008. INPOP06 : a new numerical planetary ephemeris. *Astronomy and Astrophysics*, 477 : 315–327. http://adsabs.harvard.edu/abs/2008A%26A...477..315F.

[7] E. Hairer, S. P. Norsett et G. Wanner, 1993. *Solving ordinary differential equations 1.* Springer.

[8] J. Laskar, P. Robutel, F. Joutel, M. Gastineau, A. C. M. Correia et B. Levrard, December 2004. A long-term numerical solution for the insolation quantities of the Earth. *Astronomy and Astrophysics*, 428 : 261–285. http://adsabs.harvard.edu/abs/2004A%26A...428..261L.

[9] I. Newton, 1740. *La méthode des fluxions, et les suites infinies, par M. le chevalier Newton*. Source Gallica.BnF.fr, Bibliothèque Nationale de France. Traduction en français par M. De Buffon.

Index

Cauchy-Lipschitz (théorème de –), 179, 185
conservation des aires, 23
consistance, 50
constante de Lipschitz, 177
contractante, 60
contrôle du pas de temps, 121
convergence, 46, 52

différentiation automatique, 90

équation autonome, 160
équation différentielle
    de Hamilton, 192
    de type gradient, 191
    homogène, 163
    linéaire, 163
    sans second membre, 163
équation différentielle linéaire, 162
espace des phases, 160
Euler (schéma d'–), 13
exponentielle de matrice, 164, 173

flot, 138
fonction
    de Liapounov, 189
    lipschitzienne, 177
    localement lipschitzienne, 185, 190
fonction de Liapounov, 190
fonction symplectique, 140

gradient, 191
Grönwall (lemme de –), 179

hamiltonien, 192

instabilité intrinsèque, 127

lemme de Grönwall, 179
    version discrète, 48
Liapounov (fonction de –), 190
Lipschitz (constante de –), 177

matrice symplectique, 138
méthode de Newton, 67
méthode de Picard, 171, 181

ordre, 53

pas adaptatif, 121
pas variable, 27
Picard (méthode de –), 171, 181
point d'équilibre, 161
point fixe, 60, 161
point stationnaire, 161
polygone d'Euler, 15

résolvante, 175
régularité de la fonction, 27

schéma d'Adams-Bashforth, 105
schéma d'Adams-Moulton, 106
schéma d'Euler explicite, 18
schéma d'Euler implicite, 19
schéma d'Euler modifié, 22
schéma d'Euler symplectique, 24, 141
schéma de Crank-Nicolson, 23
schéma de Runge-Kutta, 91
schéma de Stormer-Verlet, 142
schéma explicite, 25
schéma implicite, 25
schéma Leap-frog, 19
schéma numérique
    d'Euler, 13
schéma saute-mouton, 19
schéma symplectique, 23, 137
stabilité, 47
stabilité asymtotique, 161
stabilité asymptotique globale, 161
stabilité locale simple, 161

théorème
    de Cauchy-Lipschitz, 179
    de Cauchy-Lipschitz local, 185

variables séparées, 12
variation de la constante, 162, 166, 176

Wronskien, 175