Licence de Mathématiques - Université Pierre et Marie Curie
année 2018-2019

UE 3M236

Méthodes numériques pour les équations différentielles

Poly initialement de Marie Postel, remanié par Hervé Le Dret

Laboratoire Jacques-Louis Lions

17 janvier 2019

Table des matières

Introduction

Les équations différentielles sont des équations liant entre elles une fonction inconnue y d'une variable t et ses dérivées successives $y', y'', \ldots, y^{(n)}$, jusqu'à un certain ordre $n \geq 1$. On les présente le plus souvent sous la forme

$$y^{(n)}(t) = f(t, y(t), y'(t), \ldots, y^{(n-1)}(t)).$$

c'est-à-dire que la dérivée $n^{\text{ème}}$ de y au point t est fonction de t et des valeurs des dérivées de y d'ordre strictement inférieur à n en ce même point t. On se pose la question de savoir, pour f donnée, si de telles équations ont des solutions, c'est-à-dire s'il existe de telles fonctions y, puis de décrire aussi précisément que possible ces solutions éventuelles.

Les équations différentielles sont étudiées depuis l'invention du calcul différentiel par Newton [1] (1671). Des problèmes célèbres étaient à l'époque résolus soit par intégration directe, soit de manière approchée mais « à la main », le plus souvent par développement en série de la solution. Leibniz [2], Jean Bernoulli [3] et Huygens [4] (1691) résolvent ainsi plus ou moins en même temps le problème de la chaînette, c'est-à-dire le problème de déterminer la forme que prend une chaîne suspendue par ses extrémités et uniquement soumise à son poids. Jean Bernoulli pose celui de la brachistochrone en 1696, tout en en connaissant déjà la solution, comme un défi aux mathématiciens de son temps. La brachistochrone est la courbe joignant deux points telle qu'une bille roulant sans frottement sur cette courbe et lâchée à vitesse nulle du point le plus haut, atteint le point le plus bas en le moins de temps possible. Newton, Leibniz, de L'Hôpital [5] et Jacques Bernoulli [6] résolvent le problème — la courbe en question est un arc de cycloïde — et c'est l'occasion d'une brouille entre les deux frères Bernoulli, qui s'opposent sur la qualité de leur preuve respective ! Un siècle plus tard à peine, Euler [7] (1769) entreprend le recensement de toutes les équations différentielles qui peuvent être résolues de manière analytique, c'est-à-dire celles pour lesquelles on dispose de formules explicites donnant les solutions. Ces travaux occupent les tomes XXII et XXIII de ses ?uvres complètes.

La complexité des solutions de certaines équations différentielles d'apparence pourtant anodine laisse entrevoir qu'une alternative au calcul de la solution exacte est sans doute souhaitable. Mais le véritable essor de cette alternative, les méthodes numériques, a lieu au XIXème siècle quand Liouville [8] (1841) démontre que certaines équations ne peuvent pas être résolues analytiquement, c'est-à-dire que leur solution existe bien mais ne peut pas être exprimée comme une combinaison de fonctions élémentaires (c'est le cas par exemple de l'équation différentielle $y'(t) = t^2 + y(t)^2$). Ce résultat important, de nature analogue à l'impossibilité de résoudre par radicaux les équations polynomiales génériques de degré supérieur à 5, a également motivé des travaux théoriques sur l'existence et l'unicité des solutions des équations différentielles.

1. Sir Isaac Newton, 1642–1727.
2. Gottfried Wilhelm von Leibniz, 1646–1716.
3. Jean ou Johann Bernoulli, 1667–1748.
4. Christiaan Huygens, 1629–1695.
5. Guillaume François Antoine, marquis de L'Hôpital, 1661–1704.
6. Jacques ou Jakob Bernoulli, 1654–1705.
7. Leonhard Euler, 1707–1783.
8. Joseph Liouville, 1809–1882.

On se rend compte en fait que pour la plupart, en un certain sens, des équations différentielles, il est impossible d'obtenir des formules donnant leurs solutions. Les équations différentielles résolubles analytiquement sont l'exception plutôt que la règle. On peut néanmoins démontrer que les solutions existent sous des hypothèses assez générales. Et la seule façon d'obtenir des informations quantitatives sur ces solutions en l'absence de formules explicites est donc de les approcher numériquement. Il s'agit donc d'accepter philosophiquement que l'on ne peut en connaître quantitativement que des approximations, et de se donner les moyens de calculer effectivement de telles approximations tout en maîtrisant l'erreur commise autant que possible. Mais est-ce aussi dramatique que cela, alors que c'est déjà le cas ne serait-ce que pour l'immense majorité des nombres réels ? On est bien au c ?ur de l'analyse mathématique : approcher, majorer...

Depuis la méthode d'Euler (1768), une panoplie impressionnante de méthodes numériques a été mise au point pour faire face à des équations parfois très complexes. En effet parallèlement aux progrès mathématiques dans ce domaine, l'utilisation croissante des équations différentielles pour décrire des phénomènes physiques se développe également dans toutes les disciplines, de l'astronomie à la chimie, en passant par la médecine où la modélisation des épidémies avait déjà intéressé Daniel Bernoulli [9], le fils de Jean, en 1760. L'avènement des ordinateurs après la Seconde Guerre Mondiale a ensuite totalement bouleversé le paysage, de par la puissance de calcul mise à notre disposition, sans commune mesure avec celle, déjà prodigieuse, des grands calculateurs humains du passé, et qui permet de simuler numériquement avec précision des phénomènes régis par des équations différentielles d'une grande complexité.

La version 2016 du polycopié

Ce cours est une introduction aux aspects théoriques et numériques des équations différentielles au niveau licence et n'a donc en particulier pas l'ambition de présenter en détails les méthodes numériques les plus sophistiquées utilisées à l'heure actuelle. Il ne s'intéresse pas non plus aux aspects systèmes dynamiques des équations différentielles.

Il est partagé en deux grandes parties. La première partie donne une rapide introduction aux équations différentielles du point de vue théorique. Le cas particulier des équations différentielles linéaires déjà abordé en L1 est traité en premier, de façon probablement plus complète. Les principaux résultats d'existence et d'unicité des solutions dans le cas général sont ensuite présentés. Le texte contient un certain nombre de considérations parfois d'ordre historique, parfois peut-être un peu exagérément calculatoires, qui seront très certainement passées sous silence en amphi, car le temps y est assez sévèrement limité. Les passages correspondants sont en bleu, ou passeront au bleu au fur et à mesure qu'ils seront ignorés en live. Naturellement, ce n'est pas parce qu'un passage est en bleu qu'il est négligeable. Les passages bleus contiennent en fait de nombreuses choses intéressantes, il faut les lire malgré leur couleur.

La deuxième partie couvre les méthodes d'approximation numérique des solutions. Nous présentons d'abord les principes de base : discrétisation, consistance, stabilité, convergence. Les grandes catégories de méthodes sont ensuite décrites et quelques méthodes parmi celles utilisées le plus couramment sont détaillées. Les méthodes présentées sont également testées sur des exemples au moyen du logiciel `scilab` [10]. Des scripts `scilab` permettant de reproduire certains de ces exemples sont disponibles sur le site web

http://www.ljll.math.upmc.fr/~ledret/3M236/

9. Daniel Bernoulli, 1700–1782.

10. Le logiciel `scilab` est téléchargeable gratuitement sur le site web http://www.scilab.org pour OS X, diverses distributions Linux et Windows.

Signalons également que l'UPMC est partenaire de l'Université en Ligne qui propose une série de cours et d'exercices permettant au lecteur d'obtenir en temps libre les pré-requis pour suivre ce cours.

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

Enfin, ce polycopié a été initialement écrit par M. Postel, principalement à partir du livre de S. Delabrière et M. Postel [4], et d'un polycopié de S.-M. Kaber. Il a été ensuite considérablement remanié année après année par H. Le Dret. Les autres sources bibliographiques citées sont rassemblées en fin de poly.

La version 2019

Cette année le plan du cours a été remanié pour permettre une meilleure synchronisation avec les séances de travaux pratiques en python. Il est maintenant divisé en quatre parties

1. EDO etude théorique
   (a) Ce qu'il faut savoir avant de commencer
   (b) Présentation générale
   (c) Cas linéaire
   (d) Existence globale et unicité
2. Introduction des schémas numériques
   (a) Définition d'un schéma numérique explicite à un pas
   (b) Consistance
   (c) Stabilité
   (d) Convergence
3. EDO etude théorique : solutions locales
4. Etude de quelques schémas numériques

Le contenu est quasiment le même que dans la version de 2016, mais l'ordre est différent. Certaines figures ont été refaites de manière à donner des exemples de simulation numériques en python (voir la section cours dans l'UE moodle)

Le poly correspondant à cette réorganisation sera mis en ligne progressivement au même endroit sur l'UE moodle

Chapitre 1

Équations différentielles ordinaires, étude théorique

1.1 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. Inutile de rappeler ce qu'il advient des édifices construits sur du sable. Les fondations mathématiques sont établies tant bien que mal (par manque de temps) dans les premières années de licence. Tout ce qui a été vu précédemment est donc considéré comme acquis, puisque l'on ne saurait croire à cette légende urbaine de l'existence de gens qui oublient tout une fois l'examen passé. Ce qui suit est une liste plus spécifique de notions antérieures qui seront directement utilisées dans ce cours. Si par malheur, on ne se sent pas complètement au point dessus, il est toujours temps de revoir par exemple ce qui est enseigné dans l'excellent cours de Topologie et calcul différentiel donné au premier semestre de L_3 à l'UPMC.

1.1.1 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, mais l'intégrale de Lebesgue ne peut pas faire de mal non plus. Une intégrale fonction de sa borne supérieure donne une primitive de l'intégrande (par exemple si celle-ci est continue).

1.1.2 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 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 $\bar{I}$.

1.1.3 Calcul différentiel

On se doute bien que pour parler d'équations différentielles, 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 $(\nabla f(x_0))_{ij} = \frac{\partial f_i}{\partial x_j}(x_0)$, $i = 1, \dots, m$,

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

$j = 1, \dots, k$. 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 $j = 1, \dots, k$ et $l = 1, \dots, n$ 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).

1.1.4 Culture générale

Pour avoir une idée de ce à quoi peut bien servir tout ce dont nous allons parler en dehors des mathématiques elles-mêmes, il n'est pas inutile de posséder quelques restes (aussi beaux que possible) de mécanique du point matériel, de physique, de chimie, etc.

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

1.2 Présentation générale

1.2.1 Définitions et premiers exemples

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

$$\forall t \in I, \quad y'(t) = f\big(t, y(t)\big). \tag{1.2.1}$$

On rappelle qu'une fonction $y : \bar{I} \to \mathbb{R}^m$ 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. [3]

On dit que le problème (1.2.1) est un *système d'équations différentielles ordinaires* ou *système d'EDO* ou *EDO* [4] tout court, du premier ordre, car seule la dérivée première de 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 (1.2.1) est pour chaque t une égalité entre deux vecteurs de $\mathbb{R}^m$. On peut l'écrire composante par composante sous la forme

$$y_i'(t) = f_i\big(t, y_1(t), y_2(t), \ldots, y_m(t)\big), \tag{1.2.2}$$

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

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

3. Alors prononcées « y point » et « y deux points » ou « y point point ».

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

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 1.1 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.

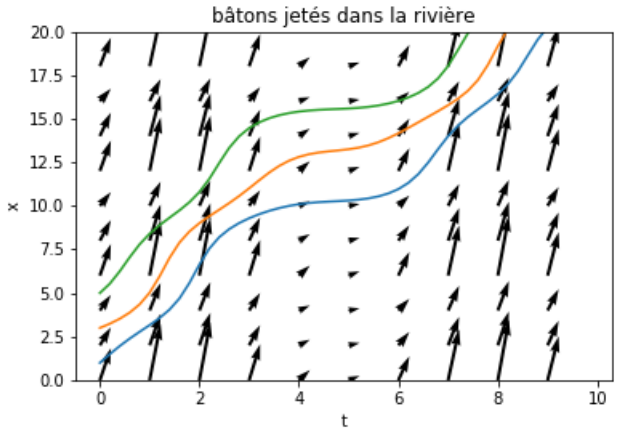


FIGURE 1.1 – Champ de vecteurs du courant dans une rivière $f(x,t) = (2 + 0.9\cos(x))(1 + 0.9\sin(t))$ et positions de trois bâtons lancés à $x_0 = 1$, $x_0 = 3$ et $x_0 = 5$ à l'instant initial (voir le notebook cours_1.ipynb).

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

1.2.2 La dimension 1

Dans le cas de la dimension un ($m = 1$), il est fréquent de renommer $t = x$ et de considérer l'équation différentielle écrite sous la forme

$$y'(x) = f(x, y(x)),\ (x,y) \in U \subset \mathbb{R} \times \mathbb{R}. \tag{1.2.3}$$

Définition 1.2.1 *À tout point $M = (x_0, y_0)$, on associe la droite D_M passant par M et de coefficient directeur $f(x_0, y_0)$*

$$D_M : y = y_0 + f(x_0, y_0)(x - x_0).$$

L'application $M \longrightarrow D_M$ est appelée champ des tangentes *associé à l'équation* (1.2.3)*. Une courbe intégrale de* (1.2.3) *est une courbe différentiable C qui a pour tangente à chaque point $M \in C$ la droite D_M du champ des tangentes passant par ce point.*

Définition 1.2.2 *Les lignes isoclines sont les courbes*

$$\Gamma_p = \{(x,y), f(x,y) = p\}$$

correspondant à l'ensemble des points M où la droite D_M a une pente donnée p.

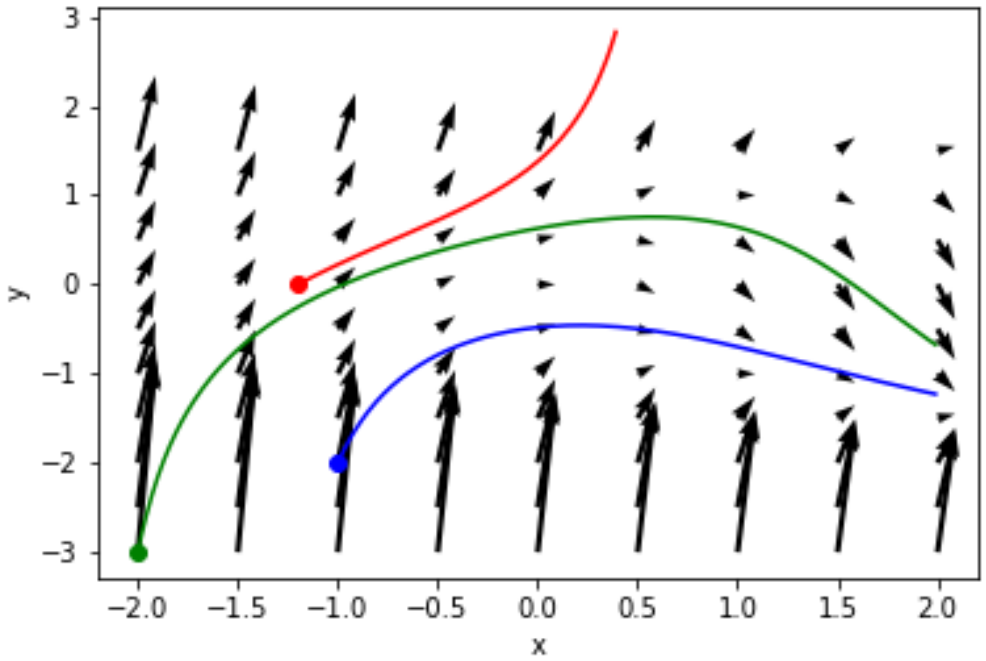


FIGURE 1.2 – Le champ des tangentes pour l'équation $y' = y^2 - x$, avec trois courbes intégrales (voir le notebook cours_1.ipynb).

La courbe Γ_0 a la propriété de partager le domaine U en trois régions

$$U = U^+ \cup U^- \cup \Gamma_0 \text{ avec } U^+ = \{M \in U, f(M) > 0\}, \quad U^- = \{M \in U, f(M) < 0\}.$$

En étudiant la position des solutions par rapport aux isoclines et à d'autres courbes caractéristiques on peut décrire le comportement des solutions très en détail.

Les EDO apparaissent dans bien d'autres domaines que la mécanique. En voici un premier exemple.

Exemple 1.2.1 La dynamique des populations s'attache à modéliser l'évolution temporelle de la taille de populations diverses (bactériennes, animales, humaines...) afin de la prédire. Le modèle le plus simple consiste à supposer que le taux de variation de cette taille est proportionnel à la population existante, c'est-à-dire que le bilan des naissances et des décès se fait à taux λ constant. C'est le cas si les taux de naissance et de décès sont eux-mêmes tous les deux fixes par exemple. Autrement dit, si $y(t)$ désigne le nombre d'individus au temps t, le nombre de naissances diminué du nombre des décès entre t et $t + \Delta t$ est approximativement proportionnel à $y(t)$ et à Δt, pour Δt petit, d'où une variation de la population $y(t + \Delta t) - y(t) \approx \lambda y(t)\Delta t$.

Divisant par Δt, puis passant à la limite quand Δt tend vers 0, on obtient que les variations de $y(t)$ sont régies par l'équation différentielle

$$\frac{dy}{dt}(t) = \lambda y(t), \tag{1.2.4}$$

dont on connaît bien les solutions $y(t) = Ce^{\lambda t}$, où C est une constante arbitraire. Si l'on connaît la population à l'instant 0, $y(0)$ supposée strictement positive, alors la solution est déterminée de façon

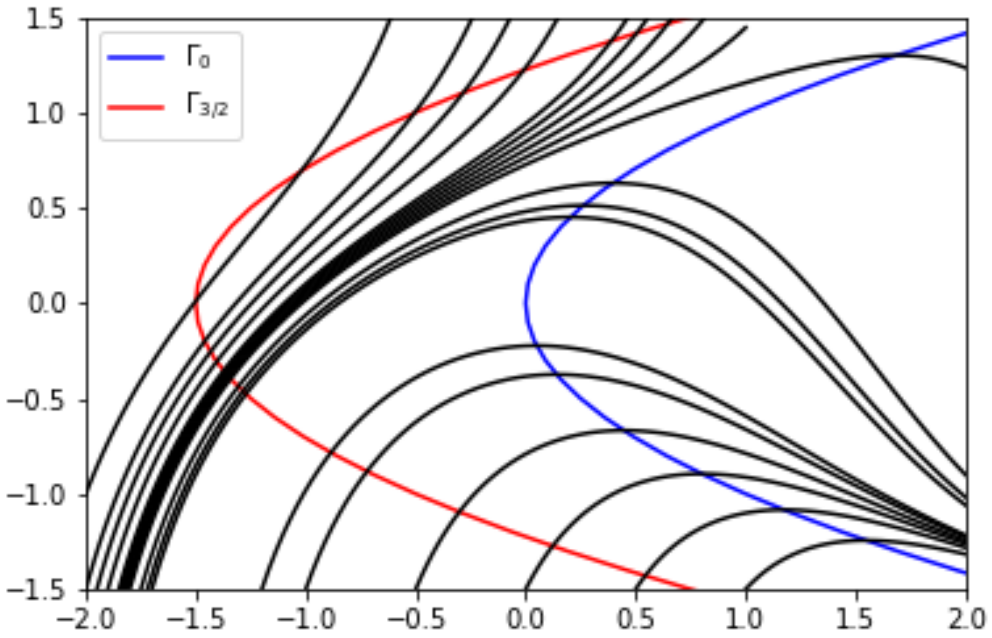


FIGURE 1.3 – Les courbes isoclines Γ_0 et $\Gamma_{3/2}$ pour l'équation $y' = y^2 - x$, avec les courbes intégrales. Les courbes intégrales ont une tangente horizontale à l'endroit où elle coupent Γ_0 (voir le notebook cours_1.ipynb).

unique $y(t) = y(0)e^{\lambda t}$. Il s'agit du fameux *modèle malthusien* [5] (1798). Il conduit à une croissance exponentielle si $\lambda > 0$, c'est-à-dire si les naissances l'emportent sur les décès et à une décroissance exponentielle dans le cas inverse où $\lambda < 0$. [6] Enfin, si $\lambda = 0$, on a affaire à une population constante $y(t) = y(0)$ pour tout t.

L'équation (1.2.4) peut être mise sous la forme générique (1.2.1) avec $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $f(t, y) = \lambda y$ (nul besoin d'un intervalle $\bar{I}$ ici), donc $m = 1$, c'est une EDO scalaire. On note que f ne dépend en fait pas de t.

La modèle malthusien est un peu simpliste, d'ailleurs la croissance exponentielle qu'il prédit dans le cas $\lambda > 0$ ne peut manifestement pas se produire pour tout temps dans une population réelle. Afin d'intégrer certains freins à la croissance infinie de tels systèmes, comme les limitations environnementales, celles des ressources naturelles qui induisent une compétition interne, on peut utiliser le *modèle logistique* suivant, aussi appelé *modèle de Verhulst-Pearl* [7].

$$\frac{dy}{dt}(t) = \lambda y(t)\Big(1 - \frac{y(t)}{K}\Big). \tag{1.2.5}$$

Le paramètre λ a la même interprétation que précédemment quand y est très petit, alors que le paramètre K représente la capacité et la taille limite du système. On notera que si $y(t)$ est proche de 0 en comparaison à la valeur K, l'équation (1.2.5) est très semblable à l'équation (1.2.4). On s'attend donc à une croissance approximativement exponentielle dans ce cas, si $\lambda > 0$, au moins dans un premier temps. Par contre, au fur et à mesure que la population $y(t)$ augmente, le facteur $\big(1 - \frac{y(t)}{K}\big)$

5. Thomas Robert Malthus, 1766–1834.
6. Dans l'affaire, on a ignoré le fait qu'une population réelle se compose d'un nombre entier d'individus.
7. Pierre-François Verhulst, 1804–1849 ; Raymond Pearl, 1879–1940.

devient sensiblement inférieur à 1, ce qui correspond à une diminution du taux de naissance effectif (ou une augmentation du taux de décès, ou les deux...), due aux contraintes environnementales. On s'attend donc à ce que la population augmente de moins en moins vite. C'est d'ailleurs bien ce qui se passe quand on résout cette EDO, et l'on s'aperçoit que la population tend vers la valeur K quand $t \to +\infty$.

Dans le cas du modèle logistique, la fonction f du second membre de l'EDO prend la forme $f(t,y) = \lambda y\big(1 - \frac{y}{K}\big)$, toujours indépendante de t et toujours avec $m = 1$. ◊

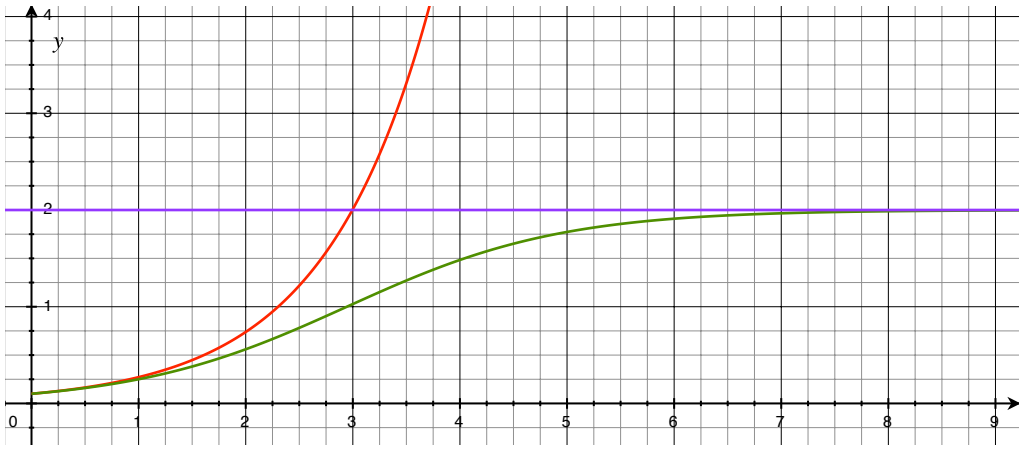


FIGURE 1.4 – Comparaison modèle malthusien-modèle logistique avec la même population initiale $y(0) = \frac{1}{10}$ et le même taux de naissance $\lambda = 1$. La population limite du modèle logistique vaut ici $K = 2$.

Exemple 1.2.2 Les variations de température à la surface d'un corps sont (en première approximation) proportionnelles à sa température relative, c'est-à-dire à l'écart entre sa propre température, et celle de l'environnement supposée constante. Plus formellement, si $y(t)$ désigne la température de ce corps au temps t, nous avons

$$\frac{dy}{dt}(t) = -\lambda(y(t) - \theta) \tag{1.2.6}$$

où θ désigne la température de l'environnement et $\lambda > 0$ une constante physique quantifiant l'efficacité du transfert de chaleur entre le corps et son environnement. Il faut prendre λ positif, car cela permet d'assurer que si le corps est plus chaud que l'environnement, alors il va y diffuser de la chaleur et se refroidir (la dérivée de la température est en effet négative dans ce cas), alors que s'il est plus froid que l'environnement, il va se réchauffer en absorbant de la chaleur venant de l'environnement. Une constante λ négative prédirait le comportement inverse, puissamment non physique.

Ici aussi la solution est bien connue et l'on a $y(t) = (y(0) - \theta)e^{-\lambda t} + \theta$ pour tout temps $t \geq 0$. Le comportement est bien celui physiquement attendu : la température du corps tend exponentiellement vite vers celle de l'environnement, le corps se refroidit s'il était initialement plus chaud que l'environnement et se réchauffe dans le cas inverse.

Dans cet exemple également, la fonction second membre ne dépend pas de t :

$$f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \quad f(t,y) = -\lambda(y - \theta).$$

Il s'agit encore d'un exemple scalaire, $m = 1$. ◊

Remarque 1.2.1 Il faut bien distinguer les EDO des équations aux dérivées partielles (EDP). Une EDP est une relation qui relie entre elles certaines des dérivées partielles d'une fonction inconnue de plusieurs variables. Par exemple, l'équation de la chaleur dans une barre mince homogène isolée de son environnement, sauf éventuellement en ses deux extrémités — une situation qui n'a rien à voir avec la précédente — s'écrit

$$\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0,$$

où $u(t,x)$ est la température au temps t et au point d'abscisse x, donc fonction inconnue des deux variables t et x, et α est une constante positive caractérisant le taux de diffusion de la chaleur dans la barre. Cette équation de la chaleur est une EDP en dimension 1 d'espace, *i.e.*, en la variable x. On a bien sûr des généralisations en dimension d quelconque d'espace.

Les EDP généralisent donc les EDO, lesquelles ne portent que sur des fonctions inconnues d'une seule variable. Les EDP et leur approximation numérique se situent également à un niveau de difficulté largement plus élevé et nous n'en parlerons plus dans ces notes. D'une certaine façon, une EDO est une EDP en dimension d'espace 0, il n'y a pas de variable x.

1.2.3 Systèmes différentiels d'ordre 1

Donnons maintenant quelques définitions de façon un peu plus formelle.

Définition 1.2.3 *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 $n^{\text{ème}}$, $y^{(n)}$, 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*

$$\forall t \in I, \quad y^{(n)}(t) = f(t, y(t), y'(t), \ldots, y^{(n-1)}(t)). \tag{1.2.7}$$

On appelle nombre de degrés de liberté *le produit mn.*

Une solution *du système* (1.2.7) *est une fonction de $\bar{I}$ à valeurs dans $\mathbb{R}^m$, n fois dérivable sur I et telle que l'égalité* (1.2.7) *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 (1.2.7) 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 (1.2.1) 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 1.2.3 n'est qu'apparente.

Proposition 1.2.4 *Soit y est une solution de* (1.2.7). *La fonction vectorielle $Y : \bar{I} \longrightarrow (\mathbb{R}^m)^n$ définie par*

$$Y(t) = (y(t), y'(t), \ldots, y^{n-1}(t))$$

est alors solution du système d'équations différentielles du premier ordre

$$\forall t \in I, \quad \frac{dY(t)}{dt} = F(t, Y(t)), \tag{1.2.8}$$

où F est la fonction définie par

$$\begin{aligned} F\colon \bar{I}\times(\mathbb{R}^m)^n &\longrightarrow (\mathbb{R}^m)^n\\ (t,Y_1,Y_2,\dots,Y_n) &\longmapsto \big(Y_2,Y_3,\dots,Y_n,f(t,Y_1,Y_2,\dots,Y_n)\big), \end{aligned}$$

les Y_i désignant des éléments génériques de $\mathbb{R}^m$.

Réciproquement, toute solution Y du système (1.2.8) *donne naissance à une solution de* (1.2.7) *en posant $y(t)=Y_1(t)$.*

Démonstration. Soit y est une solution de (1.2.7). 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 (1.2.7). On voit donc que Y est une solution du système (1.2.8).

Réciproquement, donnons-nous une solution Y de (1.2.8) 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 (1.2.8) 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 (1.2.8) pour $i=n$ et d'après la récurrence qui précède. La fonction y est par conséquent une solution de (1.2.7). ◊

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.

Exemple 1.2.3 L'équation suivante modélise la déformation u d'un fil élastique occupant l'intervalle $]a,b[$ sous l'action d'une force g qui lui est perpendiculaire

$$\forall x\in\,]a,b[,\quad -u''(x)+cu(x)=g(x), \tag{1.2.9}$$

où c est une constante donnée, sans grande signification mécanique d'ailleurs. On écrit l'équation sous cette forme car c'est celle-ci qui apparaît naturellement dans ce contexte. Par ailleurs, la variable est ici x et non t car c'est une variable d'espace dans ce modèle (ceci est très exceptionnel dans ce cours). En la réécrivant sous la forme $u''(x)=cu(x)-g(x)$, on voit qu'il s'agit d'une équation scalaire, $m=1$, et du second ordre, $n=2$, dont la fonction second membre $f\colon [a,b]\times\ \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ est donnée par

$$f(x,u_1,u_2)=cu_1-g(x).$$

Cette fonction second membre ne dépend pas de u_2 car u' n'apparaît pas dans l'équation.

Ramenons cette équation du second ordre à une équation du premier ordre. Conformément à ce qui précède, on pose d'abord

$$Y(x)=\begin{pmatrix}Y_1(x)\\ Y_2(x)\end{pmatrix}=\begin{pmatrix}u(x)\\ u'(x)\end{pmatrix},$$

en posant les calculs en colonne, plutôt qu'en ligne comme dans la proposition 1.2.4. La disposition en colonne est en effet plus naturelle pour des vecteurs mais elle prend plus de place sur une page écrite que la disposition en ligne... Dans notre contexte, toutes ces notations sont interchangeables. On écrit ensuite que la deuxième composante est la dérivée de la première $Y_2(x) = u'(x) = Y_1'(x)$, puis que la dérivée de la deuxième composante est donnée par l'EDO de départ $Y_2'(x) = u''(x) = cu(x) - g(x) = cY_1(x) - g(x)$. On a donc ainsi obtenu le système d'ordre un à deux degrés de liberté équivalent

$$\begin{cases} Y_1'(x) = Y_2(x), \\ Y_2'(x) = cY_1(x) - g(x). \end{cases}$$

Pour l'écrire sous la forme (1.2.1) ou (1.2.8), il faut lire dans l'expression précédente la fonction second membre $F\colon [a,b] \times \mathbb{R}^2 \longrightarrow \mathbb{R}^2$. Cela se fait naturellement en supprimant la dépendance en x (et donc en t dans tout le reste du cours) dans les variables Y_i, de façon à retomber sur nos pieds quand on y substituera la fonction $x \mapsto Y(x)$, [8] soit

$$\forall (x, Y) \in [a,b] \times \mathbb{R}^2, \quad F(x,Y) = \begin{pmatrix} Y_2 \\ cY_1 - g(x) \end{pmatrix}.$$

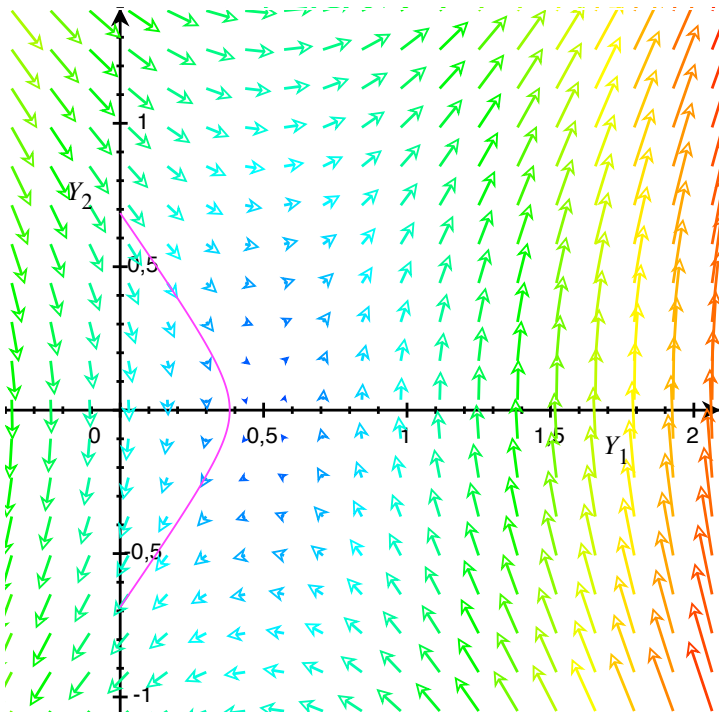


FIGURE 1.5 – Le champ de vecteurs F du fil, $c = 2$, $g(x) = 1$. Attention la solution particulière dessinée en violet n'a pas la forme du fil déformé, c'est une courbe $x \mapsto Y(x) = (u(x), u'(x))^T$, avec u une solution de l'équation (1.2.9), voir la Figure 1.6 pour la forme du fil.

Une erreur par trop courante : dire que F est définie par

$$F(x, Y(x)) = \begin{pmatrix} Y_2(x) \\ cY_1(x) - g(x) \end{pmatrix}.$$

8. On a déjà effectué cette opération à plusieurs reprises sans trop insister dessus, ne serait-ce que quelques lignes plus haut, pour donner la fonction second membre f du même exemple.

Ceci n'a aucun sens. On définit ainsi tout au plus une fonction de $[a,b] \to \mathbb{R}^2$, *i.e.,* une courbe paramétrée qui n'a donc rien à voir avec une fonction second membre $[a,b] \times \mathbb{R}^2 \to \mathbb{R}^2$ laquelle est un champ de vecteurs, voir Figure 1.5, et d'ailleurs en fait on ne définit rien du tout puisque qu'on ne sait même pas à ce stade que $x \mapsto Y(x)$ existe. À éviter... ◊

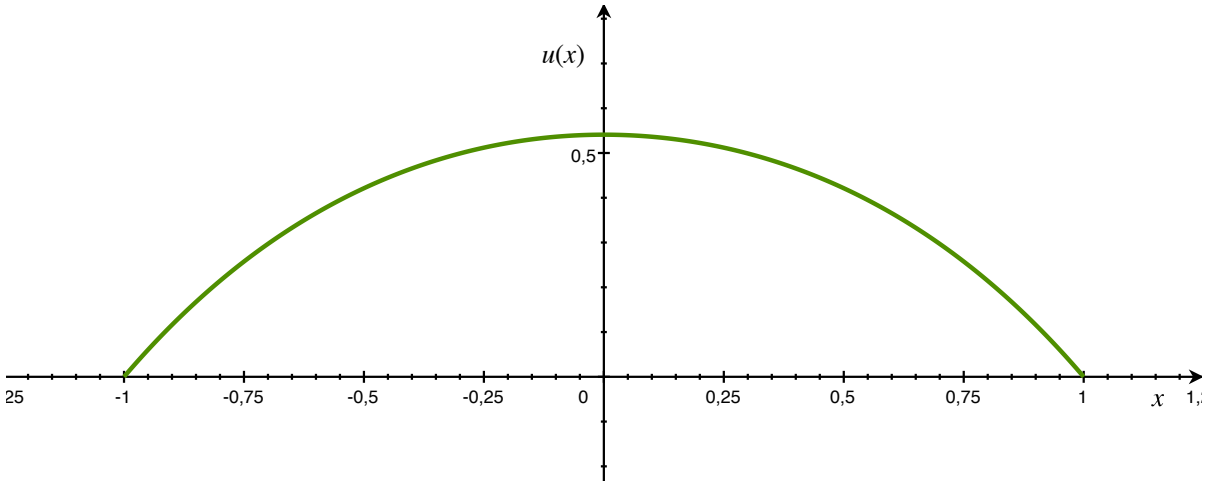


FIGURE 1.6 – Voici la forme du fil $x \mapsto u(x)$ pour les mêmes données. C'est un bon exercice de faire le lien entre cette description du fil dans l'espace physique et la description plus abstraite du même fil dans la Figure 1.5 dans ce que l'on appelle l'espace de phase.

Notons en passant que la notation usuelle $y'(t) = f(t, y(t))$ conduit à appeler *temps* la variable t, en référence aux problèmes physiques d'évolution issus de la mécanique d'où proviennent de nombreux systèmes différentiels. Mais ce n'est pas toujours le cas, comme dans l'exemple précédent, où la solution dépend de l'abscisse x le long du fil, une coordonnée spatiale, ce qui sera néanmoins très exceptionnel dans ces notes où la variable t dominera de façon écrasante.

Exemple 1.2.4 Quelle peut être la dimension des problèmes différentiels rencontrés dans la pratique ? Regardons par exemple en astronomie, le problème de calculer les trajectoires des corps célestes principaux du système solaire, pour tenter de prédire son évolution future par exemple ou d'éventuels événements catastrophiques sur la Terre... Les astronomes doivent prendre en compte, en plus du soleil et des huit planètes principales et leurs satellites, les planètes naines et les astéroïdes, ce qui conduit à un problème à N corps avec $N \approx 300$. Et encore, on néglige ici les effets gravitationnels des autres corps de taille plus petite, comètes, météorites, déchets spatiaux, poussières... dont le nombre est à proprement parler astronomique.

Plaçons-nous dans le cadre newtonien classique, c'est-à-dire en négligeant les effets de la relativité générale. À l'échelle du système solaire et en première approximation, on peut assimiler les corps célestes à des points (disons au moins tant qu'ils ne menacent pas d'entrer en collision [9]). Les équations régissant les positions à chaque instant $q_i(t) \in \mathbb{R}^3$, $i = 1, \dots, N$, de N corps célestes de masses m_i, interagissant gravitationnellement sont [10]

$$m_i \ddot{q}_i(t) = G \sum_{k \neq i}^{N} \frac{m_i m_k (q_k(t) - q_i(t))}{|q_k(t) - q_i(t)|^3}, i = 1, \dots, N,$$

9. Cela n'est pas tout à fait exact. Par exemple, les effets de marée ont une influence très importante sur le comportement en temps long du système Terre-Lune.

10. On peut bien sûr simplifier chaque ligne i par m_i.

où G est la constante gravitationnelle, qui vaut approximativement $6{,}63 \times 10^{-11}\,\mathrm{N\,m^2\,kg^{-2}}$, et la notation $|q|$ désigne la norme euclidienne dans $\mathbb{R}^3$ (mesurée en mètres). Elles sont obtenues en appliquant la loi de Newton (de la gravitation universelle) à loi de Newton (fondamentale de la dynamique) reliant l'accélération d'un point matériel à la résultante des forces qui lui sont appliquées.

En se ramenant à des équations du premier ordre par la méthode exposée ci-dessus, on obtient donc un système de $m = 6N \approx 1800$ équations différentielles scalaires couplées à résoudre. On dit qu'elles sont couplées car toutes les composantes sont mélangées entre elles dans le second membre.

Écrivons en effet $q\colon \bar{I} \to \mathbb{R}^{3N}$ en empilant les unes sur les autres les coordonnées dans une base fixée une fois pour toutes, de chacun des corps célestes, [11]

$$q(t) = \begin{pmatrix} q_1(t) \\ q_2(t) \\ \vdots \\ q_N(t) \end{pmatrix} \in \mathbb{R}^{3N} \text{ avec } q_i(t) = \begin{pmatrix} q_{i,1}(t) \\ q_{i,2}(t) \\ q_{i,3}(t) \end{pmatrix} \in \mathbb{R}^3,\ i = 1, \dots, N.$$

Le système du premier ordre équivalent va donc s'écrire avec une fonction inconnue $Q\colon \bar{I} \to \mathbb{R}^{6N}$ sous la forme $\dot{Q}(t) = F(Q(t))$. Pour décrire la fonction F, écrivons Q comme une superposition en une colonne de $2N$ blocs 3×1 correspondant pour les N premiers blocs aux positions et pour les N blocs suivants aux vitesses de chaque corps céleste, $Q_i(t) = q_i(t)$ et $Q_{i+N}(t) = \dot{q}_i(t)$ pour $i = 1, \dots, N$. La fonction second membre F, ici aussi lue en supprimant la dépendance en t dans les expressions faisant intervenir $Q_i(t)$, est alors donnée par

$$\begin{cases} F_i(Q) = Q_{i+N}, \\ F_{i+N}(Q) = G \displaystyle\sum_{k \neq i}^{N} \frac{m_k(Q_k - Q_i)}{|Q_k - Q_i|^3}, \end{cases} \qquad i = 1, \dots, N,$$

encore une fois par blocs 3×1. [12] Le seul cas général dans lequel on sache écrire explicitement les solutions de ce système est celui où $N = 2$, c'est-à-dire le problème à 2 corps. Dans ce cas, on sait depuis longtemps que les trajectoires sont planes et sont en fait des coniques : ellipses, paraboles, hyperboles, dont le centre de gravité du système est l'un des foyers et qui sont parcourues à des vitesses parfaitement connues également sous le nom de lois de Kepler [13]. Dès que $N \geq 3$, ce type de description explicite n'existe plus, sauf dans des cas très particuliers. Comment faire alors pour calculer les trajectoires, voir Figure **??** ?

Par ailleurs, si l'on souhaite planifier la trajectoire d'un engin spatial par exemple, afin de l'envoyer sur une cible mouvante, minuscule dans l'immensité de l'espace, comme une comète après quelques années de vol, les calculs doivent être d'une précision diabolique. L'état de l'art en 2008 pour certains calculs astronomiques consistait à utiliser des schémas d'Adams d'ordre (très) élevé (disons 12), voir [6]. [14] Les choses ont certainement beaucoup progressé depuis. ◊

1.2.4 Le cas des équations différentielles autonomes

Définition 1.2.5 *On appelle* équation différentielle autonome *une EDO dont le second membre ne dépend pas explicitement du temps.*

$$y^{(n)}(t) = f(y(t), y'(t), \dots, y^{(n-1)}(t)). \tag{1.2.10}$$

11. Ici aussi, on identifie donc $(\mathbb{R}^3)^N$ avec $\mathbb{R}^{3N}$ sans autre forme de procès.

12. Remarquons que cette fonction n'est définie que si $Q_i \neq Q_k$ pour tous $1 \leq i, k \leq N$, $i \neq k$. Il y a clairement des ennuis quand deux planètes entrent en collision...

13. Johannes Kepler, 1571–1630.

14. Pour se faire une idée de la complexité de ces calculs, les schémas d'Adams d'ordre 2 et 3 sont présentés plus loin dans ce cours...

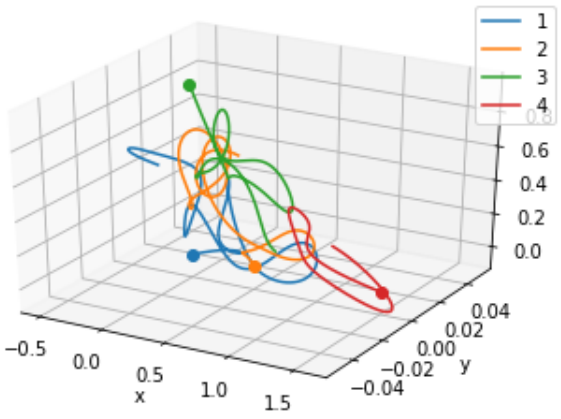


FIGURE 1.7 – Un exemple de trajectoires à quatre corps, manifestation tridimensionnelle d'une ODE dont le champ de vecteurs est de dimension 24, calculées approximativement et sans précautions particulières avec python (voir le notebook cours_1.ipynb).

Définition 1.2.6 *On appelle* espace des phases *l'espace* $\mathbb{R}^{mn}$ *où se trouvent les trajectoires des solutions de l'équation différentielle* (1.2.10) *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 [15] [1].

Définition 1.2.7 *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 1.2.8 *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*

$$\|y(0) - y_e\| < r \Rightarrow \lim_{t\to\infty} \|y(t) - y_e\| = 0.$$

15. Vladimir Igorevich Arnold, 1937–2010.

— *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 $y(t) = \big(y_1(t), y_2(t), \dots, y_m(t)\big)$ 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.

Exemple 1.2.5 Un exemple simple d'EDO est la *seconde loi de Newton* [16] qui relie l'accélération d'une particule de masse m placée dans un champ de forces F, avec sa position $q(t) \in \mathbb{R}^3$

$$m\frac{d^2q}{dt^2}(t) = F(t, q(t)).$$

On néglige ici les frottements qui ne s'écrivent pas sous cette forme. Supposons que le champ de forces soit le poids $-mge_3$, où $-ge_3$ est l'accélération de la pesanteur supposée constante ($g \approx 9{,}81 ms^{-2}$ à notre latitude) et e_3 est le troisième vecteur supposé vertical et pointant vers le haut d'une base orthonormée. En composantes dans cette base, l'équation se ramène à

$$\begin{cases} \ddot{q}_1(t) = 0, \\ \ddot{q}_2(t) = 0, \\ \ddot{q}_3(t) = -g. \end{cases}$$

Le système est découplé et se résout à la main. Les trajectoires sont des paraboles. [17] Concentrons nous sur la troisième équation donnant l'altitude du point

$$\ddot{q}_3(t) = -g.$$

Suivons le principe de la proposition 1.2.4 et notons $y_1 = q_3$ et $y_2 = \dot{q}_3$ la vitesse verticale. Il vient

$$\dot{y}_1(t) = y_2(t) \text{ et } \dot{y}_2(t) = -g. \tag{1.2.11}$$

L'espace des phases est le plan (y_1, y_2). On a représenté sur la Figure 1.8 le champ de vecteurs de vitesse de phase, qui a pour composantes $(y_2, -g)$ au point (y_1, y_2). [18] ◇

16. Encore appelée relation fondamentale de la dynamique : soit un corps de masse m (constante), l'accélération subie par ce corps dans un référentiel galiléen est proportionnelle à la résultante des forces qu'il subit, et inversement proportionnelle à sa masse m.

17. Par exemple, partant du point $(0, 0, 0)$ à vitesse initiale nulle, on obtient $q_1(t) = q_2(t) = 0$, $q_3(t) = -\frac{g}{2}t^2$. La chute est verticale.

18. Attention, l'altitude q_3 se retrouve représentée horizontalement dans l'espace des phases par y_1. La coordonnée verticale de l'espace des phases est la vitesse $\dot{q}_3$. Il faut tourner la tête vers la droite pour regarder le dessin dans l'orientation physique.

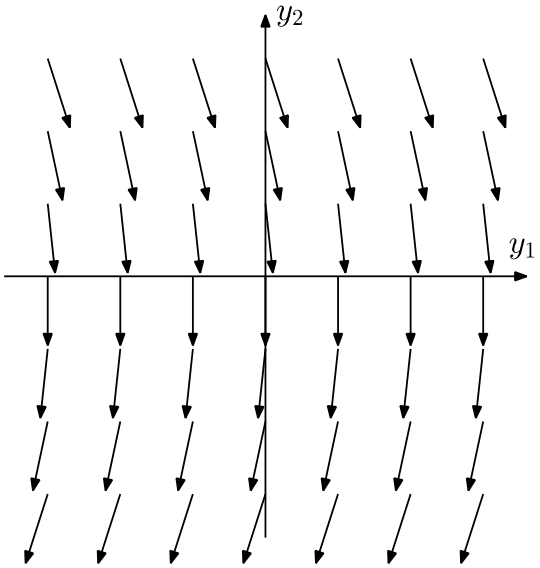


FIGURE 1.8 – L'espace des phases pour la chute verticale. Les flèches représentent le champ de vitesses de phase (voir le notebook cours_1.ipynb).

En dimension $m = 2$, le théorème de Poincaré-Bendixson [19] ([7], page 113) permet de classer les solutions bornées d'un système différentiel autonome en trois catégories. Les trajectoires peuvent

— tendre vers un point d'équilibre
— être périodique
— tendre vers un cycle limite (c'est-à-dire vers une trajectoire périodique).

Donnons un exemple de comportement périodique.

Exemple 1.2.6 Les oscillations d'un pendule dans un repère inertiel sont planes (voir Figure 1.9). Elles peuvent être décrites dans un espace des phases en dimension deux de coordonnées (y_1, y_2), où y_1 est l'angle de déviation par rapport à la verticale et y_2 la vitesse angulaire, c'est-à-dire la dérivée de l'angle par rapport au temps (voir Figure 1.10).

Toujours en raison de la loi de Newton, l'accélération est approximativement proportionnelle à l'angle de déviation quand celui-ci est petit. On montre que pour les petites oscillations du pendule

$$\ddot{y}_1(t) = -ky_1(t) \quad \text{avec } k = g/l$$

où l est la longueur du pendule et g l'accélération de la pesanteur. On transforme cette équation du second ordre en un système différentiel du premier ordre

$$\dot{y}_1(t) = y_2(t), \quad \dot{y}_2(t) = -ky_1(t),$$

On appelle *point singulier* ou *point stationnaire* dans l'espace des phases un point (y_1, y_2) où la vitesse de phase est nulle, c'est-à-dire un point d'équilibre du système du premier ordre équivalent. Dans cet exemple $(0, 0)$ est un point singulier.

Si l'on ne souhaite pas se limiter aux petites oscillations, le système modélisant plus exactement les oscillations du pendule est

$$\dot{y}_1(t) = y_2(t), \quad \dot{y}_2(t) = -k \sin y_1(t), \tag{1.2.12}$$

19. Jules Henri Poincaré, 1854–1912 ; Ivar Otto Bendixson, 1861–1935.

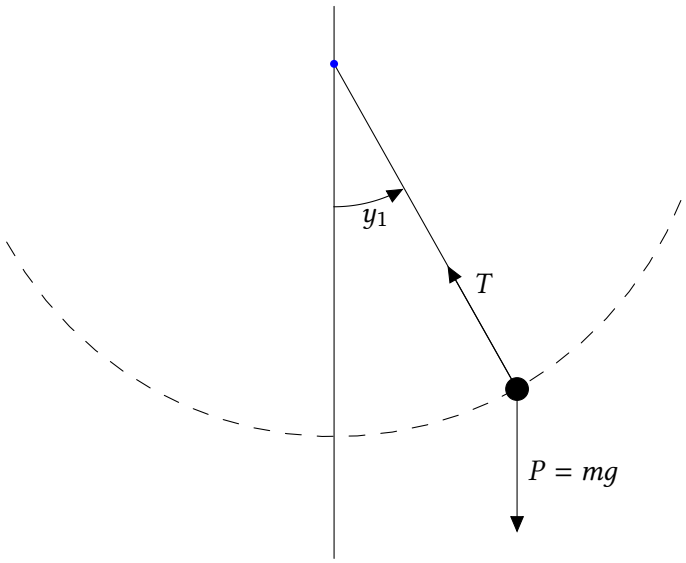


FIGURE 1.9 – Le pendule. P est le poids, T la tension du fil.

pour lequel le champ de vecteurs de vitesse de phase dans l'espace des phases est représenté sur la Figure 1.11. On notera qu'on a maintenant $(y_1, y_2) = (n\pi, 0)$ avec $n \in \mathbb{Z}$ comme points singuliers. ◊

Exemple 1.2.7 L'équation de van der Pol [20] est un exemple du troisième comportement. Elle a été initialement proposée pour décrire des oscillations spontanées dans des circuits électriques, puis a été utilisée en biologie où elle a été généralisée pour décrire le fonctionnement des neurones.

On écrit ici cette équation sous la forme générale

$$\begin{cases} x'(t) = y(t), \\ y'(t) = \varepsilon(1 - x(t)^2)y(t) - x(t), \end{cases} \tag{1.2.13}$$

avec $\varepsilon > 0$. Il y a un point critique à l'origine et par ailleurs $y'(t)$ change de signe sur la courbe

$$y = \frac{x}{\varepsilon(1 - x^2)}.$$

La Figure 1.12 montre cette courbe ainsi que deux trajectoires partant l'une de $(-2, 3)$, l'autre de $(-0.1, -0.1)$ et tendant vers un cycle limite. ◊

Exemple 1.2.8 Un exemple de système d'ordre deux dont les solutions tendent vers un point critique est

$$\begin{cases} y_1'(t) = y_2(t) - y_1^3(t), \\ y_2'(t) = y_1(t) - y_2^3(t), \\ y_1(0) = y_{0_1}, \\ y_2(0) = y_{0_2}. \end{cases} \tag{1.2.14}$$

20. Balthasar van der Pol, 1889–1959.

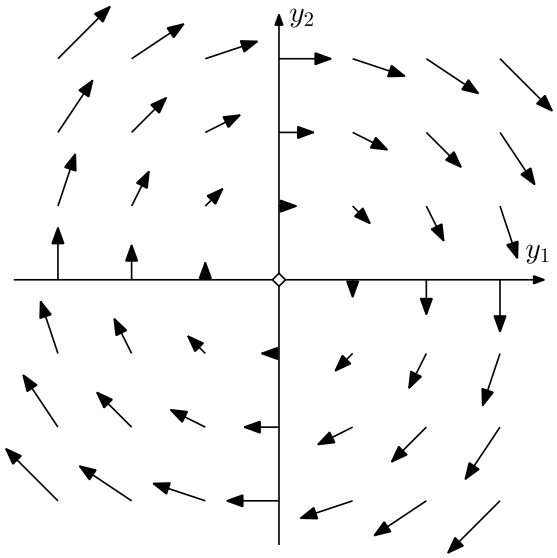


FIGURE 1.10 – L'espace des phases pour les petites oscillations du pendule. Les flèches représentent le champ de vitesses de phase, le losange indique la position du point singulier. Les trajectoires sont des ellipses toutes parcourues dans le même temps pour faire un tour complet. Un tel tour correspondant à une oscillation du pendule, on voit que la période ne dépend pas de l'amplitude : les oscillations sont harmoniques.

Quelques trajectoires sont représentées sur la Figure 1.13 en fonction du temps et dans l'espace des phases sur la Figure 1.14. Suivant la condition initiale $y(0)$, la solution converge vers l'un des trois points critiques, solution de $y'(t) = 0$, c'est-à-dire $(-1,-1)$, $(0,0)$ et $(1,1)$. ◊

En dimension supérieure à deux, la situation se complique considérablement. Certains systèmes ont des solutions bornées qui ne convergent vers aucun cycle limite ni point d'équilibre.

Exemple 1.2.9 L'exemple célèbre des équations de Lorenz [21] vient de la météorologie

$$\begin{cases} y_1'(t) = -\sigma(y_1(t) - y_2(t)), \\ y_2'(t) = -y_1(t)y_3(t) + ry_1(t) - y_2(t), \\ y_3'(t) = y_1(t)y_2(t) - by_3(t). \end{cases} \tag{1.2.15}$$

Ce système a des solutions non périodiques ou encore *chaotiques* pour certaines plages de valeurs des paramètres σ, r et b, comme la trajectoire représentée en 3D sur la Figure 1.16.

Il se trouve qu'après une brève période initiale transitoire, les trajectoires se rapprochent très vite d'une partie de $\mathbb{R}^3$ appelée « attracteur de Lorenz ». Cet attracteur est composé grossièrement de deux « lobes », sur lesquels les trajectoires évoluent en spirale de plus en plus larges. À des instants qui sont totalement imprévisibles, et inaccessibles en fait au calcul numérique, les trajectoires sautent d'un lobe à l'autre, voir Figure 1.15, c'est le côté chaotique de la chose. Le calcul numérique donne quand même une vision qualitativement correcte de l'évolution exacte, même si elle est quantitativement fausse, en un sens précis que nous ne détaillerons pas ici.

21. Edward Norton Lorenz, 1917–2008.

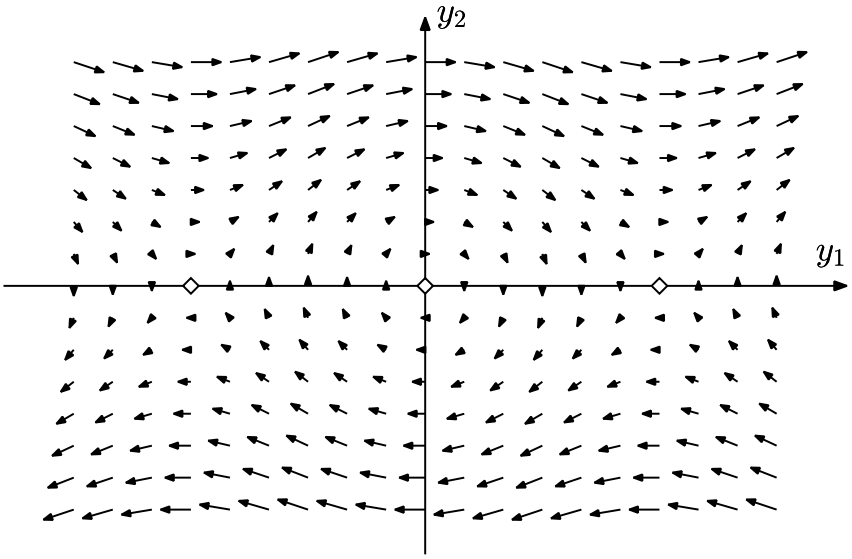


FIGURE 1.11 – L'espace des phases pour les grandes oscillations du pendule. Les flèches représentent le champ de vitesses de phase, les losanges indiquent la position des points singuliers. On distingue sur le champ de vecteurs les régions du plan de phase dans lesquelles les trajectoires correspondent à des oscillations (non harmoniques) du pendule et celles qui correspondent à des rotations complètes dans un sens ou dans l'autre.

La structure fine de l'attracteur de Lorenz n'est pas très bien connue. C'est un objet dont la dimension [22] est comprise entre 2 et 3, aux alentours de 2,06 apparemment. ◊

1.3 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* [23].

Définition 1.3.1 *On appelle* problème de Cauchy*, la conjonction d'une EDO* (1.2.1) *et d'une condition additionnelle, la* donnée initiale*,*

$$\begin{cases} y'(t) = f(t, y(t)), \text{ pour tout } t \in I, \\ y(t_0) = y_0, \end{cases} \tag{1.3.1}$$

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

22. Il s'agit ici de la dimension de Hausdorff.
23. Augustin Louis, baron Cauchy, 1789–1857.

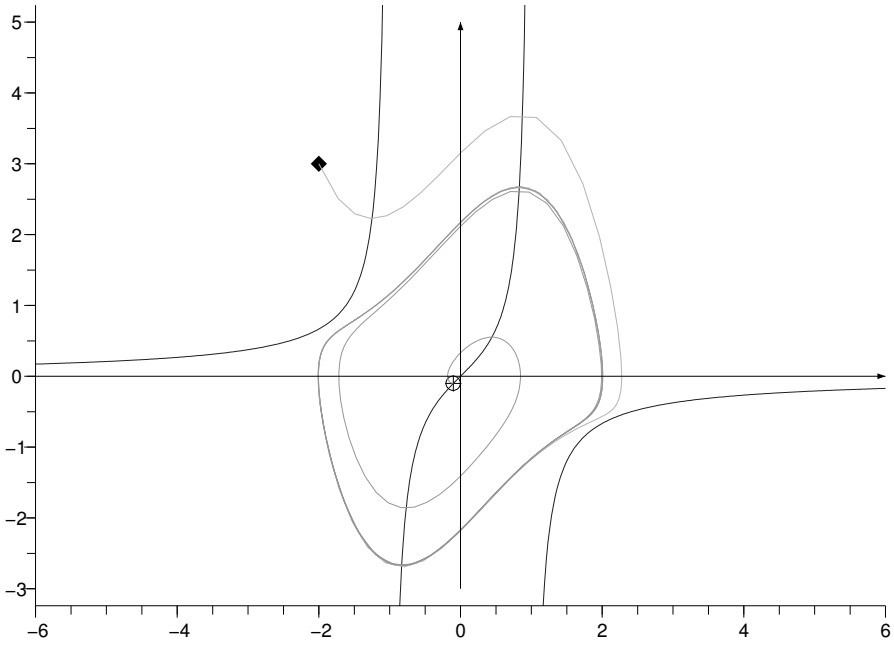


FIGURE 1.12 – Les trajectoires des solutions de 1.2.13 partant de $(-2, 3)$, et de $(-0.1, -0.1)$ pour $\varepsilon = 1$.

On considérera le plus souvent pour simplifier que $I =]0, T[$ et que $t_0 = 0$, ce qui ne nuit pas à la généralité [24]. 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. [25] 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...

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 [26]. 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,

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

24. Pourquoi d'ailleurs ?

25. 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.

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

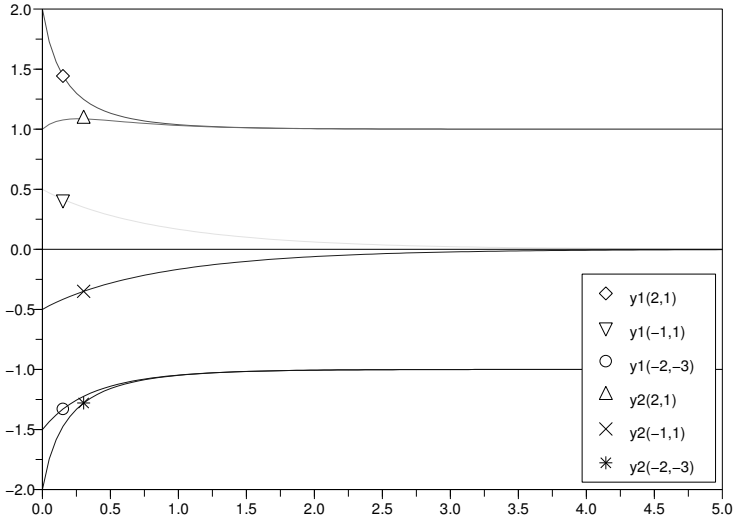


FIGURE 1.13 – Les trajectoires des solutions de 1.2.14 à partir de différentes conditions initiales.

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, [27] 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 est 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 :

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

27. Ce qui est rigoureusement interdit en mathématiques.

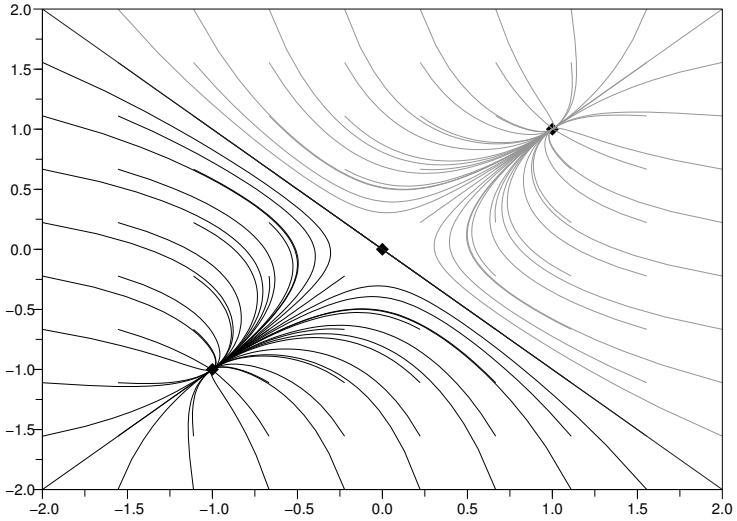


FIGURE 1.14 – Les trajectoires des solutions de 1.2.14 convergent vers un des trois points singuliers $(-1,-1)$, $(0,0)$ ou $(1,1)$.

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.

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

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.

Encore une fois, le raisonnement ci-dessus n'est pas à la hauteur, et de loin, des standards actuels en terme de rigueur mathématique. En clair, ce n'est pas une démonstration. Par contre, c'est quand même un moyen pratique de calcul, dont il se trouve qu'il donne le bon résultat — magie de la physique ! Voir néanmoins l'exemple 3.1.2 page 99 et la discussion qui suit cet exemple pour voir comment rendre l'argument ci-dessus rigoureux, ce qui n'est pas encore possible à ce stade du cours.

Un cas particulier, pas très intéressant en fait, est celui où le second membre ne dépend pas de y, c'est-à-dire $h(y) = 1$. Évidemment, dans ce cas, le problème de Cauchy $y'(t) = g(t)$, $y(0) = y_0$, se réduit à un simple calcul de primitives $y(t) = y_0 + \int_0^t g(s)\,ds$. De ce point de vue, les équations différentielles à variables séparées englobent le calcul des primitives. Ce que l'on vient de voir montre que réciproquement, les équations différentielles à variables séparées se ramènent au calcul des primitives, plus un calcul de fonction réciproque. Par contre, quand les variables ne sont pas séparées, on entre dans un territoire totalement nouveau dont on ne possède pas encore les clés.

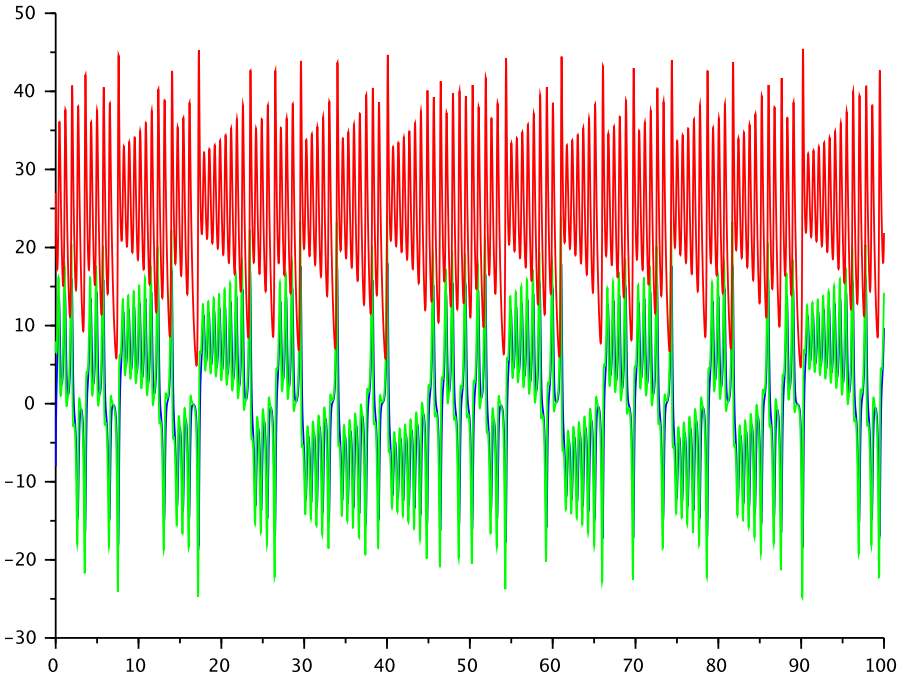


FIGURE 1.15 – Comportement chaotique des trois composantes (y_1 en bleu, y_2 en vert, y_3 en rouge, zoomer à l'écran pour mieux voir les différences entre le vert et le bleu) d'une solution des équations de Lorenz 1.2.15 pour $r = 28$, $b = 8/3$ et $\sigma = 10$.

De façon beaucoup plus intéressante, on note que les équations à variables séparées incluent celles dont le second membre ne dépend pas de t, c'est-à-dire $g(t) = 1$, que l'on appelle *équations autonomes* [28]. Dans ce cas, on prendra couramment $G(t) = t$.

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

Nous reviendrons plus loin en détail sur la résolution en général du problème de Cauchy (1.3.1). Retenons pour l'instant que, que ce soit en dimension $m > 1$ ou bien dans le cas où les variables ne sont pas séparées, les choses sont beaucoup moins simples que ce qui précède.

Donnons d'ores et déjà quelques éléments plutôt d'ordre historique dans cette direction. Pour la suite, on considère pour simplifier le cas scalaire $m = 1$. Si l'on se place dans l'état d'esprit des contemporains de Newton, qui viennent à peine de découvrir le calcul différentiel juste inventé par ledit Newton et par Leibniz [29], toutes les fonctions sont très régulières, même analytiques, c'est-à-dire développables en série entière. C'est donc le cas de la solution de l'EDO (on ne se pose même pas la question de son existence) qui est développable en série entière en la variable t au voisinage de l'origine ; elle y est égale à son développement de Taylor [30]

$$y(t) = \sum_{i=0}^{\infty} \frac{y^{(i)}(0)}{i!} t^i. \tag{1.3.2}$$

28. Nous y reviendrons.
29. Ne prenons pas trop parti dans la querelle de l'invention du calcul différentiel.
30. Brook Taylor, 1685–1731.

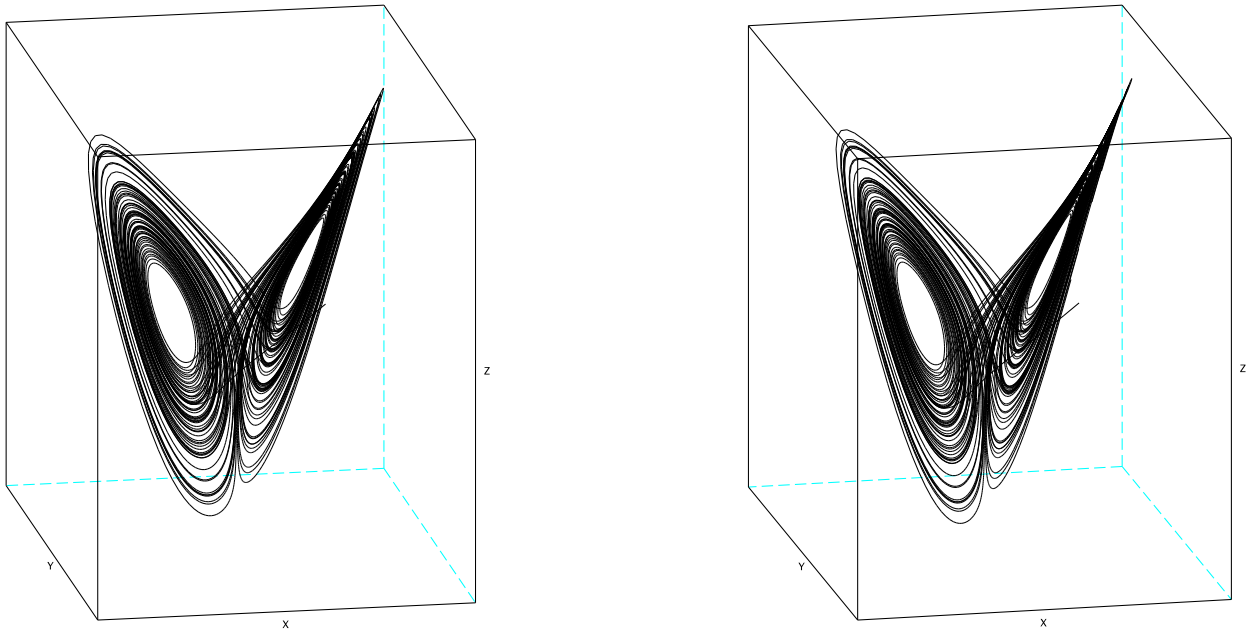


FIGURE 1.16 – Vue stéréoscopique à vision libre de l'attracteur de Lorenz : loucher vers le milieu pour le voir en 3d.

Il suffit donc de déterminer les coefficients de cette série, dont on ne se pose pas la question de la convergence non plus, pour déterminer la solution. On a donc là une approche constructive de l'existence et de l'unicité de la solution de (1.3.1) au voisinage de l'origine. C'est la méthode utilisée historiquement par Newton [9] pour calculer les solutions des équations différentielles. La Figure 1.18 donne un extrait des travaux de Newton traduits par Buffon [31]. Newton y explique comment résoudre de manière approchée et itérative l'équation différentielle $y'(x) = 1 - 3x + y(x) + x^2 + xy(x)$ avec la condition initiale $y(0) = 0$, où la variable est x et non t comme on l'aura remarqué.

Expliquons en langage moderne la façon de procéder de Newton. Il s'agit en fait d'écrire le développement de Taylor de y avec des coefficients indéterminés, $y(x) = a_0 + a_1 x + a_2 x^2 + \ldots$, et de calculer ces coefficients de proche en proche. C'est extrêmement astucieux et algorithmique.

Supposons que l'on veuille déterminer les 8 premiers termes du développement de Taylor de y en 0, juste pour aller un cran plus loin que Buffon. On connaît déjà le premier terme, qui est donné par la condition initiale, c'est 0. On forme donc un tableau à 8 colonnes (la traduction de Buffon n'en présente que deux, mais on va tracer les traits manquants pour plus de clarté). En première ligne et à partir de la deuxième colonne, on place les termes du développement de Taylor de $f(x, y)$ qui ne dépendent pas de y, par ordre de degré croissant. Ici donc 1 puis $-3x$ puis $+x^2$, puis des 0 (car f est polynomiale dans l'exemple de Newton). En première colonne et à partir de la deuxième ligne, on place les termes restants du développement de Taylor de f, puis une ligne « somme », qui va correspondre à y', puis une ligne y qui va contenir les primitives de la ligne précédente.

On remplit le tableau colonne par colonne de gauche à droite. Décrivons le remplissage des colonnes 2 et 3. Par la condition initiale nulle, on sait que $y(x) = 0 + a_1 x + a_2 x^2 + \ldots$. La colonne 2 contient les termes de degré 0 dans y, soit 0 et dans xy, soit 0 également. On somme les trois lignes pour obtenir le terme de degré 0 dans y', qui vaut donc 1. On l'intègre pour obtenir le terme de

31. Georges-Louis Leclerc, comte de Buffon, 1707–1788.

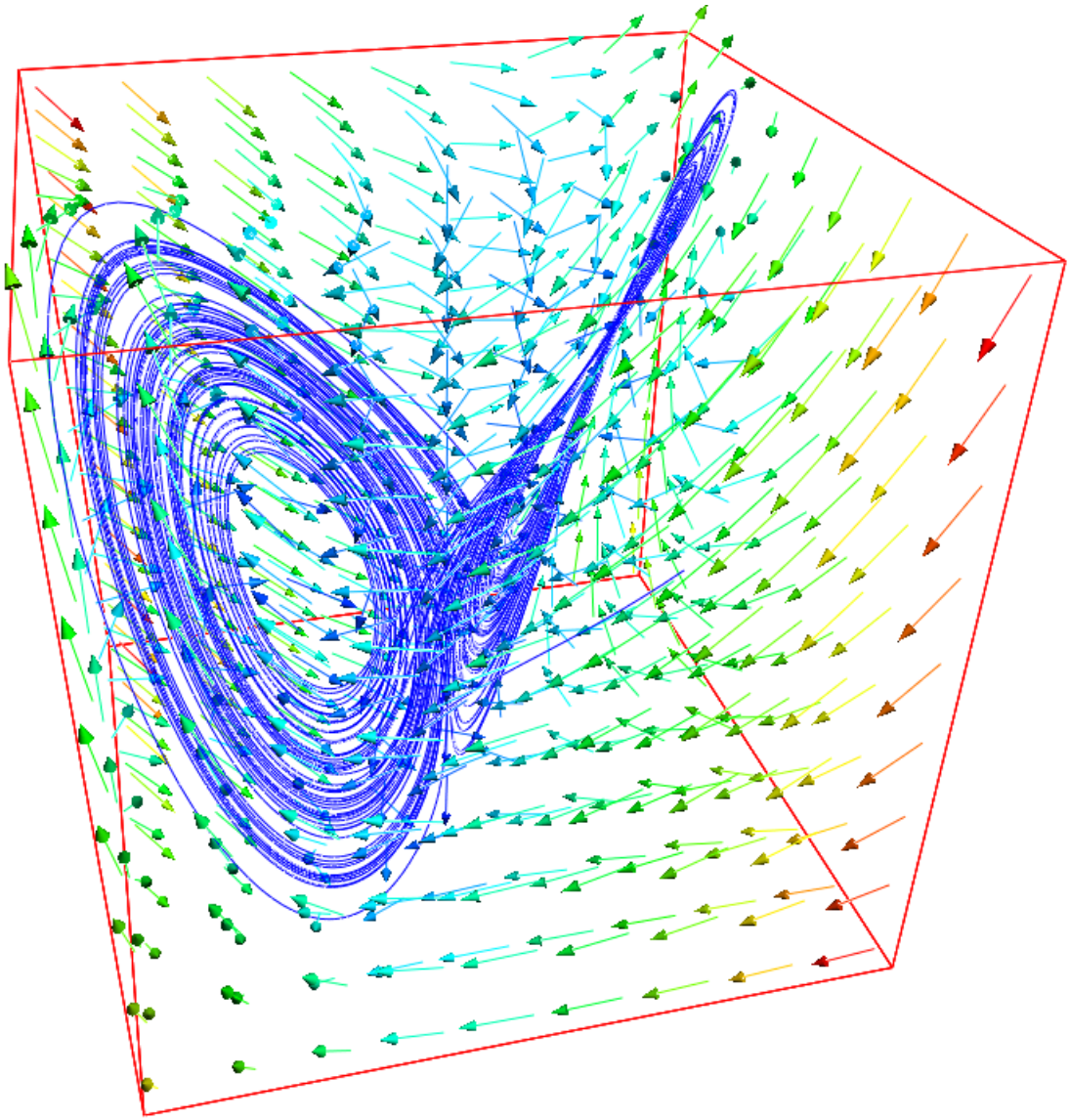


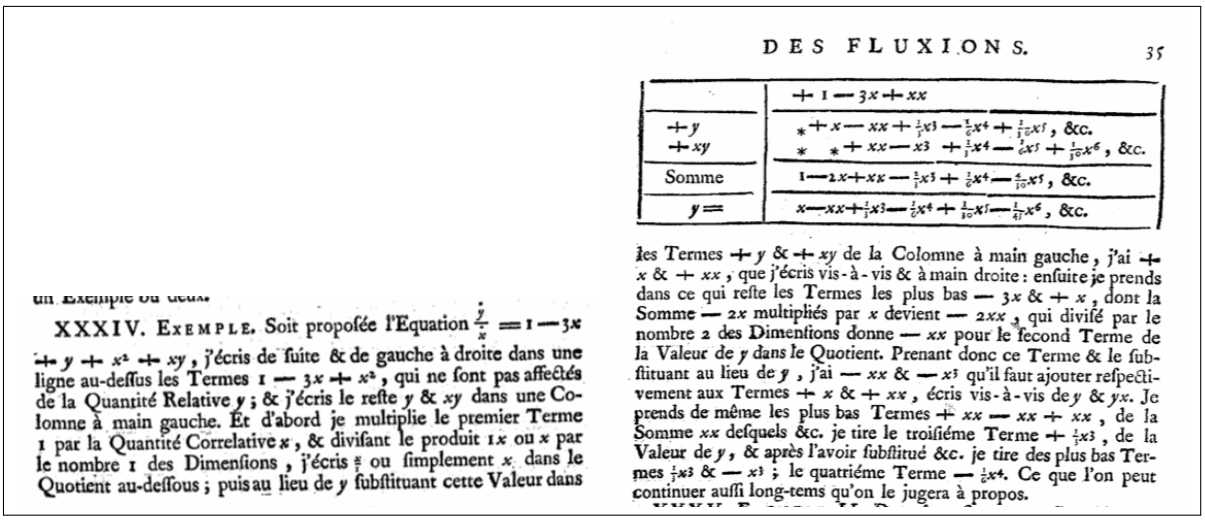
FIGURE 1.17 – L'attracteur de Lorenz plongé dans son champ de vecteurs.

degré 1 dans y, à savoir x, que l'on place en dernière ligne. On a donc déterminé $a_1 = 1$.

On passe à la colonne 3 en reportant le terme x que l'on vient de trouver en deuxième ligne (c'est le terme de degré 1 de y). Comme $xy = x(x + a_2x^2 + \ldots)$, le terme de degré 1 de xy est nul, donc on met 0 dans la case correspondante. On somme, il vient $-2x$ comme terme de degré 1 de y', donc $-x^2$ comme terme de degré 2 de y. On le reporte en ligne 2, colonne 4, et l'on recommence aussi longtemps que nécessaire.

Le résultat de l'algorithme est que $y(x) = x - x^2 + \frac{1}{3}x^3 - \frac{1}{6}x^4 + \frac{1}{30}x^5 - \frac{1}{45}x^6 + \frac{1}{630}x^7 + \&c.$, et ce calcul se fait de tête !

La méthode fut formalisée par Euler [5] et la démonstration de la convergence est due à Cauchy. En effet, on sait bien qu'une série entière peut parfaitement avoir un rayon de convergence nul, et ne pas définir une fonction. Cela ne gênait pas Newton et ses contemporains, mais du temps de Cauchy, on s'était aperçu qu'il y avait une vraie difficulté à cet endroit. De plus, on commençait à soupçonner sérieusement que toutes les fonctions ne sont peut-être pas analytiques.



DES FLUXIONS. 35

	$+1-3x+xx$
$+y$	$*+x-xx+\frac{1}{3}x^3-\frac{1}{6}x^4+\frac{1}{30}x^5$, &c.
$+xy$	$*\ *+xx-x3+\frac{1}{3}x^4-\frac{1}{6}x^5+\frac{1}{30}x^6$, &c.
Somme	$1-2x+xx-\frac{1}{3}x^3+\frac{1}{6}x^4-\frac{2}{30}x^5$, &c.
$y=$	$x-xx+\frac{1}{3}x^3-\frac{1}{6}x^4+\frac{1}{30}x^5-\frac{1}{47}x^6$, &c.

un Exemple ou deux.

XXXIV. Exemple. Soit propoſée l'Equation $\frac{\dot{y}}{\dot{x}}=1-3x+y+x^2+xy$, j'écris de ſuite & de gauche à droite dans une ligne au-deſſus les Termes $1-3x+x^2$, qui ne ſont pas affectés de la Quantité Relative y; & j'écris le reſte y & xy dans une Colomne à main gauche. Et d'abord je multiplie le premier Terme 1 par la Quantité Correlative x, & diviſant le produit $1x$ ou x par le nombre 1 des Dimenſions, j'écris $\frac{x}{1}$ ou ſimplement x dans le Quotient au-deſſous; puis au lieu de y ſubſtituant cette Valeur dans les Termes $+y$ & $+xy$ de la Colomne à main gauche, j'ai $+x$ & $+xx$, que j'écris vis-à-vis & à main droite: enſuite je prends dans ce qui reſte les Termes les plus bas $-3x$ & $+x$, dont la Somme $-2x$ multipliés par x devient $-2xx$, qui diviſé par le nombre 2 des Dimenſions donne $-xx$ pour le ſecond Terme de la Valeur de y dans le Quotient. Prenant donc ce Terme & le ſubſtituant au lieu de y, j'ai $-xx$ & $-x^3$ qu'il faut ajouter reſpectivement aux Termes $+x$ & $+xx$, écris vis-à-vis de y & yx. Je prends de même les plus bas Termes $+xx-xx+xx$, de la Somme xx deſquels &c. je tire le troiſiéme Terme $+\frac{1}{3}x^3$, de la Valeur de y, & après l'avoir ſubſtitué &c. je tire des plus bas Termes $\frac{1}{3}x^3$ & $-x^3$; le quatriéme Terme $-\frac{1}{6}x^4$. Ce que l'on peut continuer auſſi long-tems qu'on le jugera à propos.

FIGURE 1.18 – Intégration de l'équation différentielle $y' = 1 - 3x + y + x^2 + xy$ par la méthode des séries, Newton, 1671, traduction par Buffon (source Gallica.BnF.fr).

	1	$-3x$	$+x^2$	0	0	0	0
y	0	$+x$	$-x^2$	$+\frac{1}{3}x^3$	$-\frac{1}{6}x^4$	$+\frac{1}{30}x^5$	$-\frac{1}{45}x^6$
xy	0	0	$+x^2$	$-x^3$	$+\frac{1}{3}x^4$	$-\frac{1}{6}x^5$	$+\frac{1}{30}x^6$
somme ($= y'$)	1	$-2x$	$+x^2$	$-\frac{2}{3}x^3$	$+\frac{1}{6}x^4$	$-\frac{2}{15}x^5$	$+\frac{1}{90}x^6$
y	x	$-x^2$	$+\frac{1}{3}x^3$	$-\frac{1}{6}x^4$	$+\frac{1}{30}x^5$	$-\frac{1}{45}x^6$	$+\frac{1}{630}x^7$

FIGURE 1.19 – Newton en typographie LaTeX moderne.

L'existence et l'unicité de la solution du problème de Cauchy peuvent être établies sous des hypothèses beaucoup moins fortes que l'analyticité. Nous traiterons un cas relativement général grâce au théorème de Cauchy-Lipschitz [32] dans la section 1.5. En guise d'introduction, et parce que c'est le cas le plus simple, nous allons d'abord traiter le cas des équations différentielles linéaires, déjà abordé partiellement en L1.

1.4 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.

1.4.1 Définitions et propriétés générales

Définition 1.4.1 *On appelle équation différentielle linéaire sur l'intervalle I, toute équation différentielle de la forme*

$$\forall t \in I, \quad y'(t) = A(t)y(t) + b(t), \tag{1.4.1}$$

32. Rudolph Otto Sigismund Lipschitz, 1832–1903.

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

$$\begin{cases} y'(t) = A(t)y(t) + b(t), \\ y(0) = y_0. \end{cases} \tag{1.4.2}$$

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 1.4.2 (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

$$\forall t \in \bar{I}, \quad y(t) = y_0 e^{\int_0^t a(s)ds} + \int_0^t b(u) e^{\int_u^t a(s)\,ds}\, du. \tag{1.4.3}$$

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. [33] 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 [34]. Comme l'exponentielle ne

33. On prend ici celle qui s'annule en 0, mais c'est juste pour fixer les idées, ce n'est en rien nécessaire.

34. D'où le nom de la méthode.

s'annule jamais, c'est un changement de fonction inconnue complètement légitime. Une autre façon de le dire, mais dont il est moins facile de se rappeler, est de poser simplement $K(t) = e^{-\mathcal{A}(t)}y(t)$.

On écrit donc [35]

$$y'(t) = K'(t)e^{\mathcal{A}(t)} + K(t)a(t)e^{\mathcal{A}(t)} = a(t)y(t) + b(t) = a(t)K(t)e^{\mathcal{A}(t)} + b(t),$$

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

$$K'(t)e^{\mathcal{A}(t)} = b(t) \quad \text{d'où} \quad K'(t) = b(t)e^{-\mathcal{A}(t)},$$

d'où en intégrant, ce qui ne pose pas de problème puisque b est continue,

$$K(t) = K(0) + \int_0^t b(u)e^{-\mathcal{A}(u)}\,du.$$

Multipliant par $e^{\mathcal{A}(t)}$, il vient alors

$$y(t) = K(0)e^{\mathcal{A}(t)} + e^{\mathcal{A}(t)}\int_0^t b(u)e^{-\mathcal{A}(u)}\,du = K(0)e^{\mathcal{A}(t)} + \int_0^t b(u)e^{\mathcal{A}(t)-\mathcal{A}(u)}\,du,$$

et utilisant la condition initiale $y(0) = y_0$ et le fait que $\mathcal{A}(t) - \mathcal{A}(u) = \int_u^t a(s)\,ds$, on retrouve bien l'expression (1.4.3).

On a jusqu'ici établi *l'unicité* de la solution du problème de Cauchy : toute solution éventuelle ne peut être que de la forme (1.4.3). 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.

35. 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 $K' = -ae^{-\mathcal{A}}y + e^{-\mathcal{A}}y' = -ae^{-\mathcal{A}}y + e^{-\mathcal{A}}(ay + b)$.

Définition 1.4.3 *Étant donnée une équation différentielle linéaire* (1.4.1)*, on appelle équation différentielle linéaire sans second membre ou homogène associée, l'équation différentielle*

$$\forall t \in I, \quad y'(t) = A(t)y(t). \tag{1.4.4}$$

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 1.4.4 *Si y_1 et y_2 sont deux solutions de l'équation différentielle* (1.4.1)*, leur différence $y_2 - y_1$ est solution de l'équation différentielle homogène associée* (1.4.4).

Proposition 1.4.5 *L'ensemble des solutions de l'équation différentielle homogène* (1.4.4) *est un espace vectoriel sur $\mathbb{R}$.*

On en déduit que

Proposition 1.4.6 *L'ensemble des solutions de l'équation différentielle* (1.4.1) *forme un sous-espace affine [36] de l'espace vectoriel des applications de I dans $\mathbb{R}^m$.*

Démonstration. Soit S l'ensemble des solutions de l'équation différentielle (1.4.1). 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 1.4.4 à 1.4.6 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.

1.4.2 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 $Y(t) = \mathfrak{R}(Y(t)) + i\mathfrak{I}(Y(t))$, les deux fonctions à valeurs dans $\mathbb{R}^m$ ainsi obtenues sont solutions de la même EDO à valeurs réelles et l'on

36. 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.

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, $\|x\| = \left(\sum_{i=1}^m |x_i|^2\right)^{1/2}$ 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

$$\|Ax\| \leq \|A\|\|x\|, \tag{1.4.5}$$

pour tout vecteur x. Il s'agit d'une norme matricielle au sens où pour tout couple de matrices A et B, on a $\|AB\| \leq \|A\|\|B\|$. On en déduit immédiatement par récurrence que $\|A^k\| \leq \|A\|^k$ pour tout entier positif k.

On aura besoin de la notion d'exponentielle de matrices.

Proposition 1.4.7 *Pour tout 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 1.4.8 *Pour tout matrice $A \in M_m(\mathbb{C})$, on appelle* exponentielle de *A la somme de la série*

$$e^A = \sum_{k=0}^\infty \frac{1}{k!}A^k \in M_m(\mathbb{C}). \tag{1.4.6}$$

La démonstration de la proposition 1.4.7 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 va d'abord établir quelques propriétés utiles de l'exponentielle de matrice.

Théorème 1.4.9 *Si A et $B \in M_m(\mathbb{C})$ commutent alors*

$$e^{A+B} = e^A e^B.$$

Démonstration. Si A et B commutent, *i.e.*, $AB = BA$, alors on peut appliquer la formule du binôme [37] pour calculer

$$(A+B)^k = \sum_{j=0}^k C_k^j A^{k-j}B^j.$$

37. Si A et B ne commutent pas, alors cette formule est a priori fausse et la preuve s'effondre.

Par conséquent,

$$e^{A+B} = \sum_{k=0}^{\infty} \frac{1}{k!}(A+B)^k = \sum_{k=0}^{\infty} \frac{1}{k!}\Big(\sum_{j=0}^{k} C_k^j A^{k-j} B^j\Big).$$

Les coefficients du binôme sont donnés par $C_k^j = \frac{k!}{j!(k-j)!}$ (aussi notés $\binom{k}{j}$, malheureusement), par conséquent, après simplification du facteur $k!$, il vient

$$e^{A+B} = \sum_{k=0}^{\infty}\Big(\sum_{j=0}^{k}\Big(\frac{1}{(k-j)!}A^{k-j}\Big)\Big(\frac{1}{j!}B^j\Big)\Big) = \Big(\sum_{n=0}^{\infty}\frac{1}{n!}A^n\Big)\Big(\sum_{l=0}^{\infty}\frac{1}{l!}B^l\Big) = e^A e^B,$$

en reconnaissant dans la deuxième expression le produit des deux séries de la troisième expression. C'est évident pour le produit de deux séries à coefficients complexes, ça l'est moins ici où l'on manipule des matrices. On écrit donc les détails :

$$\begin{aligned}
\sum_{k=0}^{N}\Big(\sum_{j=0}^{k}\Big(\frac{1}{(k-j)!}A^{k-j}\Big)\Big(\frac{1}{j!}B^j\Big)\Big) - \Big(\sum_{n=0}^{N}\frac{1}{n!}A^n\Big)\Big(\sum_{l=0}^{N}\frac{1}{l!}B^l\Big) &= \sum_{j=0}^{N}\Big(\frac{1}{j!}B^j\Big)\sum_{k=j}^{N}\Big(\frac{1}{(k-j)!}A^{k-j}\Big) - \Big(\sum_{n=0}^{N}\frac{1}{n!}A^n\Big)\Big(\sum_{l=0}^{N}\frac{1}{l!}B^l\Big)\\
&= \sum_{j=0}^{N}\Big(\frac{1}{j!}B^j\Big)\sum_{k=0}^{N-j}\Big(\frac{1}{k!}A^k\Big) - \Big(\sum_{n=0}^{N}\frac{1}{n!}A^n\Big)\Big(\sum_{l=0}^{N}\frac{1}{l!}B^l\Big)\\
&= -\sum_{j=0}^{N}\Big(\frac{1}{j!}B^j\Big)\sum_{k=N-j+1}^{N}\Big(\frac{1}{k!}A^k\Big)
\end{aligned}$$

En passant à la norme matricielle on obtient

$$\Big\|\sum_{k=0}^{N}\Big(\sum_{j=0}^{k}\Big(\frac{1}{(k-j)!}A^{k-j}\Big)\Big(\frac{1}{j!}B^j\Big)\Big) - \Big(\sum_{n=0}^{N}\frac{1}{n!}A^n\Big)\Big(\sum_{n=0}^{N}\frac{1}{n!}B^n\Big)\Big\| \leq \sum_{j=0}^{N}\Big(\frac{1}{j!}\|B\|^j\Big)\sum_{k=N-j+1}^{N}\Big(\frac{1}{k!}\|A\|^k\Big)$$

En refaisant les mêmes calculs dans $\mathbb{R}$ pour les sommes partielles dans l'expression de $e^{\|A\|+\|B\|} = e^{\|A\|}e^{\|B\|}$ on obtient que

$$\sum_{k=0}^{N}\Big(\sum_{j=0}^{k}\Big(\frac{1}{(k-j)!}\|A\|^{k-j}\Big)\Big(\frac{1}{j!}\|B\|^j\Big)\Big) - \Big(\sum_{n=0}^{N}\frac{1}{n!}\|A\|^n\Big)\Big(\sum_{n=0}^{N}\frac{1}{n!}\|B\|^n\Big) = -\sum_{j=0}^{N}\Big(\frac{1}{j!}\|B\|^j\Big)\sum_{k=N-j+1}^{N}\Big(\frac{1}{k!}\|A\|^k\Big) \qquad (*)$$

Donc

$$\lim_{N\to+\infty}\sum_{j=0}^{N}\Big(\frac{1}{j!}\|B\|^j\Big)\sum_{k=N-j+1}^{N}\Big(\frac{1}{k!}\|A\|^k\Big) = 0 \qquad (**)$$

donc le terme de gauche dans (*), qui tend vers $\big\|e^{A+B} - e^A e^B\big\|$ quand N tend vers o, est majoré par (**) qui tend vers o. ◊

Attention, il est facile de trouver deux matrices A et B qui ne commutent pas et telles que $e^{A+B} \neq e^A e^B$! [38] La formule de Baker ?Campbell ?Hausdorff [39] 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 1.4.9.

Le théorème 1.4.9 a plusieurs conséquences utiles.

38. 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 $e^{A+B} = e^A e^B$...

39. Henry Frederick Baker, 1866–1956 ; John Edward Campbell, 1862–1924 ; Felix Hausdorff, 1868–1942.

Corollaire 1.4.10 *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 $(e^{tA})' = Ae^{tA} = e^{tA}A$.

Démonstration. i) Il est bien clair que A et $-A$ commutent, donc $e^{-A}e^A = e^Ae^{-A} = e^{A-A} = e^0 = I$.
ii) Fixons $t \in \mathbb{R}$. Pour tout $h \in \mathbb{R}^*$, on a

$$\frac{e^{(t+h)A} - e^{tA}}{h} = \frac{e^{hA}e^{tA} - e^{tA}}{h} = \frac{e^{hA} - I}{h}e^{tA},$$

car tA et hA commutent. Utilisant maintenant la série entière, on obtient

$$\frac{e^{hA} - I}{h} = \frac{\sum_{k=1}^{\infty} \frac{h^k}{k!}A^k}{h} = A + \sum_{k=2}^{\infty} \frac{h^{k-1}}{k!}A^k.$$

Or pour $|h| \leq 1$, il vient

$$\left\| \sum_{k=2}^{\infty} \frac{h^{k-1}}{k!}A^k \right\| \leq |h| \sum_{k=2}^{\infty} \frac{|h|^{k-2}}{k!}\|A\|^k \leq |h| \sum_{k=2}^{\infty} \frac{1}{k!}\|A\|^k \leq |h| \sum_{k=0}^{\infty} \frac{1}{k!}\|A\|^k \leq |h|e^{\|A\|}.$$

On voit donc que $\left\| \frac{e^{hA}-I}{h} - A \right\| \to 0$ quand $h \to 0$, c'est-à-dire

$$\frac{e^{hA} - I}{h} \longrightarrow A \text{ quand } h \to 0,$$

ce qui implique en multipliant à droite par e^{tA} que

$$\frac{e^{(t+h)A} - e^{tA}}{h} \longrightarrow Ae^{tA} \text{ quand } h \to 0.$$

En effet,

$$\left\| \frac{e^{(t+h)A} - e^{tA}}{h} - Ae^{tA} \right\| = \left\| \left(\frac{e^{hA} - I}{h} - A \right) e^{tA} \right\| \leq \left\| e^{tA} \right\| \left\| \frac{e^{hA} - I}{h} - A \right\| \to 0 \text{ quand } h \to 0.$$

On vient juste d'obtenir la dérivabilité et la première formule pour la dérivée. On obtient la deuxième formule soit en factorisant de l'autre côté au départ, soit en notant que A commute évidemment avec e^{tA}, en effet, A commute trivialement avec A^k pour tout entier k. ◊

Une autre façon de voir que e^A est toujours inversible est de montrer que $\det(e^A) = e^{\operatorname{tr} A}$ (exercice).

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 [40] 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.

40. 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 [41] 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. [42] 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 1.4.11 *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

$$y(t) = e^{tA}y_0 + \int_0^t e^{(t-s)A}b(s)\,ds, \tag{1.4.7}$$

pour tout $t \in \mathbb{R}$.

41. Hermann Amandus Schwarz, 1843–1921.

42. On a bien sûr le même résultat pour la dérivée d'un produit de matrices : $(AB)' = A'B + AB'$.

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

$$z'(t) = (e^{-tA})'y(t) + e^{-tA}y'(t) = -e^{-tA}Ay(t) + e^{-tA}(Ay(t) + b(t)) = e^{-tA}b(t),$$

d'après le corollaire 1.4.10 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 1.4.10 i), on a $y(t) = e^{tA}z(t)$, d'où la formule (1.4.7) (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 (1.4.7) 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 1.4.10 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 1.4.12 *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 1.4.11 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* [43] des matrices.

Théorème 1.4.13 *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*

$$A = D + N. \tag{1.4.8}$$

43. Nelson Dunford, 1906–1986.

On rappelle que si D est diagonalisable, il existe Δ diagonale et une matrice de passage [44] 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 $N = 0$.

Proposition 1.4.14 *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 [45] 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 (1.4.8) et du théorème 1.4.9.

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 $n \longrightarrow +\infty$, 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. ◊

44. 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.

45. On fait figurer plusieurs fois les éventuelles valeurs propres multiples de D. Notons que ce sont aussi les valeurs propres de A.

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

$$y(t) = e^{tA}y_0 = \left(\sum_{k=0}^{p-1} \frac{t^k}{k!}N^k\right)P^{-1}\mathrm{diag}\left(e^{\lambda_i t}\right)Py_0,$$

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 [46] de t. On voit apparaître une partie polynomiale par rapport à t et une 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.

Exemple 1.4.1 On cherche deux fonctions $u_1(t)$ et $u_2(t)$ solutions du système différentiel

$$\begin{cases} u_1'(t) = au_1(t) + u_2(t) \\ u_2'(t) = 4u_1(t) + au_2(t), \end{cases}$$

avec $a \in \mathbb{R}$. On le réécrit $u'(t) = Au(t)$, avec $u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ et $A = \begin{pmatrix} a & 1 \\ 4 & a \end{pmatrix}$. La somme des valeurs propres vaut $2a$ et leur produit $a^2 - 4$, donc on intuite [47] immédiatement les valeurs propres $a - 2$ et $a + 2$. Elles sont distinctes donc la matrice A est diagonalisable. Elles sont réelles, donc on a des vecteurs propres réels, que l'on peut choisir égaux respectivement $\begin{pmatrix} 1 \\ -2 \end{pmatrix}$ et $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$. En résumé

$$A = P^{-1}DP,\ D = \begin{pmatrix} a-2 & 0 \\ 0 & a+2 \end{pmatrix},\ P^{-1} = \begin{pmatrix} 1 & 1 \\ -2 & 2 \end{pmatrix},\ P = \begin{pmatrix} 1/2 & -1/4 \\ 1/2 & 1/4 \end{pmatrix},$$

et la solution u du problème de Cauchy pour la condition initiale $u(0) = u_0$ est $u(t) = e^{tA}u_0$ avec

$$e^{tA} = P^{-1}e^{tD}P = P^{-1}\begin{pmatrix} e^{(a-2)t} & 0 \\ 0 & e^{(a+2)t} \end{pmatrix}P = e^{at}\begin{pmatrix} \cosh 2t & \frac{1}{2}\sinh 2t \\ 2\sinh 2t & \cosh 2t \end{pmatrix}.$$

Ici A est diagonalisable, il n'y a pas de termes polynomiaux en t. ◊

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

46. 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.

47. Si on n'intuite pas, on calcule le polynôme caractéristique et ses racines.

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* [48] 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 1.4.15 *Soit l'équation différentielle scalaire d'ordre n à coefficients constants homogène*

$$y^{(n)}(t) + a_1 y^{(n-1)}(t) + \cdots + a_{n-1} y'(t) + a_n y(t) = 0, \tag{1.4.9}$$

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 (1.4.9) 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 1.2.4. 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, [49] 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 1.4.12. 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 !). [50] 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 (1.4.9) 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.

Plutôt que de faire appel à un argument astucieux d'algèbre linéaire, on peut également vérifier directement que tout élément de E est solution de l'EDO. Par linéarité, il suffit pour cela de vérifier

48. Marie Ennemond Camille Jordan, 1838–1922. Attention, c'est un garçon.
49. Bon, peut-être au signe près.
50. On rappelle que la dimension de l'espace des polynômes de degré inférieur à $\alpha_i - 1$ est α_i.

que $t \mapsto t^\beta e^{\lambda_i t}$ avec $\beta \leq \alpha_i - 1$ en est une. Pour cela, pour tout polynôme $P(X) = \sum_{j=0}^n a_{n-j} X^j$, on va utiliser la notation $P(v)$ pour désigner la fonction $t \mapsto \sum_{j=0}^n a_{n-j} v^{(j)}(t)$ (on remplace la puissance j par une dérivée j-ème). On veut donc montrer que, sous les hypothèses précédentes, si $v_\beta(t) = t^\beta e^{\lambda_i t}$, alors $P(v_\beta) = 0$.

On commence par noter que pour toute fonction w, $(tw)^{(j)} = jw^{(j-1)} + tw^{(j)}$ pour tout j par une récurrence immédiate. Écrivant alors $v_\beta(t) = t\left(t^{\beta-1}e^{\lambda_i t}\right) = tv_{\beta-1}(t)$, il vient

$$\begin{aligned} P(v_\beta)(t) &= \sum_{j=0}^n a_{n-j} v_\beta^{(j)}(t) \\ &= \sum_{j=0}^n j a_{n-j} v_{\beta-1}^{(j-1)}(t) + t\left(\sum_{j=0}^n a_{n-j} v_{\beta-1}^{(j)}(t)\right) \\ &= P'(v_{\beta-1})(t) + tP(v_{\beta-1})(t). \end{aligned}$$

Itérant le processus β fois, on obtient

$$\begin{aligned} P(v_\beta)(t) &= P'(v_{\beta-1})(t) + tP(v_{\beta-1})(t) \\ &= P''(v_{\beta-2})(t) + 2tP'(v_{\beta-2})(t) + t^2 P(v_{\beta-2})(t) \\ &= \sum_{m=0}^{\beta} C_\beta^m t^{\beta-m} P^{(m)}(v_0)(t). \end{aligned}$$

Or $v_0(t) = e^{\lambda_i t}$, si bien que pour n'importe quel polynôme Q, on a $Q(v_0)(t) = Q(\lambda_i)e^{\lambda_i t}$, où dans le membre de droite on prend la valeur de Q en λ_i. Par conséquent,

$$P(v_\beta)(t) = \left(\sum_{m=0}^{\beta} C_\beta^m P^{(m)}(\lambda_i) t^{\beta-m}\right) e^{\lambda_i t}.$$

Or λ_i est une racine de P de multiplicité α_i et l'on a $\beta \leq \alpha_i - 1$. Par conséquent, $P^{(m)}(\lambda_i) = 0$ pour tout $m \leq \alpha_i - 1$, donc tout m apparaissant dans cette somme, d'où $P(v_\beta) = 0$.

Bon, finalement ce n'était pas si mal, l'algèbre linéaire...

Exemple 1.4.2 Soit l'EDO d'ordre 3, $y''' + 3y'' + 3y' + y = 0$. Son polynôme caractéristique est $P(X) = X^3 + 3X^2 + X + 1 = (X+1)^3$, avec la racine triple -1, donc les solutions sont de la forme

$$y(t) = (at^2 + bt + c)e^{-t}.$$

Pour avoir les solutions réelles, on prend a, b et c réels. ◊

Exemple 1.4.3 Soit l'EDO d'ordre 4, $y^{(4)} + 2y'' + y = 0$. Son polynôme caractéristique est $P(X) = X^4 + 2X^2 + 1 = (X^2+1)^2 = (X-i)^2(X+i)^2$ dont les deux racines doubles sont i et $-i$. Donc les solutions sont de la forme

$$y(t) = (at + b)e^{-it} + (ct + d)e^{it} = (\alpha t + \beta)\cos t + (\gamma t + \delta)\sin t.$$

Pour avoir les solutions réelles, on prend α, β, γ et δ réels dans la deuxième expression. ◊

Donnons sans démonstration le résultat suivant pour le cas non homogène.

Théorème 1.4.16 *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 $\lambda = \lambda_i$.

Exemple 1.4.4 Cas de l'oscillateur harmonique
On considère un système masse-ressort de masse $m > 0$, de raideur $k > 0$, soumis à une excitation sinusoïdale de fréquence f et d'intensité F, subissant un frottement linéaire de coefficient $C \geq 0$. On suppose connues la position initiale x_0 et la vitesse initiale v_0. La position de la masse m est donnée par $x(t)$ solution de l'équation différentielle linéaire du second ordre

$$mx''(t) + Cx'(t) + kx(t) = F\sin(\omega t).$$

où $\omega = 2\pi f$ est la pulsation de l'excitation. Cette équation est une conséquence immédiate de la loi fondamentale de la dynamique de Newton. Elle peut se résoudre grâce au théorème précédent. Nous ne le ferons pas et donnerons juste les résultats (en principe bien connus). On pose $\omega_0 = \sqrt{\frac{k}{m}}$ et $f_0 = \frac{\omega_0}{2\pi}$. On distingue quatre régimes qualitativement différents.

1. **Sans excitation et sans frottement** ($F = 0$ et $C = 0$). Les solutions sont sinusoïdales

$$x(t) = x_0\cos(\omega_0 t) + \frac{v_0}{\omega_0}\sin(\omega_0 t).$$

La fréquence des oscillations f_0 est la fréquence propre de l'oscillateur, ω_0 est la pulsation propre.

2. **Sans excitation et avec frottement** ($F = 0$ et $C > 0$). Trois cas possibles suivant la nature des racines de l'équation caractéristique $m\lambda^2 + C\lambda + k = 0$. On pose $\Delta = C^2 - 4km$.
Si $\Delta > 0$, les deux racines sont simples et réelles strictement négatives : $\lambda_- = \frac{-C-\sqrt{\Delta}}{2m} < \lambda_+ = \frac{-C+\sqrt{\Delta}}{2m} < 0$ et le système revient à la position d'équilibre $x = 0$ avec un amortissement exponentiel sans oscillation

$$x(t) = Ae^{\lambda_- t} + Be^{\lambda_+ t}.$$

Ceci correspond au cas où le frottement est important vis-à-vis de la raideur du ressort et de la masse. Si $\Delta < 0$, les deux racines sont simples et complexes conjuguées, à partie réelle strictement négative. Le système revient à la position d'équilibre exponentiellement vite en oscillant à la fréquence $\frac{\sqrt{-\Delta}}{4\pi m}$

$$x(t) = e^{\frac{-Ct}{2m}}\left[A\cos\left(\frac{\sqrt{-\Delta}}{2m}t\right) + B\sin\left(\frac{\sqrt{-\Delta}}{2m}t\right)\right].$$

Ceci correspond au cas où le frottement est faible vis-à-vis de la raideur du ressort et de la masse. Si $\Delta = 0$, cas limite, il y a une racine double réelle strictement négative $\lambda = -\frac{C}{2m}$. Le système revient à la position d'équilibre sans oscillation comme une exponentielle décroissante multipliée par une fonction affine de t

$$x(t) = (At + B)e^{\frac{-Ct}{2m}}.$$

Dans les trois cas, les constantes A et B sont déterminées par les conditions initiales en résolvant un système linéaire 2×2.

3. **Avec excitation et frottement**. La solution est la somme d'une solution de l'équation sans excitation décrite au cas précédent qui tend donc exponentiellement vite vers o et décrit un régime transitoire, et d'une solution particulière de forme sinusoïdale, de même fréquence que l'excitation et déphasée par rapport elle

$$x(t) = a\sin(\omega t + \varphi),$$

d'amplitude

$$a = \frac{F}{\sqrt{m^2(\omega^2 - \omega_0^2)^2 + C^2\omega^2}}.$$

Le déphasage est solution de l'équation

$$\tan\varphi = \frac{C\omega}{m(\omega^2 - \omega_0^2)}.$$

Le maximum de l'amplitude de la réponse sinusoïdale à l'excitation s'obtient quand cette dernière a une fréquence égale à $f = f_0$ la fréquence propre de l'oscillateur, qu'on appelle pour cette raison la fréquence de résonance.

4. **Résonance sans frottement** ($f = f_0$, $C = 0$). Si la fréquence de l'excitation est égale à la fréquence de résonance , la solution est la somme d'une fonction sinusoïdale et d'une fonction de la forme $At\sin(\omega_0 t)$ dont les valeurs absolues maximales tendent vers l'infini linéairement avec le temps : c'est le phénomène de résonance pure.

◊

1.4.3 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. [51]

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 (1.3.1).

Proposition 1.4.17 *Soit y une solution du problème de Cauchy* (1.4.2)*, continue sur* $[0,T]$*. Alors on a pour tout* $t \in [0,T]$

$$y(t) = y_0 + \int_0^t \big(A(s)y(s) + b(s)\big)\,ds. \tag{1.4.10}$$

Réciproquement, soit y une fonction continue sur $[0,T]$ *à valeurs dans* $\mathbb{R}^m$ *et satisfaisant l'équation intégrale* (1.4.10)*. Alors, y est dérivable sur* $]0,T[$ *et solution du problème de Cauchy* (1.4.2)*.*

51. Essayer de s'en convaincre en tentant de le démontrer : on doit buter très vite sur des problèmes de non commutativité.

Démonstration. Soit y une solution du problème de Cauchy (1.4.2), 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 (1.4.10).

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 1.4.18 *On suppose A et b continues sur $\bar{I} = [0,T]$. Le problème de Cauchy* (1.4.2) *admet une solution unique y.*

Démonstration. On va utiliser la forme intégrale de la proposition 1.4.17. On définit sur $\bar{I}$ par récurrence une suite de fonctions vectorielles

$$\forall t \in \bar{I}, \quad y_0(t) = y_0, \tag{1.4.11}$$

$$y_{n+1}(t) = y_0 + \int_0^t \big(A(s)y_n(s) + b(s)\big)\,ds. \tag{1.4.12}$$

Le premier terme de la suite est donc la fonction constante égale à la condition initiale, puis on applique l'itération (1.4.12) 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 (1.4.10), 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 1.5.2) et est connue dans la littérature, pour la partie existence, sous le nom de *méthode de Picard* [52].

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 (1.4.5), 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

$$y_{n+1}(t) - y_n(t) = \int_0^t A(s)\big(y_n(s) - y_{n-1}(s)\big)\,ds.$$

52. Charles Émile Picard, 1856–1941.

en soustrayant l'égalité (1.4.12) 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

$$\forall t \in [0,T], \quad \|y_{n+1}(t)-y_n(t)\| \leq \frac{\alpha\|y_0\|+\beta}{\alpha}\frac{(\alpha t)^{n+1}}{(n+1)!}. \tag{1.4.13}$$

On procède par récurrence. Tout d'abord, l'estimation (1.4.13) 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 (1.4.13) étant maintenant établie, il s'ensuit que pour tout $n \geq 0$,

$$\|y_{n+1}-y_n\|_{C^0([0,T];\mathbb{R}^m)} = \max_{t\in[0,T]}\|y_{n+1}(t)-y_n(t)\| \leq \max_{t\in[0,T]}\left(\frac{\alpha\|y_0\|+\beta}{\alpha}\frac{(\alpha t)^{n+1}}{(n+1)!}\right) = \frac{\alpha\|y_0\|+\beta}{\alpha}\frac{(\alpha T)^{n+1}}{(n+1)!}.$$

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

$$\sum_{n=1}^{\infty}\|y_n-y_{n-1}\|_{C^0([0,T];\mathbb{R}^m)} \leq \frac{\alpha\|y_0\|+\beta}{\alpha}\sum_{n=1}^{\infty}\frac{(\alpha T)^n}{n!} = \frac{\alpha\|y_0\|+\beta}{\alpha}\big(e^{\alpha T}-1\big).$$

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 [53]

$$y_n = y_0 + \sum_{k=1}^{n}(y_k-y_{k-1}),$$

on voit que la suite y_n converge uniformément sur $[0,T]$ vers la fonction continue

$$y = y_0 + \sum_{k=1}^{\infty}(y_k-y_{k-1}).$$

53. Cette astuce de passer par une somme télescopique qui se trouve être une série convergente est assez courante.

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 (1.4.12) pour obtenir (1.4.10).

Prenons maintenant deux solutions y et $\tilde{y}$. Leur différence $z = y - \tilde{y}$ vérifie

$$z(t) = \int_0^t A(s)z(s)\,ds. \tag{1.4.14}$$

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 (1.4.17) 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 [54] suivant le pays dans lequel on se trouve) sur les applications strictement contractantes dans un espace métrique complet. [55] 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 1.20 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 $y(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

Corollaire 1.4.19 *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 1.4.5. [56] 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 1.4.18, 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 $\dim S = m$. ◊

54. Stefan Banach, 1892–1945.

55. 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.

56. Considérée comme évidente, donc donnée sans preuve...

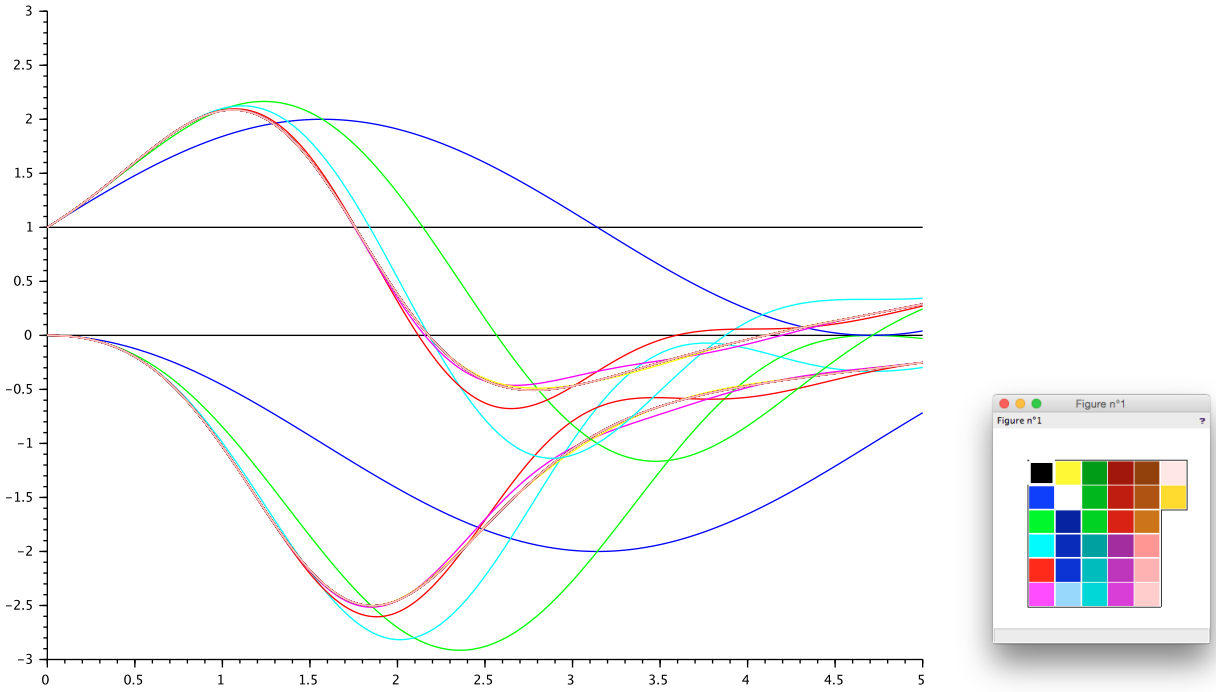


FIGURE 1.20 – Un exemple d'itérations de Picard. Les courbes de même couleur correspondent aux deux composantes de la même itération, en commençant par noir, bleu foncé, etc. À droite, l'ordre des couleurs de la palette par défaut de scilab.

Exemple 1.4.5 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 1.4.18 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

$$y_2(t) = y_0 + \int_0^t Ay_1(s)\, ds = y_0 + \int_0^t A(y_0 + sAy_0)\, ds$$
$$= Iy_0 + tAy_0 + \frac{t^2}{2}A^2y_0 = \left[(tA)^0 + (tA)^1 + \frac{(tA)^2}{2}\right]y_0.$$

Une récurrence immédiate (mais ce n'est pas une raison de ne pas la faire en détail) conduit alors à

$$y_n(t) = \left[(tA)^0 + (tA)^1 + \frac{(tA)^2}{2} + \cdots + \frac{(tA)^n}{n!}\right]y_0 = \left[\sum_{k=0}^{n} \frac{(tA)^k}{k!}\right]y_0.$$

Comme on l'a vu dans le cas général, la suite $(y_n)_{n\in\mathbb{N}}$ 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 1.4.20 *Soient y_0 et $\tilde{y}_0$ deux conditions initiales pour le problème de Cauchy* (1.4.2)*, 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 (1.4.2) 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 [57] 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. ◊

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

57. Dans le cas complexe, il faut prendre la partie réelle du produit scalaire hermitien sur $\mathbb{C}^m$, pour lequel $(u|v) = \overline{(v|u)}$.

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 1.4.18 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 1.4.21 *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 1.4.19. 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 1.4.22 *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 1.4.23 *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* [58], est solution de l'EDO scalaire $\frac{d}{dt}(\det W)(t) = \operatorname{tr} A(t) \det W(t)$. Cette équation à 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 1.4.24 *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 1.4.25 *Supposons que pour tous $t, s \in I$, $A(t)A(s) = A(s)A(t)$. Alors on a*

$$R(t,t_0) = e^{\int_{t_0}^{t} A(s)\,ds}.$$

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. [59] Reprenant la série entière qui définit l'exponentielle, on en déduit que

$$\frac{d}{dt}\big(e^{F(t)}\big) = e^{F(t)}F'(t) = F'(t)e^{F(t)},$$

58. Josef Hoëné-Wronski, 1776–1853.

59. Ces formules sont bien sûr violemment fausses si $F(t)$ et $F'(t)$ ne commutent pas, ce qui est le cas général.

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, $R(t, t_0) = e^{(t-t_0)A}$.

1.5 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 1.5.2 le cas où f est *globalement lipschitzienne*, et où il y a existence et unicité d'une solution sur tout l'intervalle de temps $\bar{I}$.

1.5.1 Approche historique par la méthode d'Euler

Dans ce paragraphe, nous allons décrire sans démonstration et sans être très précis, une première approche pour l'existence d'une solution du problème de Cauchy. Son intérêt en tant que telle est plus historique qu'autre chose et elle sera rapidement supplantée par des théorèmes plus généraux, précis et complètement démontrés. Cette approche annonce également la problématique qui sera la nôtre dans la deuxième partie 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 (1.3.1).

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. [60] 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*

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

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, $i = 0, \ldots, n$, 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

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

60. 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.*

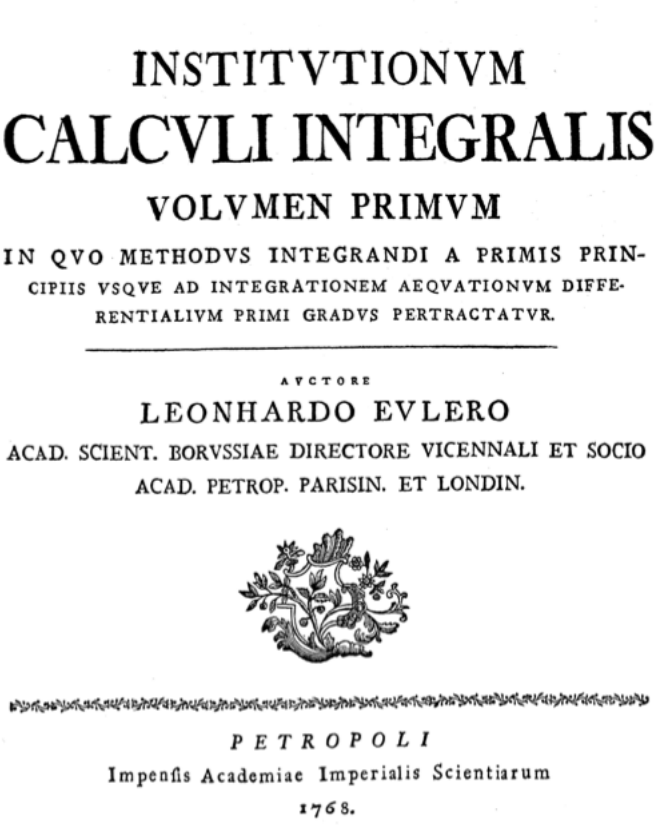

INSTITVTIONVM
CALCVLI INTEGRALIS
VOLVMEN PRIMVM
IN QVO METHODVS INTEGRANDI A PRIMIS PRINCIPIIS VSQVE AD INTEGRATIONEM AEQVATIONVM DIFFERENTIALIVM PRIMI GRADVS PERTRACTATVR.
AVCTORE
LEONHARDO EVLERO
ACAD. SCIENT. BORVSSIAE DIRECTORE VICENNALI ET SOCIO ACAD. PETROP. PARISIN. ET LONDIN.
PETROPOLI
Impensis Academiae Imperialis Scientiarum
1768.

FIGURE 1.21 – La source.

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, d'autant plus que l'on n'a pas encore montré que la solution y existe...

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

$$y_h(t) = y_i + (t - t_i) f(t_i, y_i), \quad \text{pour } t \in [t_i, t_{i+1}]. \tag{1.5.1}$$

La Figure 1.22 représente ce polygone pour $n = 5$, pour la fonction $y(t) = t(1 - t)$ solution 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 1.5.1 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,

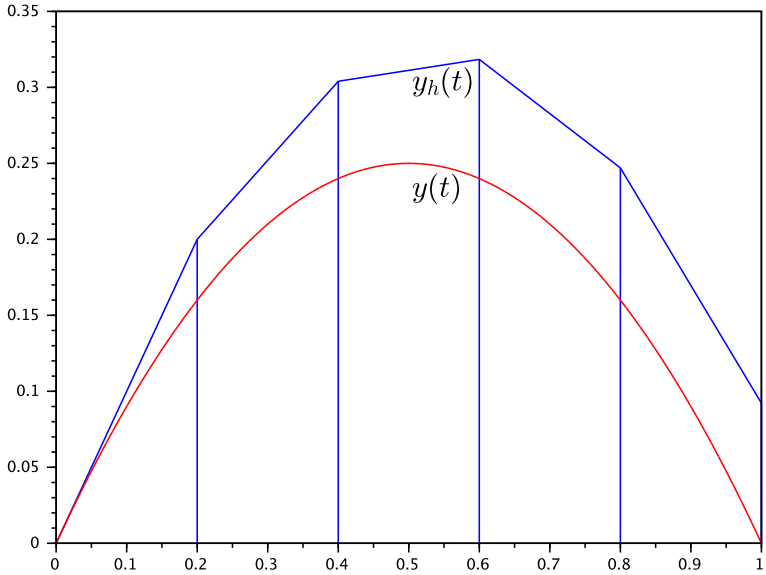


FIGURE 1.22 – Polygone d'Euler pour $y'(t) = 1 - 2t^2 - 2y(t)$, $T = 1$ et $n = 5$.

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 $n \longrightarrow +\infty$. 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 1.23, dans le cas général modulo quelques hypothèses supplémentaires. On ne va pas les donner ici. [61] 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.

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.

61. 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.

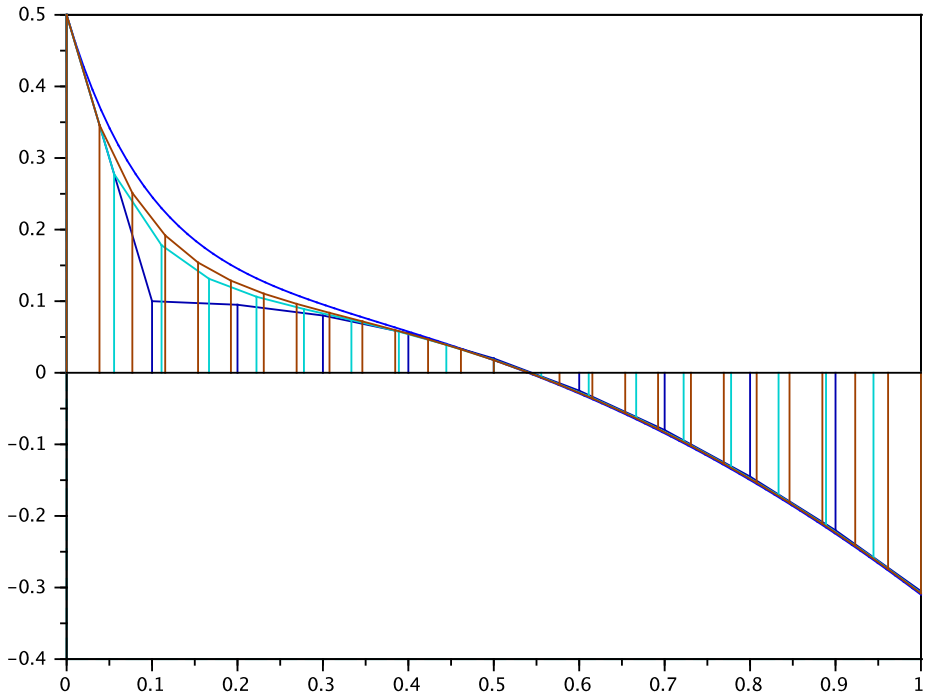


FIGURE 1.23 – Convergence des polygones d'Euler pour le problème de Cauchy $y'(t) = 1 - 5t^2 - 10y(t)$, $y(0) = 0,5$. On a tracé la solution exacte en bleu et trois polygones de couleurs différentes correspondant respectivement à $n = 10$ (bleu foncé), 18 (cyan) et 26 (brun).

1.5.2 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 1.5.1 *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 1.5.2 *Supposons que f possède des dérivées partielles par rapport à y_i, $i = 1, \dots, m$, 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

$$\begin{aligned} g'(s) &= \sum_{i=1}^{m} \frac{\partial f}{\partial y_i}(t, sy + (1 - s)z) \frac{d}{ds}(sy + (1 - s)z)_i \\ &= \sum_{i=1}^{m} \frac{\partial f}{\partial y_i}(t, sy + (1 - s)z)(y_i - z_i). \end{aligned}$$

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 1.5.1 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 1.5.2 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 1.5.3 *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 1.5.4 (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* (1.3.1).

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 1.5.5 *Toute fonction $f\colon [0,T] \times \mathbb{R}^m \to \mathbb{R}^m$ 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 1.5.4 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 1.4, *cf.* théorème 1.4.18, page 47. 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 [62] (ou une de ses versions les plus simples).

Proposition 1.5.6 *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 $t \in [0,T]$,

$$v(t) + \frac{\beta}{\alpha} \leq \left(v(0) + \frac{\beta}{\alpha}\right) e^{\alpha t}.$$

62. Thomas Hakon Grönwall, 1877–1932.

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 1.5.4. On notera de façon générique L la constante de Lipschitz de f par rapport à y.

Proposition 1.5.7 *Sous les hypothèses du Théorème* 1.5.4*, 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 1.4.20, mais nous la recopions quand même ici.

Soient y et $\tilde{y}$ deux solutions de (1.3.1). 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,

$$\big(f(t, y(t)) - f(t, \tilde{y}(t))\big|y(t) - \tilde{y}(t)\big) \leq \|f(t, y(t)) - f(t, \tilde{y}(t))\|\|y(t) - \tilde{y}(t)\| \leq L\|y(t) - \tilde{y}(t)\|^2.$$

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 1.5.8 (Cauchy-Lipschitz, unicité) *Sous les hypothèses du Théorème* 1.5.4*, le problème de Cauchy* (1.3.1) *a au plus une solution.*

Démonstration. On applique la proposition 1.5.7 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 1.5.7 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* [63]. 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 1.5.9 *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* (1.3.1)*, continue sur* $[0,T]$*. Alors on a pour tout* $t \in [0,T]$

$$y(t) = y_0 + \int_0^t f\big(s,y(s)\big)\,ds. \tag{1.5.2}$$

Réciproquement, soit y une fonction continue sur $[0,T]$ *à valeurs dans* $\mathbb{R}^m$ *et satisfaisant l'équation intégrale* (1.5.2)*. Alors, y est dérivable sur I et solution du problème de Cauchy* (1.3.1)*.*

Démonstration. La démonstration est strictement identique au cas linéaire, voir proposition 1.4.17. ◊

Les deux formulations, problème de Cauchy et équation intégrale, sont donc équivalentes.

Proposition 1.5.10 (Cauchy-Lipschitz, existence) *Sous les hypothèses du théorème* 1.5.4*, le problème de Cauchy* (1.3.1) *a au moins une solution.*

Démonstration. On utilise encore la méthode des approximations successives de Picard. [64] On définit donc une suite de fonctions continues $(y_n)_{n\in\mathbb{N}}$, en posant pour tout $t \in [0,T]$,

$$y_0(t) = y_0, \quad y_{n+1}(t) = y_0 + \int_0^t f(s,y_n(s))\,ds.$$

63. 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.

64. D'ailleurs, en anglais le théorème de Cauchy-Lipschitz est connu sous le nom de *Picard-Lindelöf theorem*. Ernst Leonard Lindelöf, 1870–1946.

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, [65] 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 $y : [0, T] \to \mathbb{R}^m$.

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 1.5.9, 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 1.5.8 et 1.5.10. 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 1.5.11 *i) Si $f : [t_0, T] \times \mathbb{R}^m \to \mathbb{R}^m$ 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 $f : [-T, 0] \times \mathbb{R}^m \to \mathbb{R}^m$, $T > 0$, 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 $y_0 \in \mathbb{R}^m$.

65. 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 [66], 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 [67], 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 1.5.12 *Sous les hypothèses du théorème* 1.5.4*, 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.

66. Si ce n'est pas clair, ne pas hésiter à le vérifier.
67. Si ce n'est pas clair, ...

Chapitre 2

Approximation numérique des équations différentielles ordinaires

2.1 Principes généraux

On a vu au paragraphe 1.5.1 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.

2.1.1 La notion de schéma numérique

On se donne un problème de Cauchy générique de la forme (1.3.1) 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 4.5.1.

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 (1.2.1) 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

$$y'(t_n) \approx \frac{y(t_{n+1}) - y(t_n)}{h} \tag{2.1.1}$$

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

$$y_{n+1} = y_n + hf(t_n, y_n), \tag{2.1.2}$$

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 (2.1.2) 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 [1]) 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 ?uvre du schéma, on dit que l'on a affaire à un schéma explicite.

(b) L'approximation

$$y'(t_{n+1}) \approx \frac{y(t_{n+1}) - y(t_n)}{h} \tag{2.1.3}$$

1. Voir remarque plus haut sur cet exposant implicite.

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

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}). \tag{2.1.4}$$

Cette fois-ci, pour calculer y_{n+1}, il faut résoudre une équation, en général non linéaire (sauf si $y \longrightarrow f(t, y)$ 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 (2.1.4) est appelée *schéma d'Euler implicite* ou rétrograde.

(c) L'approximation centrée

$$y'(t_n) \approx \frac{y(t_{n+1}) - y(t_{n-1})}{2h} \tag{2.1.5}$$

conduit au schéma

$$y_{n+1} = y_{n-1} + 2hf(t_n, y_n), \tag{2.1.6}$$

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 (2.1.1), (2.1.3) et (2.1.5) 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 (1.2.1) sur un intervalle $[t_n, t_{n+1}]$, on obtient

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} y'(s)\, ds = \int_{t_n}^{t_{n+1}} f\big(s, y(s)\big)\, ds. \tag{2.1.7}$$

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

$$I_n = \int_{t_n}^{t_{n+1}} f\big(s, y(s)\big)\, ds. \tag{2.1.8}$$

Évidemment, on ne sait pas calculer cette intégrale puisqu'on ne connait pas la solution exacte [2]. 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 2.5 pour l'interprétation 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 (2.1.8) par la méthode des rectangles à gauche (avec un seul rectangle).

On obtient donc

$$I_n \approx hf\big(t_n, y(t_n)\big).$$

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

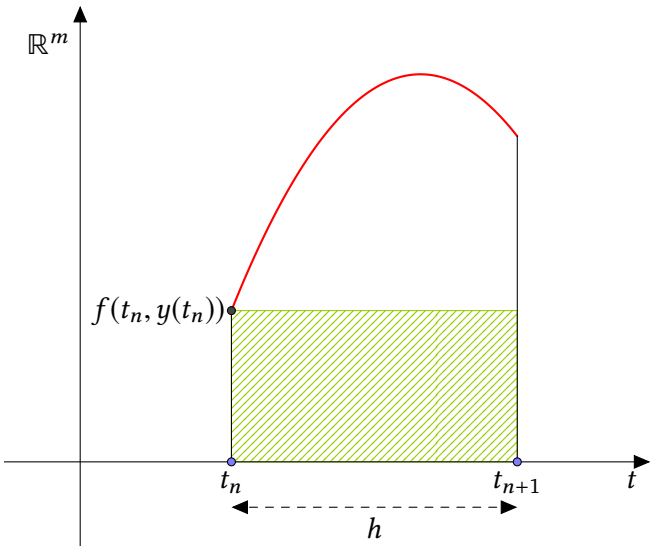


FIGURE 2.1 – À gauche, la fonction $t \mapsto f(t, y(t))$ étant tracée en rouge.

Remplaçant maintenant dans (2.1.7) et dans l'approximation précédente les valeurs exactes $y(t_n)$ par des valeurs que l'on espère approchées y_n, il vient

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

soit

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

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 2.5. 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

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

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 2.5, $y_n = h\sum_{k=1}^{n} g(t_k)$.

(c) Si l'on approche (2.1.8) par la méthode du point milieu, [3] on trouve l'approximation

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

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

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

3. 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.

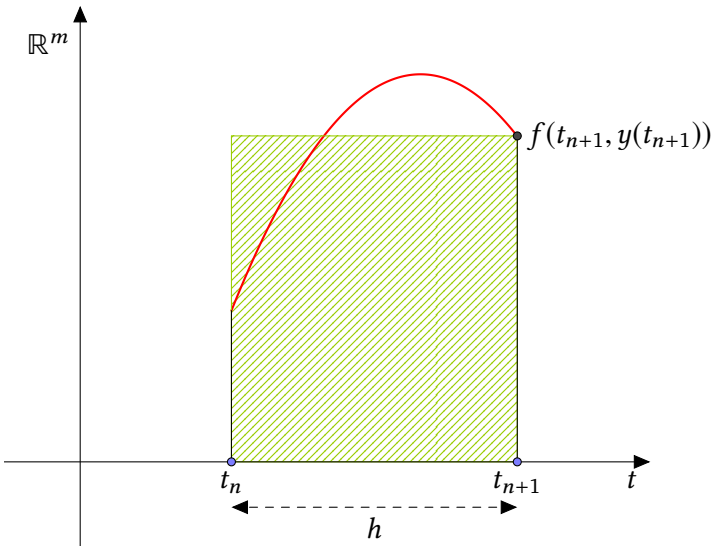


FIGURE 2.2 – À droite.

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

$$y_{n+1} = y_n + hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f(t_n, y_n)\right), \tag{2.1.9}$$

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 2.5, $y_n = h\sum_{k=0}^{n-1} g(t_k + \frac{h}{2})$.

(d) En approchant (2.1.8) par la méthode des trapèzes (avec un seul trapèze, cf. Figure 2.4), on obtient le schéma

$$y_{n+1} = y_n + \frac{h}{2}\big(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\big), \tag{2.1.10}$$

appelé *schéma de Crank-Nicolson* [4]. C'est un schéma implicite. Dans le cas où $f(t, y) = g(t)$ [...] la méthode des trapèzes composée [...] Figure 2.5, $y_n = h\big(\frac{1}{2}g(t_0) + \sum_{k=1}^{n-1} g(t_k) + \frac{1}{2}g(t_n)\big)$.

À chaque discrétisation de (2.1.8), on peut ainsi associer un schéma numérique pour le problème de Cauchy.

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. [5]

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

4. John Crank, 1916–2006 ; Phyllis Nicolson, 1917–1968.
5. La géométrie cachée derrière les systèmes hamiltoniens s'appelle la géométrie symplectique.

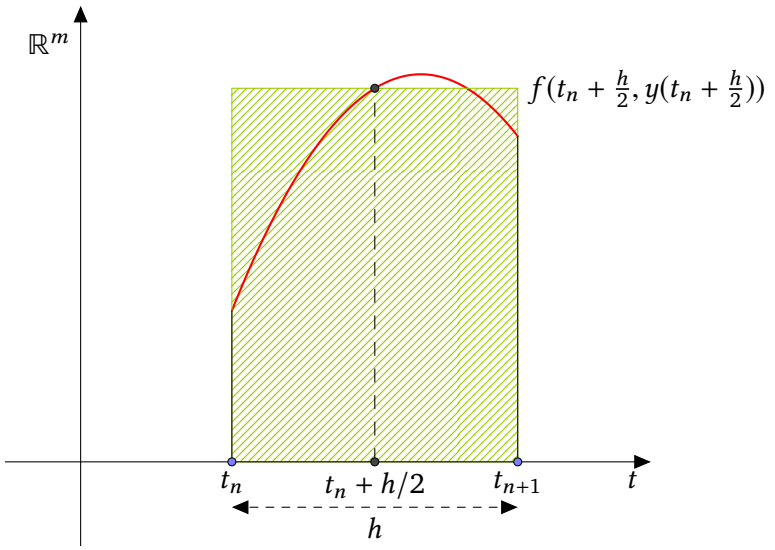


FIGURE 2.3 – Au milieu.

structure supplémentaire en question. C'est le cas pour les systèmes hamiltoniens. On rappelle que ces systèmes différentiels s'écrivent sous la forme générique

$$\begin{cases} \dot{q} = \dfrac{\partial H}{\partial p}, \\ \dot{p} = -\dfrac{\partial H}{\partial q}, \end{cases} \tag{2.1.11}$$

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. On a vu que ces systèmes avaient 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 d'une partie de $\mathbb{R}^k$ en la découpant en tous petits parallélotopes... [6] 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 (2.1.11) 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.

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

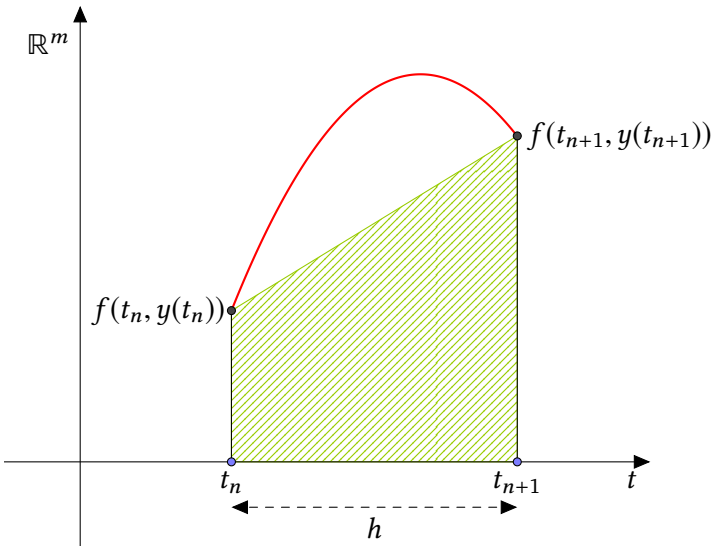


FIGURE 2.4 – Avec un trapèze.

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 1.5.1 montre que l'on converge bien vers ce qu'on veut. Pour le schéma d'Euler implicite, on trouve

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

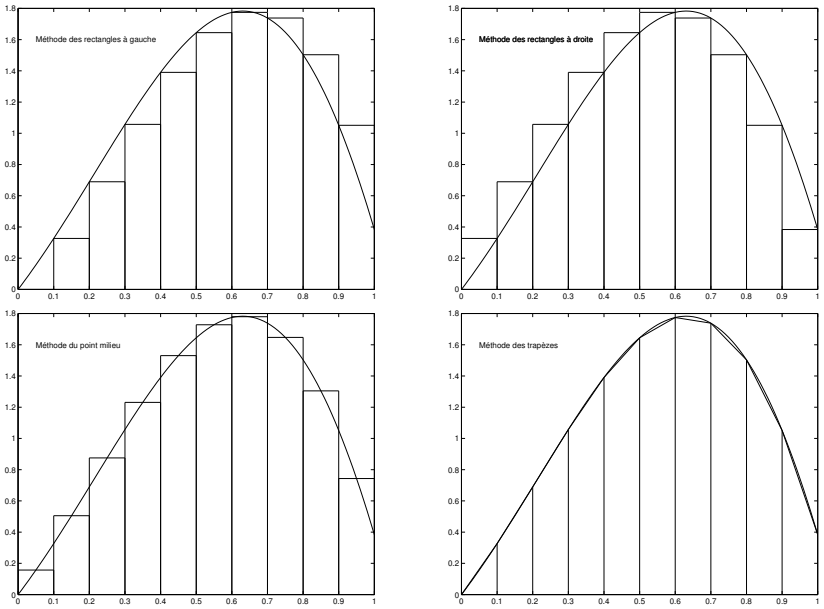


FIGURE 2.5 – Illustration des formules de quadratures utilisées dans le contexte de l'intégration numérique : on approche l'aire située sous la courbe par la somme des aires des rectangles (ou des trapèzes).

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

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

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 1.5.1 que l'on converge bien aussi dans ces trois derniers cas vers ce que l'on veut, voir Figure 2.6. [7]

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

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

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

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.

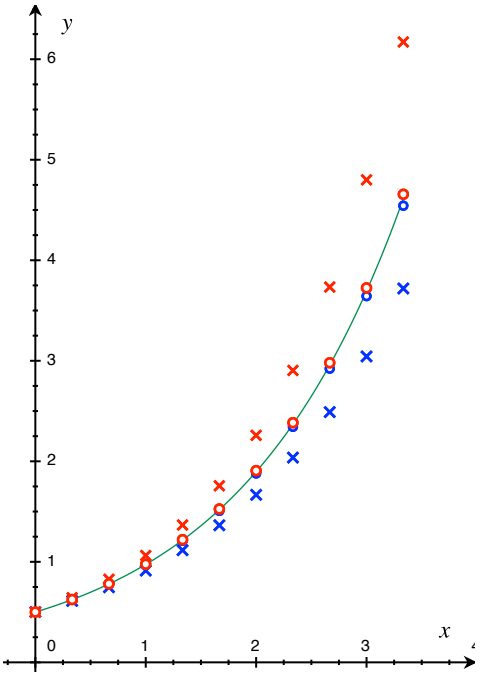


FIGURE 2.6 – × : Euler, × : Euler implicite, ○ : Euler modifiée, ○ : Crank-Nicolson, tous avec le même pas $h = \frac{1}{3}$. En vert, la solution exacte. On en ressort avec l'impression nette que Euler modifiée et Crank-Nicolson donnent de meilleurs résultats que Euler et Euler implicite. Intuition à confirmer plus tard.

2.1.2 Schémas numériques généraux

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

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

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 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}. [8] 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). [9] 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 $y_0, \dots y_q$. 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 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 maintenant 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.

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

9. 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.

2.1.3 Schémas explicites à un pas

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

$$y_{n+1} = y_n + hF(t_n, y_n, h), \tag{2.1.12}$$

où F est une fonction définie et continue [10] sur $[0,T]\times\mathbb{R}^m\times[0,1]$ à valeurs dans $\mathbb{R}^m$, que l'on sait écrire explicitement [11].

Par exemple, le schéma d'Euler correspond à $F(t,y,h) = f(t,y)$ et le schéma d'Euler modifié à $F(t,y,h) = f\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right)$. 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. [12] 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 (2.1.12) ? 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 [13], 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 $y(t)$ 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 $[0,1]$ [14]. 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 (2.1.12) 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 2.1.1. [15]

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

Définition 2.1.1 *Le schéma (2.1.12) est dit convergent si, pour toute donnée initiale $y(0)$ du problème continu,*

$$\lim_{\substack{h\to 0\\ y_0^N\to y(0)}} \sup_{0\le n\le N} \|y_n^N - y(t_n)\| = 0. \tag{2.1.13}$$

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

10. Donc continue par rapport au triplet (t,y,h), on l'avait notée Φ dans la section précédente.
11. 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.
12. 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.
13. Même h ne l'est pas, puisqu'il est de la forme T/N avec N entier.
14. Ou plus généralement $[0,h_0]$ pour un certain $h_0 > 0$.
15. 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.

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 2.1.1 Le problème de Cauchy

$$y'(t) = 6\big(y(t) - \varphi(t)\big) + \varphi'(t), \tag{2.1.14}$$

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 2.7 montre l'approximation obtenue par un

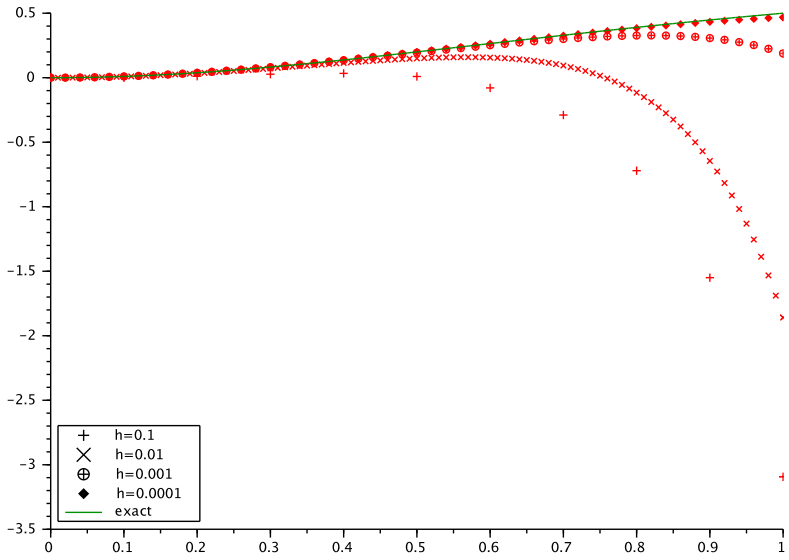


FIGURE 2.7 – Résolution de (2.1.14) avec le schéma d'Euler. On n'a pas tracé tous les points pour les deux plus petites valeurs de h (1 point sur 20 et 1 point sur 200 respectivement), il y en a trop et on ne voit plus rien.

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.

2.1.4 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 2.1.2 *Le schéma* (2.1.12) *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

$$\max_{0\le n\le N} \|z_n - y_n\| \le C \sum_{n=0}^{N} \|\eta_n\|. \tag{2.1.15}$$

La constante C est appelée *constante de stabilité* du schéma. [16] 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.

Proposition 2.1.3 *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*

$$\|F(t,y,h) - F(t,z,h)\| \le M\|y - z\|, \tag{2.1.16}$$

alors le schéma (2.1.12) *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 lemme qui nous sera très utile dans toute la suite du cours, et qui est le pendant discret du lemme de Grönwall 1.5.6, décliné en deux versions.

Lemme 2.1.4 (Lemme de Grönwall discret, version 1) *Soit $(u_n)_{n\ge 0}$ une suite de réels. On suppose qu'il existe deux réels λ et μ, avec $\lambda \ne 0$, $1+\lambda \ge 0$ et $u_0 + \frac{\mu}{\lambda} \ge 0$, tels que*

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

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

$$u_n + \frac{\mu}{\lambda} \le \left(u_0 + \frac{\mu}{\lambda}\right) e^{\lambda n}. \tag{2.1.17}$$

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

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

16. 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.

Lemme 2.1.5 (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

$$u_n \leq e^{\lambda n} u_0 + \sum_{k=0}^{n-1} e^{\lambda k} \mu_{n-k-1}. \tag{2.1.18}$$

Démonstration. Montrons par récurrence sur n que

$$u_n \leq (1+\lambda)^n u_0 + \sum_{k=0}^{n-1} (1+\lambda)^k \mu_{n-1-k}. \tag{2.1.19}$$

L'inégalité (2.1.19) 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 (2.1.18) devient

$$u_n \leq e^{\lambda n} u_0 + \mu \frac{e^{\lambda n} - 1}{e^{\lambda} - 1}, \tag{2.1.20}$$

majoration plus fine que (2.1.17) 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 2.1.3. Pour deux suites $(y_n)_n$ et $(z_n)_n$ telles que celles définies à la définition 2.1.2, 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 (2.1.16), 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 2.1.5 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, (2.1.15), 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 2.1.2 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 2.1.3 Pour le schéma d'Euler modifié, et toujours pour une fonction f globalement lipschitzienne par rapport à y, uniformément par rapport à t, [17] 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.

2.1.5 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 2.1.6 *On appelle* erreur de consistance *(ou erreur de discrétisation locale) du schéma* (2.1.12) *la quantité $\varepsilon_n \in \mathbb{R}^m$ définie par*

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

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

où y est une solution de l'EDO. [18]

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 2.8. 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.

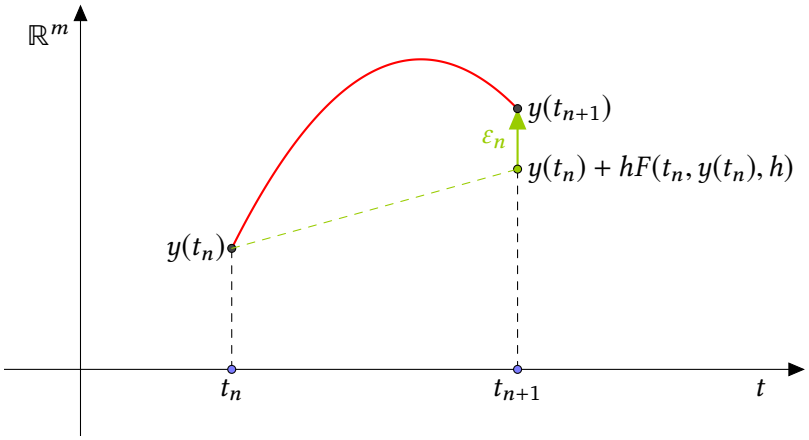


FIGURE 2.8 – Erreur de discrétisation locale, a.k.a. de consistance.

Définition 2.1.7 *Le schéma* (2.1.12) *est dit* consistant *avec l'EDO* (1.2.1) *si pour toute solution y de* (1.2.1)*, 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

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

l'éradiquer... [19] [20] La consistance, c'est quand même plus adapté à la bouillie.

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.

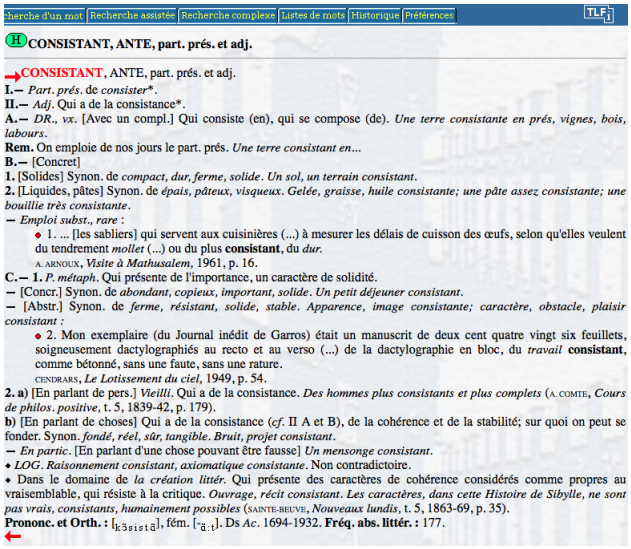
Proposition 2.1.8 *On suppose l'application F continue sur $[0,T]\times\mathbb{R}^m\times[0,1]$. Le schéma (2.1.12) est consistant si et seulement si, pour tout $(t,y)\in I\times\mathbb{R}^m$, on a*

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

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

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

19. À ce propos, voir le TLFi : http://atilf.atilf.fr/tlf.htm

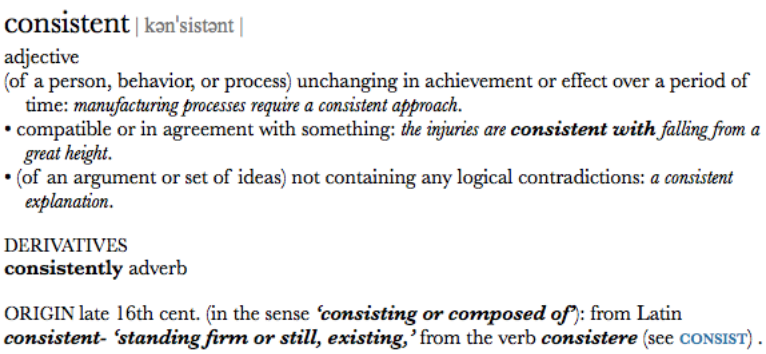


herche d'un mot | Recherche assistée | Recherche complexe | Listes de mots | Historique | Préférences | TLFi

H **CONSISTANT, ANTE, part. prés. et adj.**

→**CONSISTANT**, ANTE, part. prés. et adj.
I.— *Part. prés.* de *consister**.
II.— *Adj.* Qui a de la consistance*.
A.— *DR., vx.* [Avec un compl.] Qui consiste (en), qui se compose (de). *Une terre consistante en prés, vignes, bois, labours.*
Rem. On emploie de nos jours le part. prés. *Une terre consistant en...*
B.— [Concret]
1. [Solides] Synon. de *compact, dur, ferme, solide. Un sol, un terrain consistant.*
2. [Liquides, pâtes] Synon. de *épais, pâteux, visqueux. Gelée, graisse, huile consistante; une pâte assez consistante; une bouillie très consistante.*
— *Emploi subst., rare* :
◆ 1. ... [les sabliers] qui servent aux cuisinières (...) à mesurer les délais de cuisson des œufs, selon qu'elles veulent du tendrement *mollet* (...) ou du plus **consistant**, du *dur.*
A. ARNOUX, *Visite à Mathusalem*, 1961, p. 16.
C.— **1.** *P. métaph.* Qui présente de l'importance, un caractère de solidité.
— [Concr.] Synon. de *abondant, copieux, important, solide. Un petit déjeuner consistant.*
— [Abstr.] Synon. de *ferme, résistant, solide, stable. Apparence, image consistante; caractère, obstacle, plaisir consistant :*
◆ 2. Mon exemplaire (du Journal inédit de Garros) était un manuscrit de deux cent quatre vingt six feuillets, soigneusement dactylographiés au recto et au verso (...) de la dactylographie en bloc, du *travail* **consistant**, comme bétonné, sans une faute, sans une rature.
CENDRARS, *Le Lotissement du ciel*, 1949, p. 54.
2. a) [En parlant de pers.] *Vieilli.* Qui a de la consistance. *Des hommes plus consistants et plus complets* (A. COMTE, *Cours de philos. positive*, t. 5, 1839-42, p. 179).
b) [En parlant de choses] Qui a de la consistance (*cf.* II A et B), de la cohérence et de la stabilité; sur quoi on peut se fonder. Synon. *fondé, réel, sûr, tangible. Bruit, projet consistant.*
— *En partic.* [En parlant d'une chose pouvant être fausse] *Un mensonge consistant.*
◆ *LOG. Raisonnement consistant, axiomatique consistante.* Non contradictoire.
◆ Dans le domaine de *la création littér.* Qui présente des caractères de cohérence considérés comme propres au vraisemblable, qui résiste à la critique. *Ouvrage, récit consistant. Les caractères, dans cette Histoire de Sibylle, ne sont pas vrais, consistants, humainement possibles* (SAINTE-BEUVE, *Nouveaux lundis*, t. 5, 1863-69, p. 35).
Prononc. et Orth. : [kɔ̃sistɑ̃], fém. [-ɑ̃t]. Ds *Ac.* 1694-1932. **Fréq. abs. littér. :** 177.
←

20. Alors qu'en anglais, cela va très bien :



consistent | kən'sistənt |

adjective
(of a person, behavior, or process) unchanging in achievement or effect over a period of time: *manufacturing processes require a consistent approach.*
• compatible or in agreement with something: *the injuries are **consistent with** falling from a great height.*
• (of an argument or set of ideas) not containing any logical contradictions: *a consistent explanation.*

DERIVATIVES
consistently adverb

ORIGIN late 16th cent. (in the sense ***'consisting or composed of'***): from Latin ***consistent-*** ***'standing firm or still, existing,'*** from the verb ***consistere*** (see CONSIST).

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

$$\varepsilon_n = h\int_0^1 f\big(t_n + sh, y(t_n + sh)\big)\,ds - hF(t_n, y(t_n), h)$$
$$= h\int_0^1 \big(F\big(t_n + sh, y(t_n + sh), 0\big) - F(t_n, y(t_n), h)\big)\,ds,$$

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

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

Soit $K = \{(t,s,h) \in [0,T]\times[0,1]\times[0,1];\, t+h \le T\}$. 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 $h \le h_0$, $G(t,s,h) = |G(t,s,h) - G(t,s,0)| \le \eta$ pour tous t et s. [21] Il s'ensuit que pour $h \le h_0$, on a

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

d'où

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

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 $t + \frac{0}{2} \le T$, donc $F(t,y,0) = f\big(t + \frac{0}{2}, y + \frac{0}{2}f(t,y)\big) = f(t,y)$.

2.1.6 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 [22], quoique probablement pas dû à Lax dans le contexte des EDO.

Théorème 2.1.9 *Si le schéma (2.1.12) est stable et consistant alors il est convergent.*

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

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

21. 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 $|t_1 - t_2| + |s_1 - s_2| + |h_1 - h_2| \le h_0$ alors $|G(t_1,s_1,h_1) - G(t_2,s_2,h_2| \le \eta$. On prend ici $t_1 = t_2 = t$, $s_1 = s_2 = s$, $h_1 = h$ et $h_2 = 0$.

22. Peter David Lax, 1926–

par définition de l'erreur de consistance du schéma ε_n. On a donc affaire à une suite perturbée au sens de la définition 2.1.2 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

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

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 2.1.1. ◊

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. [23]

2.1.7 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 2.1.10 *Le schéma (2.1.12) est dit d'ordre au moins $p \in \mathbb{N}^*$ si, pour toute solution y de l'EDO (1.2.1), il existe une constante C indépendante de h telle que*

$$\forall N, \quad \sum_{n=0}^{N-1} \|\varepsilon_n\| \le Ch^p. \tag{2.1.21}$$

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 2.1.11 *On suppose le schéma (2.1.12) stable et d'ordre $p \ge 1$ et qu'il existe une constante $C > 0$ telle que $\|y_0 - y(0)\| \le Ch^p$. Alors il existe $\tilde{C} > 0$ telle qu'on ait l'estimation d'erreur suivante*

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

Démonstration. Reprendre la démonstration du théorème 2.1.9 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)

◊

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

Remarque 2.1.1 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* 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 2.1.12 *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

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

Supposons que la solution y considérée est de classe C^2. [24] On a donc, par l'inégalité de Taylor-Lagrange [25]

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

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, [26] cette fois-ci par rapport à h en $h = 0$, en supposant f de classe C^2, ce qui donne

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

Or on a (voir aussi la proposition 2.1.13)

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

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

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

24. On reviendra plus loin sur ces questions de régularité de la solution, voir proposition 2.1.13.
25. Joseph Louis, comte de Lagrange (en italien Giuseppe Lodovico de Lagrangia), 1736–1813.
26. En effet, inutile d'aller plus loin, puisque ce terme va être multiplié par h dans l'erreur de consistance.

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, [27] 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 2.1.14. 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. [28]

Remarque 2.1.2 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 2.1.3 On a vu au paragraphe **??** une illustration numérique de l'ordre du schéma d'Euler. 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 2.9 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 $O(h^2)$.

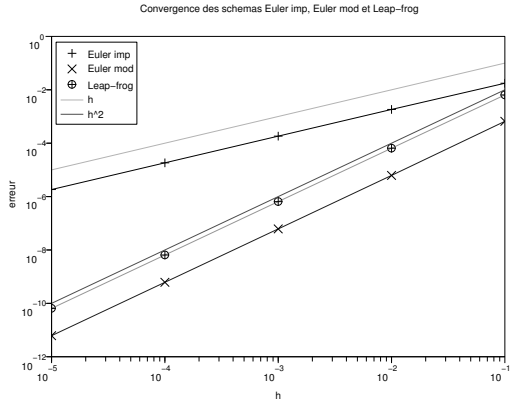


FIGURE 2.9 – Convergence des schémas d'Euler implicite, Euler modifié et leapfrog sur l'EDO $y'(t) = -y(t)$. Erreur estimée numériquement et allures théorique en $O(h)$ et $O(h^2)$.

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.

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

Proposition 2.1.13 *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 $0 \leq k \leq p$,*

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

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

$$\forall (t,y) \in [0,T] \times \mathbb{R}^m, \quad \begin{cases} f^0(t,y) = f(t,y), \\ f^{k+1}(t,y) = \dfrac{\partial f^k}{\partial t}(t,y) + \displaystyle\sum_{j=1}^{m} \frac{\partial f^k}{\partial y_j}(t,y) f_j(t,y), \ \textit{pour } 0 \leq k \leq p-1. \end{cases} \tag{2.1.22}$$

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 $k = 0, \ldots, p$, 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 1.5.3. De plus, on a $f^0 = f$ de classe C^{p-0}.

Supposons donc maintenant y de classe C^{k+1} pour un certain $k \leq p-1$ avec $y^{(k+1)}(t) = f^k(t, y(t))$ et f^k de classe C^{p-k}. Comme $p - k \geq 1$, 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, $i = 1, \ldots, m$,

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

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 2.1.14 *On suppose que F est de classe C^p. Le schéma (2.1.12) est d'ordre au moins p si et seulement si pour tout $k = 0,\dots,p$, on a*

$$\frac{\partial^k F}{\partial h^k}(t,y,0) = \frac{1}{k+1} f^k(t,y), \tag{2.1.23}$$

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 (2.1.23) 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 2.1.13. La condition (2.1.23) implique donc que

$$\frac{\partial^k F}{\partial h^k}(t,y(t),0) = \frac{1}{k+1} y^{(k+1)}(t), \tag{2.1.24}$$

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 (2.1.24), 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

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

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

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. ◊

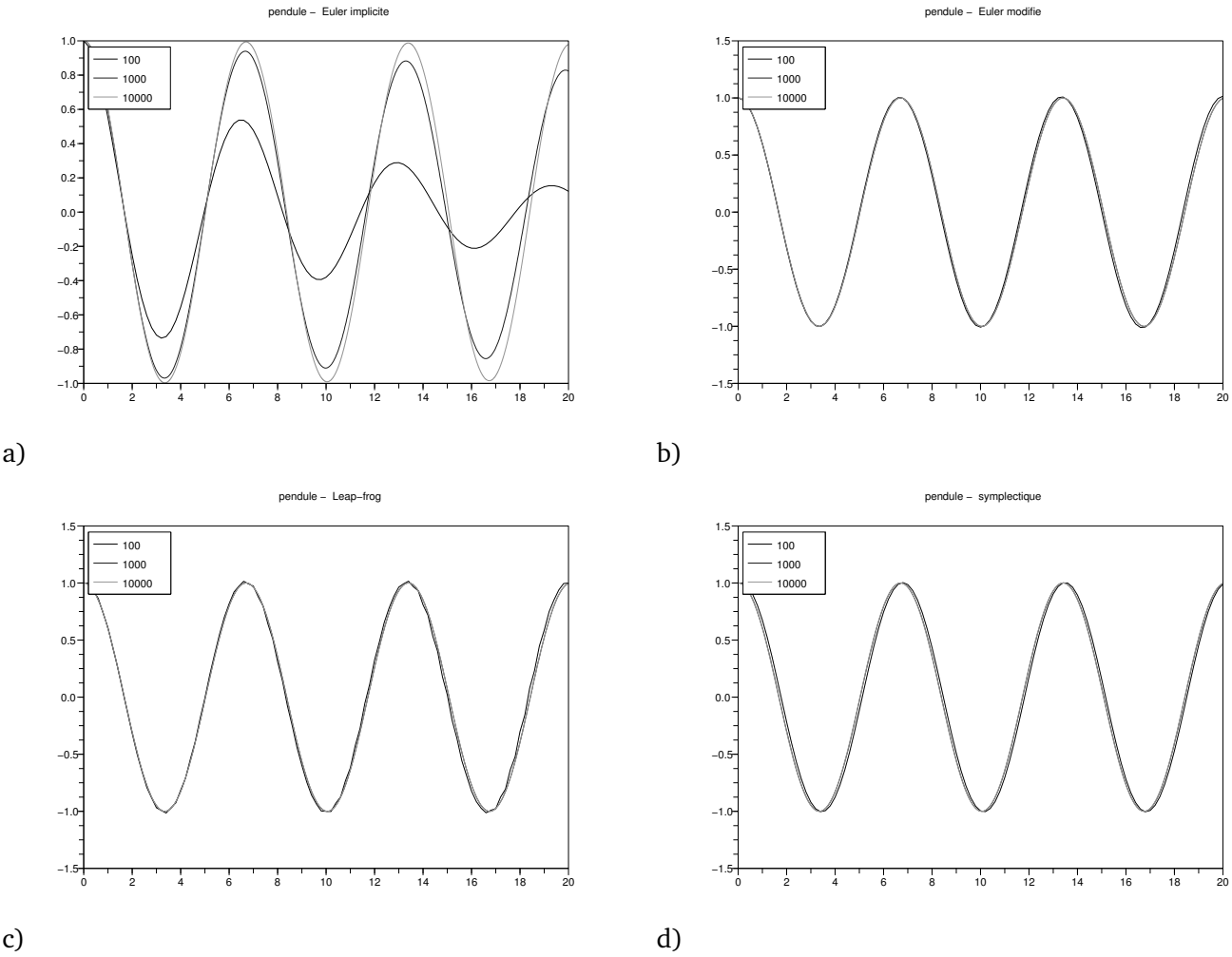


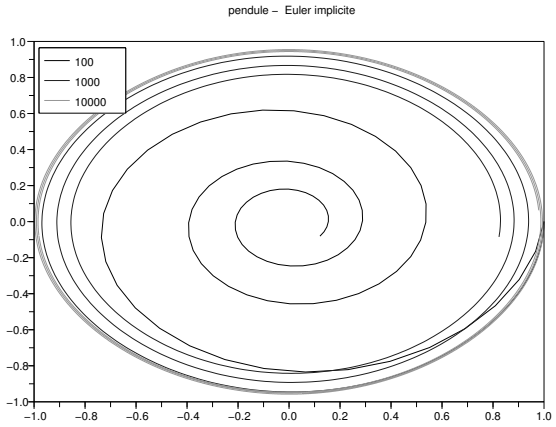
FIGURE 2.10 – Variations de l'inclinaison du pendule en fonction du temps pour trois discrétisations différentes et les schémas a) d'Euler implicite, b) Euler modifié, c) leapfrog d) symplectique.

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.

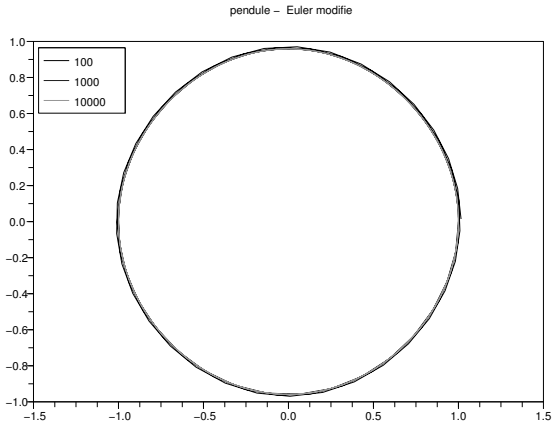
Pour conclure ce paragraphe, reprenons de nouveau notre exemple favori [29] et utilisons-le pour tester les trois schémas de la Figure 2.9, plus le schéma symplectique. La Figure 2.10 montre les variations de l'inclinaison en fonction du temps pour trois discrétisations différentes et chacun des algorithmes. La Figure 2.11 montre les trajectoires dans l'espace des phases et la Figure 2.12 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.

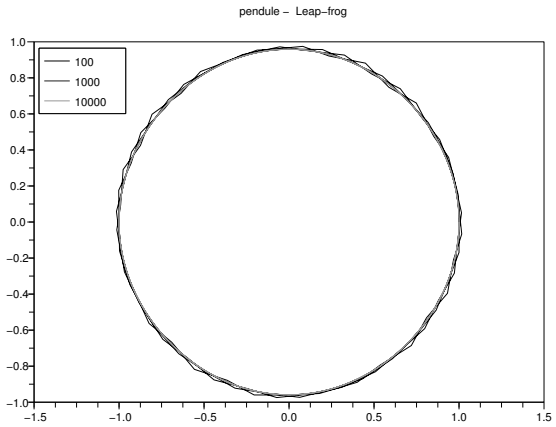
29. Celui du pendule bien sûr !



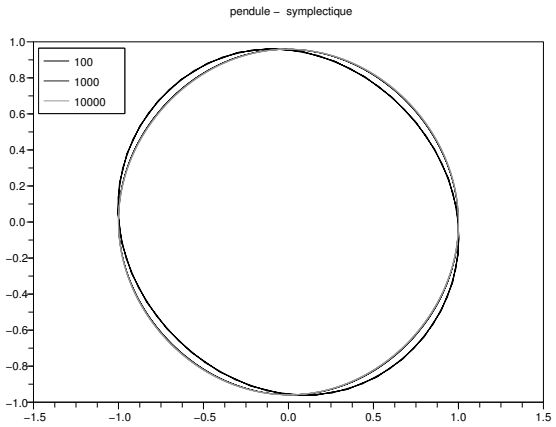
a)



b)

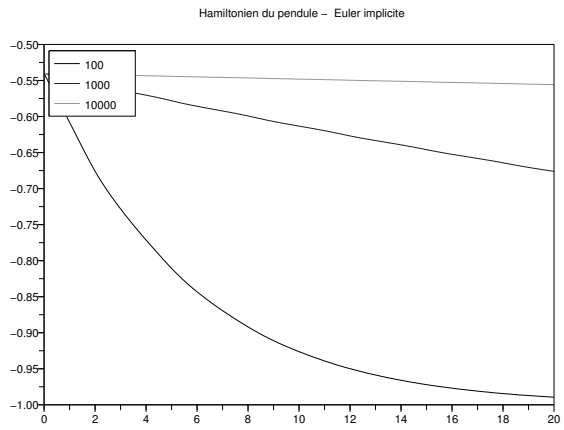


c)

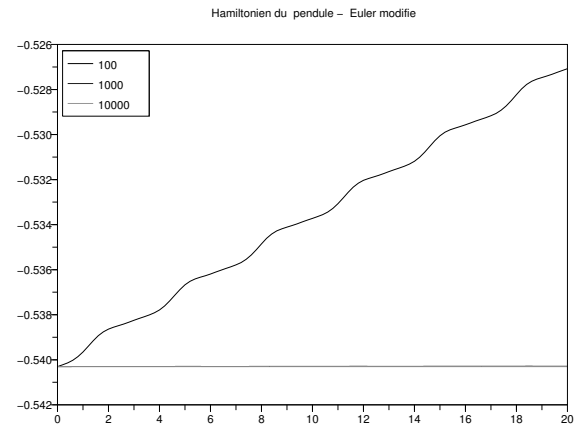


d)

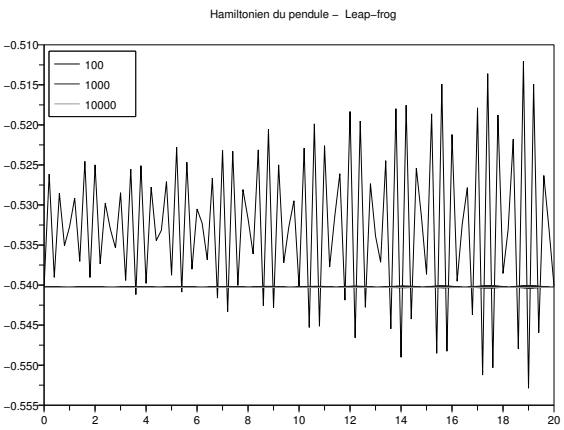
FIGURE 2.11 – Trajectoires du pendule dans l'espace des phases pour trois discrétisations différentes et les schémas a) d'Euler implicite, b) Euler modifié, c) leapfrog, d) symplectique.



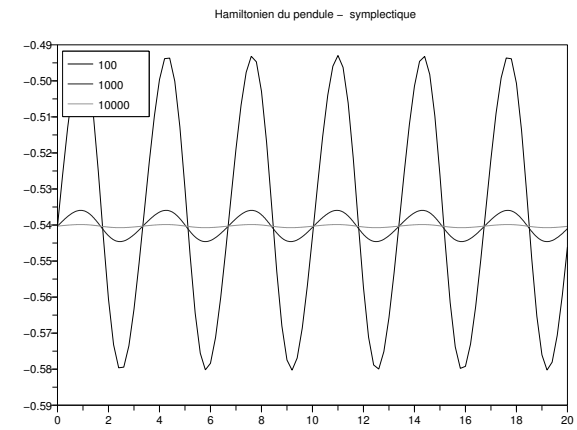
a)



b)



c)



d)

FIGURE 2.12 – 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.

Chapitre 3

Retour à l'étude théorique

Nous revenons maintenant à l'étude théorique des EDO. On se souvient qu'on a obtenu dans le premier chapitre l'existence et l'unicité d'une solution dans le cas où la fonction second membre du problème de Cauchy est globalement lipschitzienne. Nous affaiblissons maintenant cette hypothèse au cas où f est seulement *localement lipschitzienne* et introduisons les notions de solutions locales et maximales. Enfin dans le paragraphe 3.1.1, nous utilisons la notion de fonction de Liapounov pour obtenir des résultats d'existence globale pour des cas particuliers d'équations différentielles de type gradient et hamiltonien.

3.1 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.

Plus précisément, 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 3.1.1 *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 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 3.1.2 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.

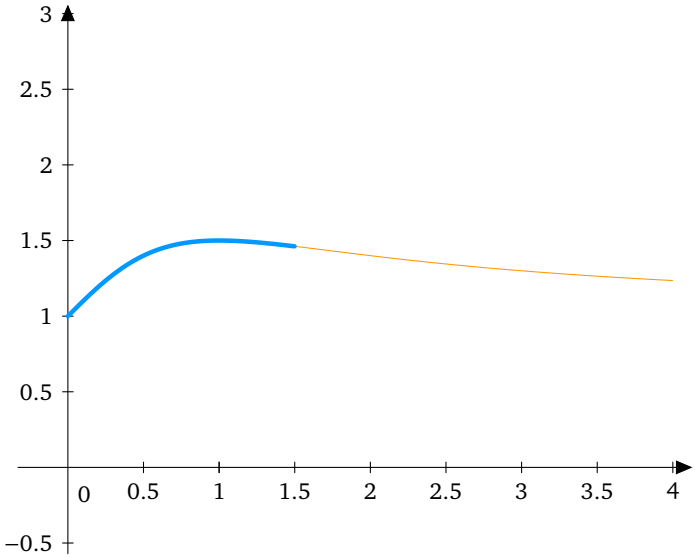


FIGURE 3.1 – La solution orange sur $[0, 4]$ prolonge strictement la solution bleue sur $\left[0, \frac{3}{2}\right]$.

La notion d'unicité d'une solution locale est précisée par la définition suivante.

Définition 3.1.3 *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 3.5 plus loin pour un exemple de non unicité locale.

Proposition 3.1.4 *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 3.1.5 *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 3.1.5, une solution locale (I,y) telle que y est bornée sur $I=[0,\tau[$ avec $\tau<T$ n'est certainement pas maximale.

Exemple 3.1.1 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 3.1.6 *On dit que $f\colon [0,T[\times V\to\mathbb{R}^m$ 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 $t\in[0,\tau]$,*

$$\|f(t,y)-f(t,z)\|\leq L\|y-z\|.$$

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 3.1.7 (Cauchy-Lipschitz local) *Soit $f\colon [0,T[\times V\to\mathbb{R}^m$ 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 3.1.8 *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 3.1.7. 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 1.5.4, 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 $0<\tau_*\leq\tau$ 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 3.1.3. Soient $([0,\tau_1[,y_1)$ et $([0,\tau_2[,y_2)$ deux solutions locales avec $\tau_2\geq\tau_1$. Montrons que $y_{2|[0,\tau_1[}=y_1$. En effet, soit

$$\sigma=\sup\{s<\tau_1, y_1(t)=y_2(t)\text{ pour tout }0\leq t\leq s\}.$$

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 [1] $[\sigma,\sigma+\alpha[$ 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. ◊

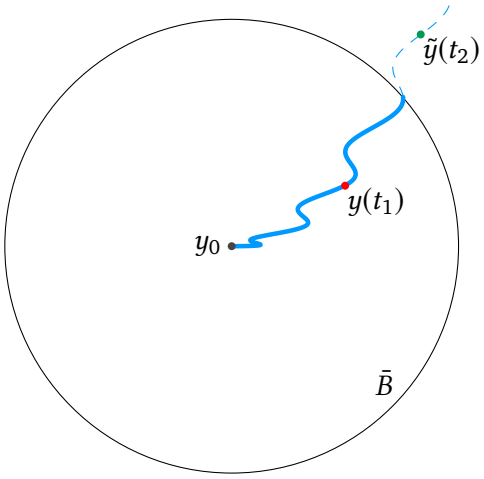


FIGURE 3.2 – Tant que la solution du problème de Cauchy prolongé reste dans la boule où $\tilde{f}$ et f coïncident, c'est-à-dire pour $t_1 \leq \tau_*$, c'est une solution locale du problème de Cauchy de départ. Peu importe ce qu'elle fait ensuite pour $t_2 > \tau_*$, on la laisse vivre sa vie en pointillés.

Au vu de cette démonstration, il est bien clair que bon nombre de propriétés globales, comme les corollaires 1.5.7 et 1.5.12, 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.

Démonstration du lemme 3.1.8. 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

$$\forall (t,y) \in [0,\tau]\times\mathbb{R}^m, \quad \tilde{g}(t,y) = \inf_{u\in\bar{B}} (g(t,u) + L\|y-u\|). \tag{3.1.1}$$

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 (3.1.1), 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 1.5.5 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 $t' \in [0,\tau]$,

$$\tilde{g}(t',y) - \tilde{g}(t,y) \leq g(t',v) + L\|y-v\| - g(t,v) - L\|y-v\| \leq |g(t',v) - g(t,v)|.$$

1. On applique l'existence à partir de l'instant initial σ plutôt que 0, ce qui n'est clairement pas un problème.

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 (3.1.1) s'appelle *prolongement de McShane-Whitney* [2] quand g ne dépend pas de t, voir figure 3.3. [3] Notons que l'on peut donner une démonstration de la continuité du prolongement par rapport à 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 $v_{n_{pq}}\to w$ quand $q\to+\infty$ pour un certain $w\in\bar{B}$. Comme $(t,v_{n_{pq}})\to(t,w)$ et $(t_{n_{pq}},v_{n_{pq}})\to(t,w)$ 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 $\tilde{g}(t,y)$.

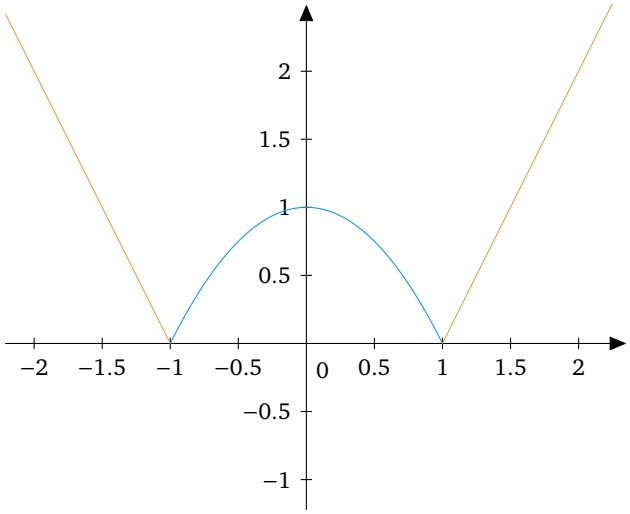


FIGURE 3.3 – Le prolongement de McShane-Whitney de la fonction $y\mapsto 1-y^2$ à l'extérieur de la boule de dimension un, $\bar{B}=[-1,1]$.

Pour pouvoir appliquer le théorème de Cauchy-Lipschitz local, nous avons également besoin de moyens pratiques d'en vérifier les hypothèses. On 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 $(t,y)\in[0,\tau]\times\bar{B}$.

2. Edward James McShane, 1904 –1989 ; Hassler Whitney, 1907–1989.

3. 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.

Proposition 3.1.9 *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 1.5.2. 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 3.1.9 donne une condition suffisante pour pouvoir appliquer le théorème 3.1.7 de Cauchy-Lipschitz local. Ce n'est pas du tout une condition nécessaire. On a une condition suffisante encore plus facile à vérifier.

Proposition 3.1.10 *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 3.1.9. ◊

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 $\mathbb{R}^m$.

Exemple 3.1.2 Soit le problème de Cauchy

$$\begin{cases} y'(t)=y(t)^2,\\ y(t_0)=y_0. \end{cases}$$

La fonction $f(t,y)=y^2$ est de classe C^1 donc localement lipschitzienne sur $\mathbb{R}$ d'après ce qui précède (pas de dépendance en t ici). On sait par conséquent qu'il existe une unique solution locale dans un voisinage de t_0. Par contre, la fonction f n'est pas globalement lipschitzienne sur $\mathbb{R}$, sa dérivée n'étant pas bornée. L'EDO est autonome, on peut la résoudre par séparation des variables, et c'est l'occasion de revenir sur cette méthode « physicienne » à la lumière du théorème de Cauchy-Lipschitz.

On l'a dit tout au début, la séparation des variables ne constitue pas seule une preuve convaincante, car on commence par diviser potentiellement par zéro en écrivant $\frac{y'}{y^2}=1$, ce qui est mal, on en

conviendra. Montrons qu'une telle division par zéro ne peut pas se produire si $y_0 \neq 0$. Tout d'abord, par Cauchy-Lipschitz, il existe une solution locale et une seule, y. Cette solution ne peut pas s'annuler sur son intervalle d'existence $[t_0, \tau[$. En effet, s'il existe $\tau_0 \in [t_0, \tau[$ tel que $y(\tau_0) = 0$, il existe donc $\tau_1 \in [t_0, \tau[$ tel que $\tau_1 = \inf\{t \in [t_0, \tau[; y(t) = 0\}$. Par continuité de la fonction y, on a $y(\tau_1) = 0$. Comme $y(t_0) = y_0 \neq 0$, on a donc $\tau_1 > t_0$, donc $y(t) \neq 0$ pour tout $t \in [t_0, \tau_1[$. Mais le problème de Cauchy

$$\begin{cases} z'(s) = -z(s)^2, \\ z(0) = 0, \end{cases}$$

a une solution et une seule $z(s) = 0$ sur $[0, \alpha[$, pour un certain $\alpha > 0$ que l'on peut choisir tel que $\tau_1 - \alpha \geq t_0$, toujours par Cauchy-Lipschitz local. Or on vérifie facilement que $z(s) = y(\tau_1 - s)$ est solution de ce problème de Cauchy, c'est donc la solution nulle par unicité de Cauchy-Lipschitz local. On en déduit que $0 = z(\alpha) = y(\tau_1 - \alpha)$, avec $t_0 \leq \tau_1 - \alpha < \tau_1$, ce qui contredit le fait que $y(t) \neq 0$ pour tout $t \in [t_0, \tau_1[$. On reconnaît l'argument de renversement du temps utilisé au corollaire 1.5.12. [4]

Les physiciens ont donc raison de diviser allègrement par y, puisque ce dernier ne s'annule jamais à cause du théorème de Cauchy-Lipschitz. De l'égalité

$$\frac{d}{dt}\left(-\frac{1}{y(t)}\right) = \frac{y'(t)}{y^2(t)} = 1,$$

on déduit donc par intégration des deux membres

$$\begin{aligned} \left[-\frac{1}{y(s)}\right]_{t_0}^{t} &= t - t_0, \\ \frac{1}{y_0} - \frac{1}{y(t)} &= t - t_0, \\ y(t) &= \frac{y_0}{1 - y_0(t - t_0)}. \end{aligned}$$

Donc si $y_0 > 0$, l'unique solution du problème de Cauchy n'est définie que sur $[t_0, t_0 + y_0^{-1}[$, elle n'est pas globale (mais prolongeable jusqu'à $-\infty$ pour $t < t_0$). Si $y_0 < 0$, l'unique solution est par contre globale (mais seulement prolongeable jusqu'à $t_0 + y_0^{-1}$ pour $t < t_0$). Et si $y_0 = 0$, que se passe-t-il ?

Voir la figure 3.4 pour l'application du prolongement de McShane-Whitney dans ce cas particulier. ◊

Dans le même ordre d'idées, reprenons l'équation à variables séparées $y'(t) = g(t)h(y(t))$ et la méthode physicienne toujours sous l'éclairage du théorème de Cauchy-Lipschitz. La fonction $(t, y) \mapsto f(t, y) = g(t)h(y)$ est telle que $|f(t, y) - f(t, z)| = |g(t)||h(y) - h(z)|$, elle n'est donc continue et localement lipschitzienne par rapport à y uniformément par rapport à t que si g est continue et h est localement lipschitzienne (sauf si $g = 0$ ou $h = 0$ qui ne sont pas des cas très intéressants). Faisons maintenant ces hypothèses sur $\mathbb{R}$ tout entier pour simplifier, le problème de Cauchy admet donc une solution locale pour toute donnée initiale y_0.

Cette solution est-elle donnée par la méthode physicienne ? Tout d'abord, g est continue, elle admet donc une primitive G sur $\mathbb{R}$. Si h ne s'annule pas, alors $\frac{1}{h}$ est également continue et admet aussi une primitive R sur $\mathbb{R}$ tout entier. Par contre, si h s'annule en au moins un point de $\mathbb{R}$, comme c'était le cas pour $h(y) = y^2$, on est bien ennuyés. Mais le même raisonnement d'unicité

4. On peut aussi raisonner plus brutalement comme suit : par Cauchy-Lipschitz, il existe une solution locale et une seule. On se cache alors pour faire le calcul de séparation des variables sans que personne ne voie ce que l'on manigance, et l'on n'en exhibe que le résultat, dont on vérifie directement à la main que c'est bien une solution locale du problème de Cauchy. C'est donc la bonne par unicité de Cauchy-Lipschitz. Pas très beau mais logiquement inattaquable.

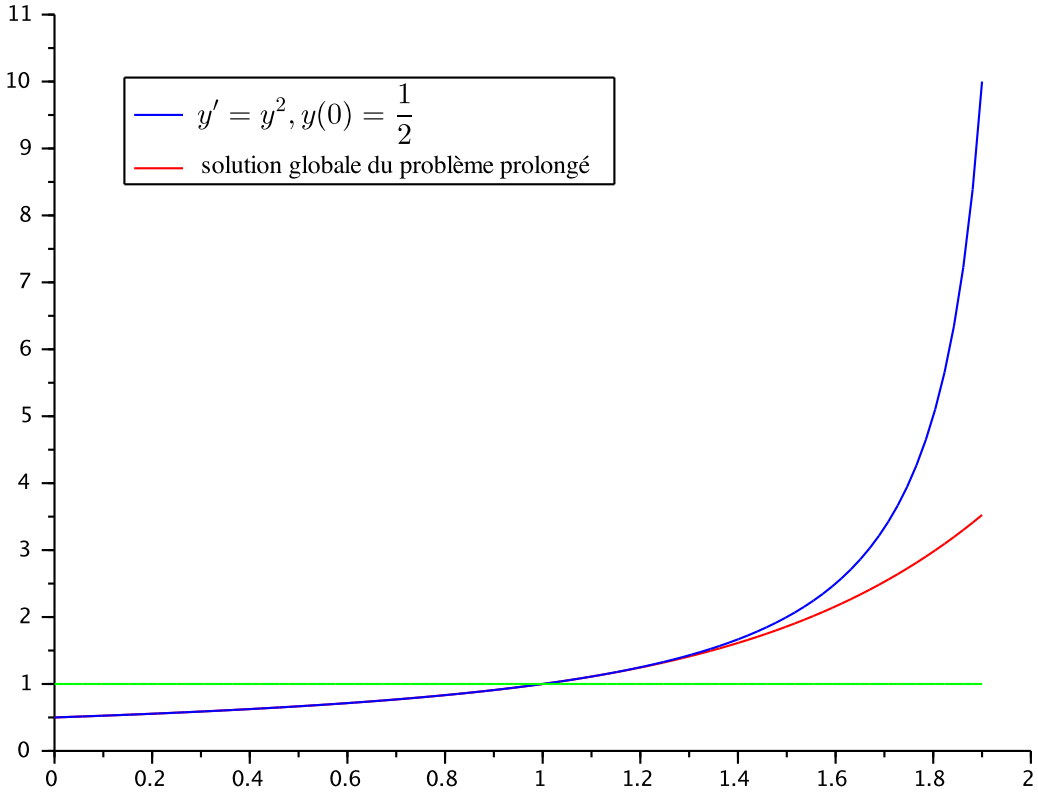


FIGURE 3.4 – On a pris ici le problème de Cauchy $y' = y^2$, $y(0) = \frac{1}{2}$, dont la solution explose à $t = 2$. Le prolongement de McShane-Whitney de la fonction $y \mapsto y^2$ en dehors de $[-1, 1]$ vaut $y \mapsto 2|y| - 1$. On a calculé les solutions des deux problèmes de Cauchy, qui coïncident jusqu'à $\tau_* = 1$, avec la fonction ode de scilab, bien que celles-ci s'expriment analytiquement ici.

par Cauchy-Lipschitz montre que $h(y(t))$ est soit identiquement nul, si $h(y_0) = 0$, soit ne s'annule jamais si $h(y_0) \neq 0$. Donc quand $h(y_0) \neq 0$, la solution locale — qui existe, ce n'est pas la question — se débrouille pour ne jamais annuler h tant qu'elle existe, c'est-à-dire qu'elle prend ses valeurs dans un intervalle où h ne s'annule pas, plus précisément, le plus grand intervalle J contenant y_0 et où h ne s'annule pas [5]. Comme $\frac{1}{h}$ est continue sur J, elle y admet donc une primitive R_J et l'on a bien pour tout temps t d'existence $\big(R_J(y(t))\big)' = \frac{y'(t)}{h(y(t))}$ et l'on peut intégrer sans problème pour trouver $R_J(y(t)) = R_J(y_0) + G(t) - G(0)$. De plus comme h est continue, $\frac{1}{h}$ garde un signe constant sur J. Il s'ensuit que R_J, application continue, y est soit strictement croissante, soit strictement décroissante, et dans les deux cas y admet une application réciproque R_J^{-1}, bien sûr dérivable. La formule $y(t) = R_J^{-1}\big(R_J(y_0) + G(t) - G(0)\big)$ est donc correcte et la physique avait une fois de plus raison contre toute attente.

Nous pouvons maintenant assez facilement démontrer les propositions 3.1.4 et 3.1.5, 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 3.1.11 *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 $k \to +\infty$.*

5. Par le théorème des valeurs intermédiaires.

Démonstration. Soit $([0,\tau[,y)$ une solution locale fournie par le théorème de Cauchy-Lipschitz local 3.1.7. 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 3.1.12 (Peano) *On suppose la fonction f continue au voisinage du point $(0,y_0)$. Alors le problème de Cauchy* (1.3.1) *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 3.1.3 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 [6] 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 3.5. ◊

Il existe un théorème d'existence encore plus général, le théorème de Carathéodory [7] 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

6. Giuseppe Peano, 1858–1932.
7. Constantin Carathéodory, 1873–1950.

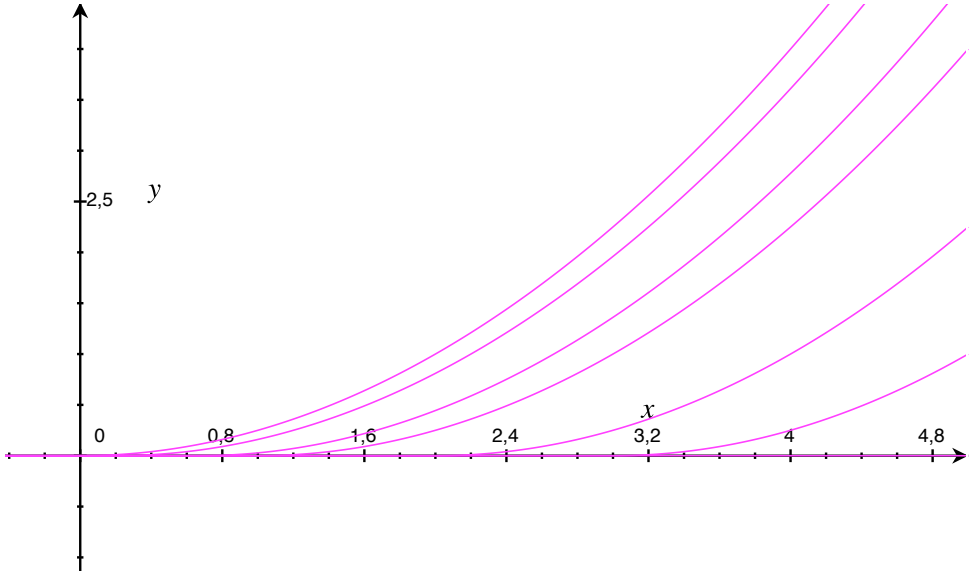


FIGURE 3.5 – Quelques-unes parmi l'infinité (non dénombrable) des solutions du problème de Cauchy $y'(t) = \sqrt{|y(t)|}$, $y(0) = 0$, avec une infinité de branchements contredisant l'unicité locale.

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.

3.1.1 Existence globale à l'aide des fonctions de Liapounov

Quand f est seulement localement lipschitzienne, mais pas globalement lipschitzienne, [8] 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 [9] associée à l'équation différentielle. On se place dans le cas où $V = \mathbb{R}^m$ pour simplifier.

Définition 3.1.13 *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. *$U(y) \to +\infty$ quand $\|y\| \to +\infty$,*
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

8. Qui est une condition quand même bien restrictive.
9. Alexandre Mikhaïlovitch Liapounov, 1857–1918.

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 3.1.14 *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 3.1.4 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 3.1.5, 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 1.5.6 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 1.2.7, où elle n'a pas vraiment d'autre choix que d'être constante.

Exemple 3.1.4 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 [10] (voir [2])

$$dU(y) = \left(\frac{\partial U}{\partial y_1}(y) \quad \cdots \quad \frac{\partial U}{\partial y_m}(y)\right),$$

avec le produit scalaire canonique de $\mathbb{R}^m$

$$dU(y)z = (\nabla U(y)|z).$$

Définition 3.1.15 *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 3.1.13, 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.

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 1.2.6, 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 3.1.5 Équations de Hamilton [11]. 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

$$J\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} 0_m & I_m \\ -I_m & 0_m \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} p \\ -q \end{pmatrix},$$

10. 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.

11. Sir William Rowan Hamilton, 1805–1865.

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 3.1.13, 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.

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 [12], 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* [13] 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 1.2.6). 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,

12. Voir définition 1.2.6 un peu plus loin.
13. Siméon-Denis Poisson, 1781–1840.

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. ◊

Chapitre 4

Méthodes numériques (suite et fin)

4.1 Schémas implicites

On a introduit au début de ce chapitre 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 ?uvre 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

$$y_0 = y(0), \quad y_{n+1} = y_n + h\Phi(t_{n+1}, y_{n+1}, t_n, y_n, h), \tag{4.1.1}$$

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

$$z = \varphi(z), \text{ avec } \varphi(z) = y_n + h\Phi(t_{n+1}, z, t_n, y_n, h). \tag{4.1.2}$$

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 4.1.1 *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 4.1.2 *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 $d(x^*, \tilde{x}^*) \leq k d(x^*, \tilde{x}^*)$, 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 $x^* = \tilde{x}^*$. [1]

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

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

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

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, [2] 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 4.1.3 *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, 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 4.1.4 *Le schéma implicite (4.1.1) 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

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

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 (2.1.12) avec

$$F(t,y,h) = \Phi(t + h, \Theta(t + h, t, y, h), t, y, h). \tag{4.1.3}$$

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. [3] 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 4.1.5 *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 $\mathbb{R}^m$ satisfaisant respectivement*

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

et

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

sont telles que

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

L'erreur de consistance du schéma (4.1.1) *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 (4.1.1) *est consistant si pour toute solution y de* (1.2.1)*, 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 implicite

Théorème 4.1.6 *Un schéma implicite* (4.1.1) *stable et consistant est convergent.*

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

Proposition 4.1.7 *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* (4.1.1) *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.*

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

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 $h \leq h_0$.

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

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

Par conséquent,

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

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

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

On applique alors le lemme de Grönwall discret version 2 2.1.5, 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

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

voir la démonstration de la Proposition 2.1.3 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 ?uvre 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 $p \to +\infty$ 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. [4]

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.

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 [5] 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 (4.1.3). 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 4.5, mais *in fine*, c'est la stabilité du schéma qui permet d'avoir (une certaine) confiance dans le résultat qui sort de la machine.

On verra à la section suivante qu'il existe des méthodes plus efficaces que les itérations de point fixe pour approcher y_{n+1}.

4.1.1 Méthodes de résolution des équations non linéaires

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*.

Décrivons dans un premier temps la méthode en dimension un. On notera plutôt $z = x$. Soit $\bar{x}$ une racine de g, c'est-à-dire une solution de l'équation $g(\bar{x}) = 0$ avec g au moins de classe C^1 sur un intervalle de $\mathbb{R}$. On suppose qu'on connaît une valeur approchée x_0 de la racine, d'une façon ou d'une autre. L' idée 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 $y = 0$ soit

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

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

$$x_{p+1} = x_p - \frac{g(x_p)}{g'(x_p)}.$$

4. De plus, $\|y_{n+1} - y_{n+1}^{p+1}\| \leq \frac{\alpha}{1-hM}$.
5. C'est normal, on ne peut faire de calcul numérique autre qu'explicite.

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 $\bar{x}$, alors on a $\bar{x} = \bar{x} - \frac{g(\bar{x})}{g'(\bar{x})}$, soit $g(\bar{x}) = 0$.

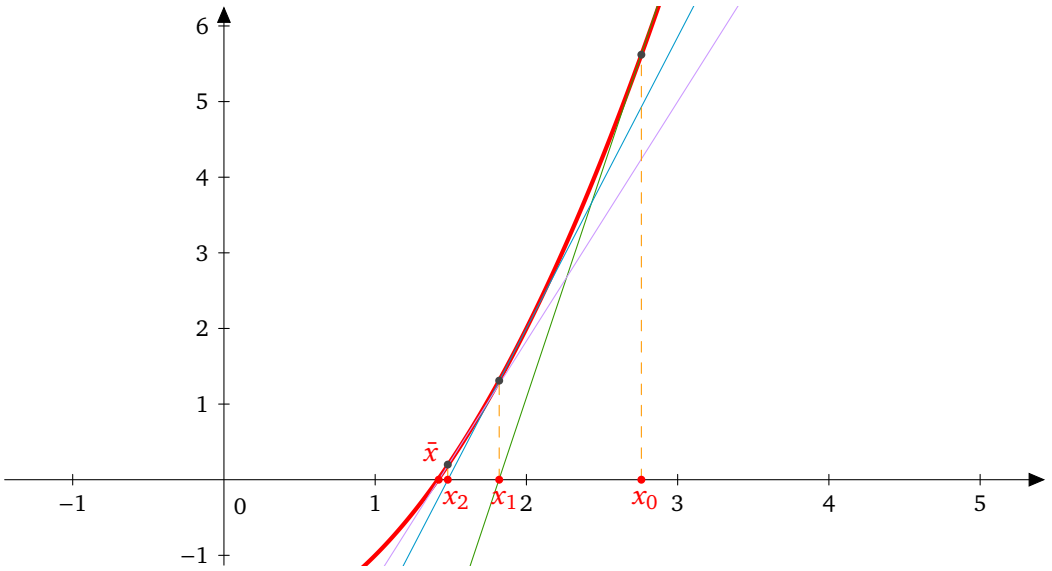


FIGURE 4.1 – Principe de la méthode de Newton en dimension un.

Analysons plus précisément la méthode de Newton.

Théorème 4.1.8 *On suppose que g est de classe C^2 sur l'intervalle $I = [\bar{x} - r, \bar{x} + r]$ 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 $x_0 \in]\bar{x} - \alpha, \bar{x} + \alpha[$, la suite de Newton x_p est bien définie pour tout p et converge vers $\bar{x}$ quand $p \longrightarrow +\infty$, avec l'estimation

$$|x_p - \bar{x}| \leq \frac{1}{K} (K|x_0 - \bar{x}|)^{2^p}, \tag{4.1.5}$$

où $K = \frac{M}{2m}$.

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

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

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

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

On voit donc que

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

Si $x \in]\bar{x} - \alpha, \bar{x} + \alpha[\subset I$, on a donc $|x - \bar{x}| < \alpha$, d'où

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

puisque $\alpha \leq \frac{2m}{M}$. On en déduit que $\psi(x) \in]\bar{x} - \alpha, \bar{x} + \alpha[$. La suite de Newton x_p est donc bien définie, puisque l'intervalle $]\bar{x} - \alpha, \bar{x} + \alpha[$ est invariant par ψ.

De plus, on obtient pour tout p

$$K|x_{p+1} - \bar{x}| \leq (K|x_p - \bar{x}|)^2.$$

d'où l'estimation (4.1.5) par récurrence. Comme $K|x_0 - \bar{x}| < 1$, on voit que $x_p \to \bar{x}$ quand $p \to +\infty$. ◊

Non seulement la méthode de Newton converge sous les hypothèses précédentes, mais elle converge extrêmement rapidement. Mettons pour fixer les idées que $K = 1$. La partie entière de la quantité $-\log_{10}(|x_p - \bar{x}|)$ nous donne le nombre de décimales exactes à l'itération p (peut-être à une unité près...), et cette quantité double à chaque itération. On peut donc dire que le nombre de décimales exactes double grosso-modo à chaque itération, c'est-à-dire qu'il croît exponentiellement, ce qui va très vite. C'est à comparer avec les itérations de point fixe pour lesquelles on ne s'attend a priori qu'à une croissance linéaire du nombre de décimales exactes avec les itérations. Par contre, la méthode de Newton demande plus de calculs et sa convergence n'est pas assurée pour toute valeur initiale de l'itération. Pour les applications aux schémas implicites, on a besoin de résoudre des équations à valeurs dans $\mathbb{R}^m$ avec $m \geq 1$. Ainsi le schéma d'Euler implicite $y_{n+1} = y_n + hf(t_{n+1}, y_{n+1})$ va conduire à résoudre à chaque pas de temps un système d'équations non linéaires $G(z) = 0$, où la fonction $G : \mathbb{R}^m \longrightarrow \mathbb{R}^m$ est définie par $G(z) = z - y_n - hf(t_{n+1}, z)$. Pour cela on va utiliser l'extension de la méthode de Newton au cas vectoriel, connue sous le nom de *méthode de Newton-Raphson* [6].

On veut résoudre numériquement une équation $G(x) = 0$ où $G : U \to \mathbb{R}^m$ est une application de classe C^2 définie sur un ouvert U de $\mathbb{R}^m$, connaissant une valeur approchée x_0 de la solution $\bar{x}$. Comme dans la méthode de Newton scalaire, l'idée est d'approcher G par sa partie linéaire au voisinage de x_0

$$G(x) = G(x_0) + \nabla G(x_0)(x - x_0) + o(\|x - x_0\|).$$

Ici $\nabla G(x)$ est la matrice jacobienne

$$(\nabla G(x))_{ij} = \frac{\partial G_i}{\partial x_j}(x).$$

On cherche à résoudre l'équation $G(x_0) + \nabla G(x_0)(x_1 - x_0) = 0$. Si la matrice $\nabla G(x_0)$ est inversible, on a une solution unique x_1 simplement donnée par

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

et si $\nabla G(x_1)$ est encore inversible, on pourra itérer pour calculer x_2, etc. La méthode de Newton-Raphson consiste donc à construire la suite

$$x_{p+1} = x_p - (\nabla G(x_p))^{-1} G(x_p).$$

6. Joseph Raphson, ≈ 1648–1715.

Il est facile de voir que si G est de classe C^2 et que $\nabla G(\bar{x})$ est inversible, alors il existe une boule centrée en $\bar{x}$ telle que pour toute donnée initiale dans cette boule, la suite est bien définie et a la même convergence quadratique qu'en dimension 1

$$\|x_{p+1} - \bar{x}\| \leq K\|x_p - \bar{x}\|^2,$$

pour un certain K. En fait la démonstration est exactement la même, en remplaçant la formule de Taylor-Lagrange utilisée dans le cas $m = 1$ par une version valable pour $m > 1$, par exemple avec reste intégral, et en multipliant à gauche par $(\nabla G(x))^{-1}$ au lieu de diviser par $g'(x)$.

Revenons à notre schéma implicite et voyons si $\nabla G(z)$ est bien inversible dans des conditions raisonnables. On a $G(z) = z - y_n - hf(t_{n+1}, z)$ d'où

$$\nabla G(z) = I - h\nabla f(t_{n+1}, z).$$

On rappelle un résultat d'algèbre linéaire bien connu.

Proposition 4.1.9 *Pour toute matrice $M \in M_m(\mathbb{R})$ telle que $\|M\| < 1$, la matrice $I + M$ est inversible.*

Par conséquent, en choisissant un pas de discrétisation h inférieur à

$$h_0 < \frac{1}{\max_x \|\nabla f(t_{n+1}, x)\|},$$

on s'assure que ∇G est inversible. Cette condition n'est bien sûr pas sans rapport avec celle assurant le caractère bien posé de la méthode d'Euler implicite. Notons qu'en pratique, dès que m est un peu grand, on ne calcule jamais $(\nabla G(x))^{-1}$, car c'est beaucoup trop coûteux, mais on résout le système linéaire $\nabla G(x_p)x_{p+1} = \nabla G(x_p)x_p - G(x_p)$, ce qui est beaucoup plus efficace.

4.2 Stabilité absolue

Nous avons vu l'importance de la notion de stabilité donnée par la définition 2.1.2, qui assure que des perturbations du schéma entraînent des perturbations des solutions calculées qui restent contrôlables. Le contrôle en question peut-être néanmoins être 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,

$$\begin{cases} y'(t) = \lambda(y(t) - g(t)) + g'(t), \\ y(0) = y_0 \end{cases} \tag{4.2.1}$$

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 (4.2.1) 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 2.1.2 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. [7]

Nous ne considérons dans cette section que le problème de Cauchy scalaire suivant

$$\begin{cases} y'(t) = \lambda y(t) \text{ dans } I, \\ y(0) = y_0 \neq 0, \end{cases} \tag{4.2.2}$$

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 (4.2.2), $y'(t) = Ay(t)$. Mais nous avons vu qu'après diagonalisation, si celle-ci est possible, on obtient un système découplé d'équations scalaires de type (4.2.2), voir l'exemple en page 42. Ces dernières jouent donc un rôle important.

Supposons que l'on commette une erreur sur la donnée initiale de (4.2.2) 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 (4.2.2) 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. [8]

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 (2.1.12). [9] Comme l'EDO est linéaire et autonome, on suppose que la fonction F est linéaire par

7. 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.

8. 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 $h \to 0$.

9. Dans le cas implicite, on donne simplement un nom à la fonction implicite définissant le schéma.

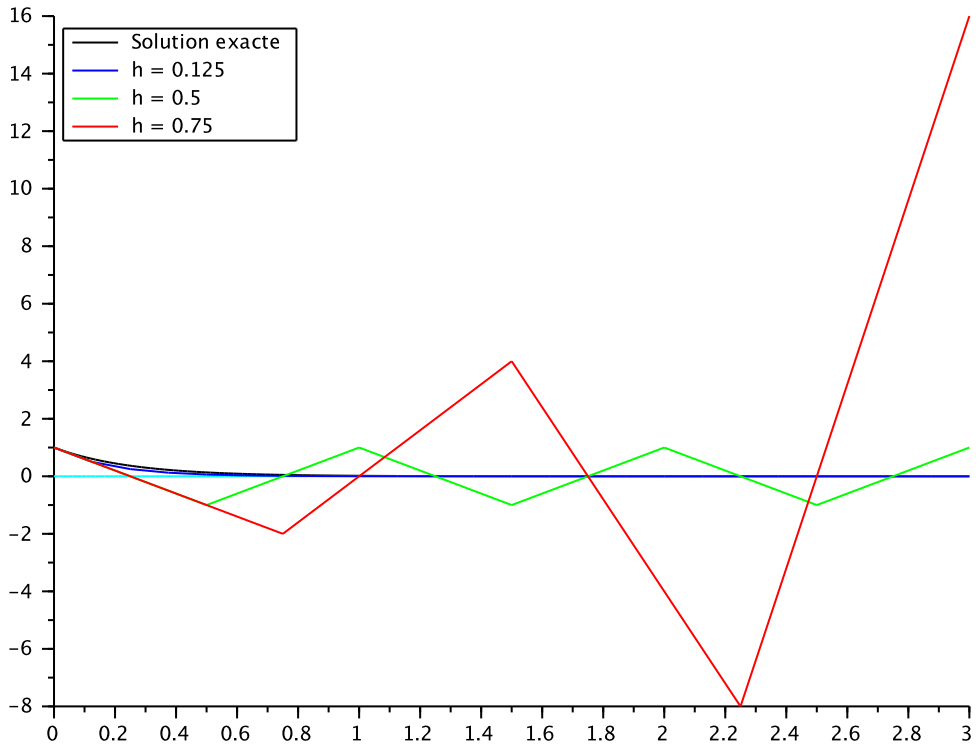


FIGURE 4.2 – Illustration de la problématique de la stabilité absolue avec $y_0 = 1$, $\lambda = -4$ et diverses valeurs de h avec le schéma d'Euler explicite. Il n'y a pas de rapport avec la stabilité au sens précédent.

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

$$y_{n+1} = G(\lambda h)y_n, \tag{4.2.3}$$

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 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 $G(z) = \frac{2+z}{2-z}$.

4.2.1 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 4.2.1 *L'ensemble des $z \in \mathbb{C}$ tels $|G(z)| < 1$ est appelé* domaine de stabilité *(ou domaine de stabilité absolue) du schéma* (4.2.3).

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 4.3).

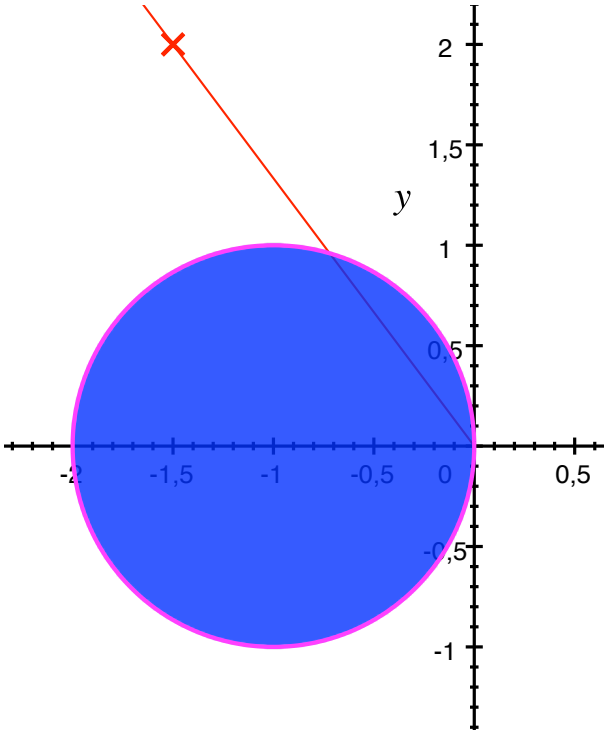


FIGURE 4.3 – Domaine de stabilité absolue du schéma d'Euler explicite.

Soit $\lambda \in \mathbb{C}$, à partie réelle $\Re(\lambda)$ négative. Pour discrétiser l'équation (4.2.2) 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 4.3, 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 (4.2.2). 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

$$0 < h < -2\frac{\cos\theta}{\varrho}. \tag{4.2.4}$$

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 (4.2.2), 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 (4.2.4) soit vérifiée.

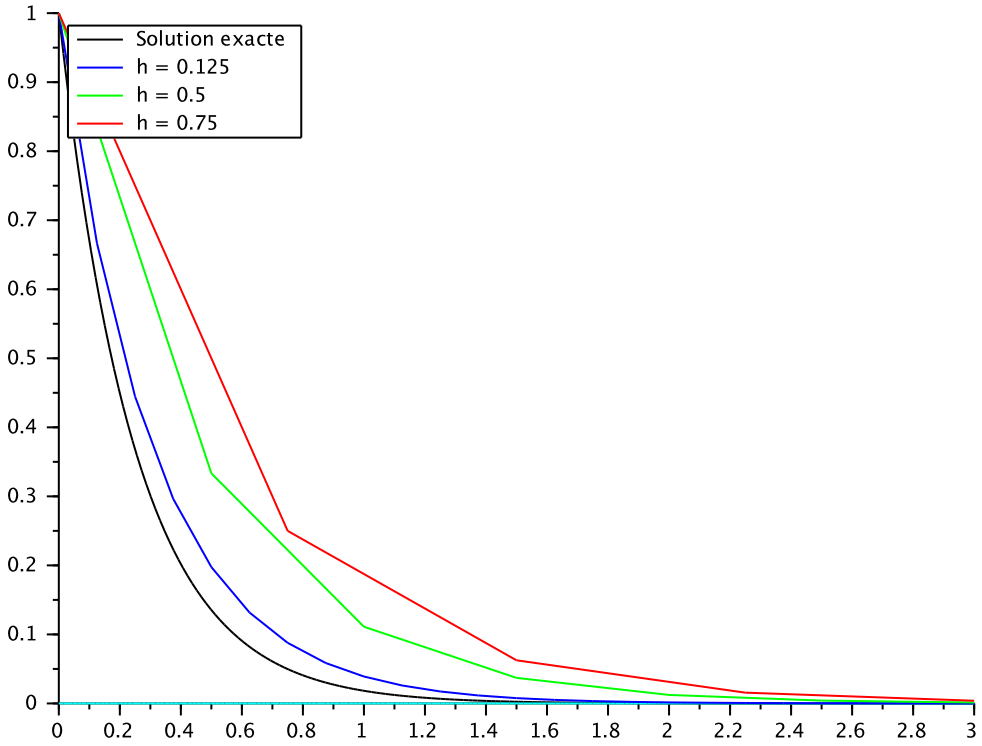


FIGURE 4.4 – Le même calcul qu'à la Figure 4.2 avec Euler implicite.

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 4.5.

4.3 Diverses familles de schémas d'ordre aussi élevé qu'on veut

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.

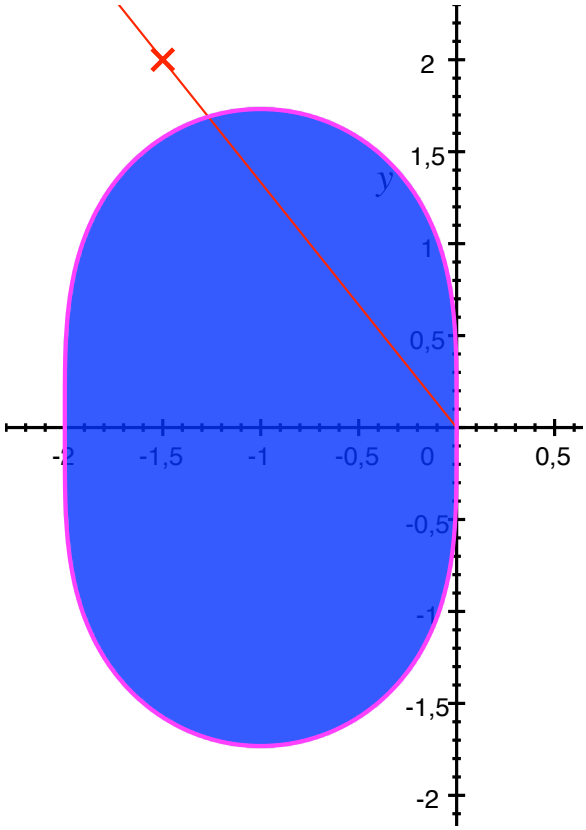


FIGURE 4.5 – Domaine de stabilité absolue du schéma d'Euler modifié.

4.3.1 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,

$$y(t_{n+1}) = y(t_n) + \sum_{k=1}^{p} \frac{h^k}{k!} y^{(k)}(t_n) + O(h^{p+1}). \tag{4.3.1}$$

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 (2.1.22). Le développement de Taylor (4.3.1). 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, [10]

$$y_{n+1} = y_n + h \sum_{k=1}^{p} \frac{h^{k-1}}{k!} f^{k-1}(t_n, y_n). \tag{4.3.2}$$

10. On ne fait pas apparaître cette indexation par p dans la notation.

Ces schémas sont de la forme (2.1.12), 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. 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 4.3.1 *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 (4.3.2) 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'amusement de longue soirée d'hiver, 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 (4.3.2) se simplifie en

$$y_{n+1} = \left(\sum_{k=0}^{p} \frac{(\lambda h)^k}{k!}\right) y_n. \tag{4.3.3}$$

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. [11] 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...

11. 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.

4.3.2 Méthodes d'Adams

Les méthodes d'Adams [12] sont des méthodes à pas multiples qui sont fondées sur l'interpolation polynomiale de Lagrange (voir [4] chapitre 1). Remarquons que nous n'avons pas traité la théorie générale des schémas à pas multiples, consistance, ordre, stabilité et convergence. On agitera un peu les mains à ces sujets... D'abord un petit rappel sur l'interpolation de Lagrange.

Théorème 4.3.2 *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, $0 \leq j \leq q$. Il s'écrit :*

$$P(t) = \sum_{j=0}^{q} \beta_j \prod_{k \neq j} \left(\frac{t - \alpha_k}{\alpha_j - \alpha_k} \right). \tag{4.3.4}$$

Une forme plus algorithmique est la forme de Newton du polynôme, utilisant les différences divisées

$$P(t) = \beta[\alpha_0] + \sum_{i=1}^{q} \beta[\alpha_0, \ldots, \alpha_i] \prod_{k=0}^{i-1} (t - \alpha_k),$$

où lesdites différences divisées sont définies par récurrence

$$\begin{aligned} \beta[\alpha_i] &= \beta_i, \\ \beta[\alpha_i, \ldots, \alpha_{i+p}] &= \frac{\beta[\alpha_{i+1}, \ldots, \alpha_{i+p}] - \beta[\alpha_i, \ldots, \alpha_{i+p-1}]}{\alpha_{i+p} - \alpha_i}. \end{aligned}$$

Cette forme du polynôme d'interpolation sera plus adaptée à la définition des schémas d'ordre élevé.

Les méthodes d'Adams entrent dans la deuxième catégorie de schémas présentée au chapitre précédent, 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 4.3.2 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 (2.1.8) 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 (4.3.4). 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 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

12. John Couch Adams, 1819–1892.

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

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} P_{n,q}(t)\,dt. \tag{4.3.5}$$

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 $n - q + d \leq j \leq n + d$. 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 [13]. Si $d = 1$, les schémas obtenus, appelés schémas d'Adams-Moulton [14], 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 $s = \frac{t-t_n}{h}$,

$$\int_{t_n}^{t_{n+1}} P_{n,1}(t)\,dt = h\int_0^1 \big(f_n + (f_n - f_{n-1})s\big)\,ds = h\left(f_n\left[1 + \frac{s^2}{2}\right]_0^1 - f_{n-1}\left[\frac{s^2}{2}\right]_0^1\right),$$

d'où le schéma

$$\text{(AB2)} \qquad y_{n+1} = y_n + h\left(\frac{3}{2}f_n - \frac{1}{2}f_{n-1}\right). \tag{4.3.6}$$

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 4.3.3 *Le schéma* AB2 *est d'ordre deux.*

Démonstration. Comme toujours pour ces questions d'ordre de schémas, il s'agit de développements de Taylor brutaux, ayant une confiance aveugle et touchante [15] dans l'uniformité des O, mais sans

13. 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.

14. 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*.

15. Mais en fait justifiée.

malice.

$$\begin{aligned}\varepsilon_n &= y(t_{n+1}) - y(t_n) - \frac{h}{2}(3f(t_n, y(t_n)) - f(t_{n-1}, y(t_{n-1})))\\ &= y(t_n + h) - y(t_n) - \frac{h}{2}\big(3y'(t_n) - y'(t_n - h)\big)\\ &= hy'(t_n) + \frac{h^2}{2}y''(t_n) + O(h^3) - \frac{h}{2}\Big[3y'(t_n) - \Big(y'(t_n) - hy''(t_n) + O(h^2)\Big)\Big]\\ &= O(h^3).\end{aligned}$$

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 $j = 0, 1, 2$. On peut utiliser l'algorithme de Newton des différences divisées, mais comme il y a encore peu de points ici, la formule utilisant les polynômes de base de l'interpolation de Lagrange (4.3.4) reste accessible. Ainsi, 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

$$\text{(AB3)} \qquad y_{n+1} = y_n + h\left(\frac{23}{12}f_n - \frac{4}{3}f_{n-1} + \frac{5}{12}f_{n-2}\right). \tag{4.3.7}$$

Proposition 4.3.4 *Le schéma* AB3 *est d'ordre trois.*

Démonstration. Comme toujours pour ces questions d'ordre de schémas, [...], mais sans malice.

$$\begin{aligned}
\varepsilon_n &= y(t_{n+1}) - y(t_n) - \frac{h}{12}(23f(t_n, y(t_n)) - 16f(t_{n-1}, y(t_{n-1})) + 5f(t_{n-2}, y(t_{n-2})) \\
&= y(t_n + h) - y(t_n) - \frac{h}{12}\big(23y'(t_n) - 16y'(t_n - h) + 5y'(t_n - 2h)\big) \\
&= hy'(t_n) + \frac{h^2}{2}y''(t_n) + \frac{h^3}{6}y'''(t_n) + O(h^4) \\
&\qquad - \frac{h}{12}\Big[23y'(t_n) - 16\Big(y'(t_n) - hy''(t_n) + \frac{h^2}{2}y'''(t_n) + O(h^3)\Big) \\
&\qquad\qquad + 5\big(y'(t_n) - 2hy''(t_n) + 2h^2y'''(t_n) + O(h^3)\big)\Big] \\
&= O(h^4).
\end{aligned}$$

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, ou avec les différences divisées, 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 4.3.5 *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, beaucoup plus compliquées comme on va le voir incessamment.

Avant cela, qu'en est-il de la stabilité des schémas d'Adams ? Il s'agit, comme dans le cas des schémas à un pas, de regarder l'influence de perturbations introduites à chaque pas de temps. 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

$$\begin{aligned}
z_0 &= y_0 + \eta_0 \\
z_1 &= y_1 + \eta_1 \\
z_{n+1} &= z_n + h\left(\frac{3}{2}f(t_n, z_n) - \frac{1}{2}f(t_{n-1}, z_{n-1})\right) + \eta_{n+1}, \quad \text{pour } n = 1, \dots, N-1.
\end{aligned}$$

on a donc pour $n = 1, \dots, N-1$

$$\begin{aligned}
z_{n+1} - y_{n+1} &= z_n - y_n + h\frac{3}{2}\Big(f(t_n, z_n) - f(t_n, y_n)\Big) - \frac{1}{2}\Big(f(t_{n-1}, z_{n-1}) - f(t_{n-1}, y_{n-1})\Big) + \eta_{n+1}, \quad \text{d'où} \\
\|z_{n+1} - y_{n+1}\| &\le \|z_n - y_n\| + h\frac{3}{2}\|f(t_n, z_n) - f(t_n, y_n)\| + h\frac{1}{2}\|f(t_{n-1}, z_{n-1}) - f(t_{n-1}, y_{n-1}\| + \|\eta_{n+1}\|
\end{aligned}$$

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

$$\|z_{n+1} - y_{n+1}\| \quad \leq \quad (1 + hL\frac{3}{2})\|z_n - y_n\| + hL\frac{1}{2}\|z_{n-1}) - y_{n-1}\| + \|\eta_{n+1}\|.$$

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

$$\theta_n \quad = \quad \max_{k=0,\ldots,n} \|\eta_k\|.$$

Pour $n \geq 1$ on a $\|z_n - y_n\| \leq \theta_n$ et $\|z_{n-1}) - y_{n-1}\| \leq \theta_n$ donc

$$\|z_{n+1} - y_{n+1}\| \quad \leq \quad (1 + hL\frac{3}{2})\theta_n + hL\frac{1}{2}\theta_n + \|\eta_{n+1}\| = (1 + 2hL)\theta_n + \|\eta_{n+1}\|$$

Par ailleurs $\theta_n \leq (1 + 2hL)\theta_n$ donc

$$\theta_{n+1} = \max(\|z_{n+1} - y_{n+1}\|, \theta_n) \quad \leq \quad (1 + 2hL)\theta_n + \|\eta_{n+1}\|.$$

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 $\|\eta_k\|$.

Domaine de stabilité des schémas à pas multiples. 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 (4.2.2), 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 $z = h\lambda$. [16] 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,

16. On a supposé implicitement que ces racines sont simples, ce qui est le cas générique.

Définition 4.3.6 *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 (4.2.2)

$$y_{n+1} = y_n + h\lambda\left(\frac{3}{2}y_n - \frac{1}{2}y_{n-1}\right).$$

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 [17]. On a utilisé cette description pour construire la figure 4.6. La frontière du domaine de stabilité est incluse dans la courbe $\theta \mapsto F(e^{i\theta})$. [18]

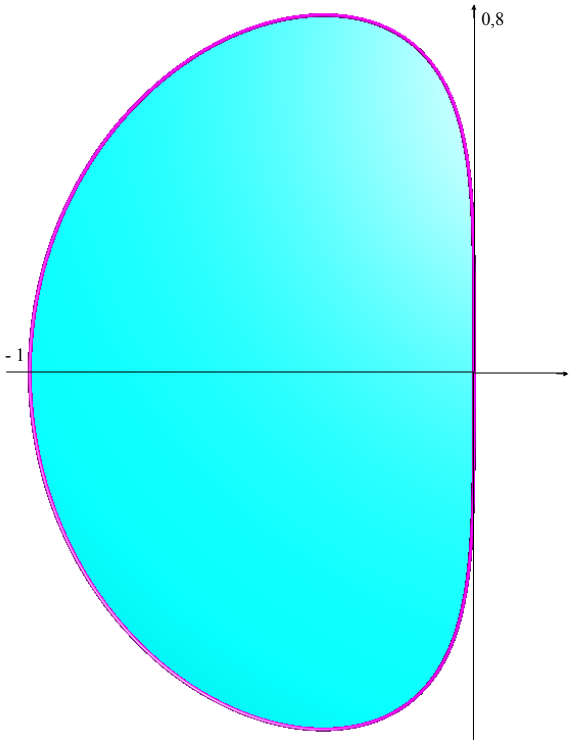


FIGURE 4.6 – Domaine de stabilité absolue du schéma d'Adams-Bashforth d'ordre 2 en bleu.

17. 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.

18. Elle lui est égale dans les premiers exemples considérés ici, mais pas toujours, voir plus loin.

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 (4.2.2), il s'écrit

$$y_{n+1} = y_n + \frac{h\lambda}{2}\big(y_{n+1} + y_n\big),$$

et son équation caractéristique est donc

$$P(X) = X - 1 - \frac{z}{2}(X+1),$$

d'où

$$F(X) = 2\left(\frac{X-1}{X+1}\right).$$

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 :

$$\text{(AB}_3\text{)} \quad F(X) = \frac{12(X^3 - X^2)}{23X^2 - 16X + 5},$$
$$\text{(AM}_3\text{)} \quad F(X) = \frac{12(X^2 - X)}{5X^2 + 8X - 1}.$$

Les domaines de stabilité correspondants sont représentés dans les figures 4.7 et 4.8. 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

$$F(X) = \frac{24(X^4 - X^3)}{55X^3 - 59X^2 + 37X - 9}.$$

On a dessiné dans la figure 4.9, 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.

4.3.3 Schémas de Runge-Kutta.

Les méthodes de Runge-Kutta [19] sont aussi fondées sur l'intégration de l'EDO entre t_n et t_{n+1}, mais cette fois-ci, on ne va plus se servir que de la valeur en t_n, et d'un certain nombre de

19. Carl David Tolmé Runge, 1856–1927 ; Martin Wilhelm Kutta, 1867–1944.

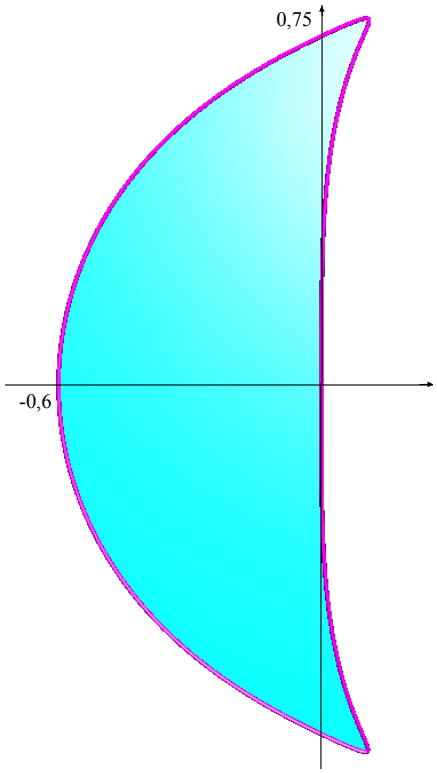


FIGURE 4.7 – Domaine de stabilité absolue du schéma d'Adams-Bashforth d'ordre 3 (seul le côté à partie réelle négative compte).

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, [20] sur l'intervalle $[0, 1]$

$$\int_0^1 \varphi(s)\, ds \approx \sum_{j=1}^{q} b_j \varphi(c_j), \tag{4.3.8}$$

dont les c_j sont les nœuds et les b_j les poids. On suppose que cette formule est exacte pour les fonctions constantes, c'est-à-dire que

$$\sum_{j=1}^{q} b_j = 1. \tag{4.3.9}$$

Les c_i servent à définir q instants intermédiaires $t_{n,i}$ (pas forcément distincts) entre t_n et t_{n+1},

$$t_{n,i} = t_n + c_i h, \quad 1 \leq i \leq q. \tag{4.3.10}$$

20. C'est-à-dire une formule d'approximation numérique de l'intégrale. La nature regorge de formules de quadrature.

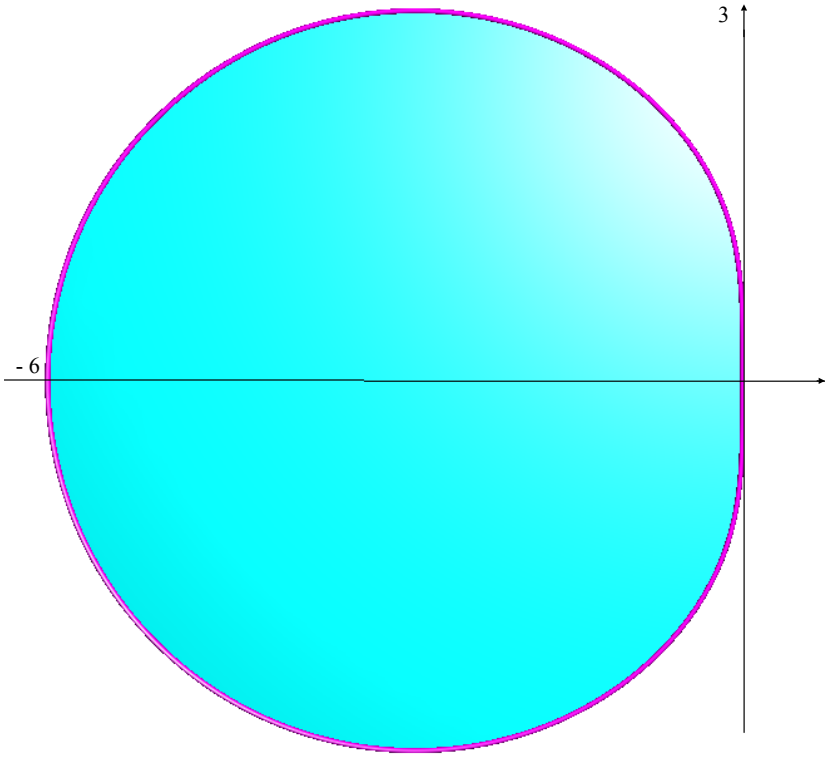


FIGURE 4.8 – Domaine de stabilité absolue du schéma d'Adams-Moulton d'ordre 3

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

$$\int_0^{c_i} \varphi(s)\,ds \approx \sum_{j=1}^{q} a_{ij}\varphi(c_j), \tag{4.3.11}$$

qu'on supposera aussi exacte pour les constantes

$$\sum_{j=1}^{q} a_{ij} = c_i. \tag{4.3.12}$$

En intégrant l'EDO (1.2.1) entre les instants t_n et $t_{n,i}$, on obtient

$$y(t_{n,i}) - y(t_n) = \int_{t_n}^{t_{n,i}} f\big(t, y(t)\big)\,dt = h\int_0^{c_i} f\big(t_n + sh, y(t_n + sh)\big)\,ds, \tag{4.3.13}$$

via le changement de variable qui s'impose, $t = t_n + sh$. L'utilisation de la formule de quadrature (4.3.11) pour discrétiser l'intégrale apparaissant dans (4.3.13), 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

$$y_{n,i} = y_n + h\sum_{j=1}^{q} a_{ij} f\big(t_{n,j}, y_{n,j}\big), \tag{4.3.14}$$

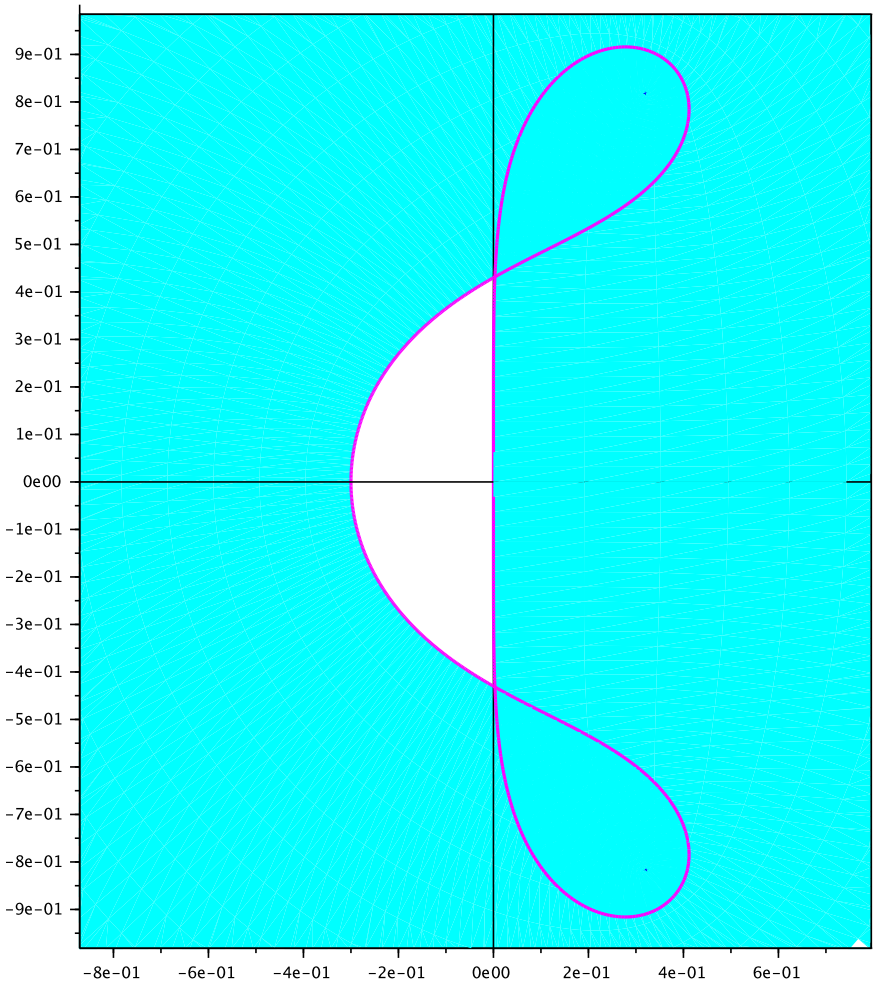


FIGURE 4.9 – Domaine de stabilité absolue du schéma d'Adams-Bashforth d'ordre 4.

pour $i = 1, \dots, q$. Notons que si $a_{ij} = 0$ pour $i \leq j$, les relations (4.3.14) 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

$$\begin{pmatrix} y_{n,1} \\ \vdots \\ y_{n,q} \end{pmatrix} = \begin{pmatrix} y_n \\ \vdots \\ y_n \end{pmatrix} + hA \begin{pmatrix} f(t_{n,1}, y_{n,1}) \\ \vdots \\ f(t_{n,q}, y_{n,q}) \end{pmatrix}, \tag{4.3.15}$$

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).

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 (4.3.8), on remplace les valeurs exactes par les approximations supposées et l'on obtient le schéma de Runge-Kutta

$$y_{n+1} = y_n + h \sum_{j=1}^{q} b_j f\big(t_{n,j}, y_{n,j}\big), \tag{4.3.16}$$

les valeurs intermédiaires $y_{n,j}$ étant, rappelons le, calculées par (4.3.14). 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 4.1, appelés *tableaux de Butcher* [21].

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1q} \\ \vdots & \vdots & & \vdots \\ c_q & a_{q1} & \dots & a_{qq} \\ \hline & b_1 & \dots & b_q \end{array}$$

TABLE 4.1 – 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 4.3.7 *Soit $f : [0,T] \times \mathbb{R} \to \mathbb{R}$ lipschitzienne par rapport à t 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 $\|A\|_\infty$

21. John Charles Butcher, 1933–.

la norme matricielle subordonnée. [22] Fixons y, t et h. On a donc pour tous z et $\tilde{z}$ dans $\mathbb{R}^q$,

$$\begin{aligned}\|\psi_y(z)-\psi_y(\tilde{z})\|_\infty &= h\|A(G(z)-G(\tilde{z})\|_\infty\\ &\le h\|A\|_\infty\|G(z)-G(\tilde{z})\|_\infty\\ &= h\|A\|_\infty \max_{1\le i\le q}|f(t+c_ih,z_i)-f(t+c_ih,\tilde{z}_i)|\\ &\le h\|A\|_\infty L\max_{1\le i\le q}|z_i-\tilde{z}_i|\\ &= hL\|A\|_\infty\|z-\tilde{z}\|_\infty.\end{aligned}$$

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. [23] Au vu de (4.3.16), on a

$$F(t,y,h)=\sum_{j=1}^{q}b_jf(t+c_jh,\mathrm{int}\,(y)_j), \tag{4.3.17}$$

d'où pour tout couple de réels $(y,\tilde{y})$,

$$\begin{aligned}|F(t,y,h)-F(t,\tilde{y},h)| &\le \sum_{j=1}^{q}|b_j||f(t+c_jh,\mathrm{int}\,(y)_j)-f(t+c_jh,\mathrm{int}\,(\tilde{y})_j)|\\ &\le L\sum_{j=1}^{q}|b_j||\mathrm{int}\,(y)_j-\mathrm{int}\,(\tilde{y})_j|\\ &\le L\max_j|\mathrm{int}\,(y)_j-\mathrm{int}\,(\tilde{y})_j|\sum_{j=1}^{q}|b_j|\\ &= L\|b\|_1\|\mathrm{int}\,(y)-\mathrm{int}\,(\tilde{y})\|_\infty.\end{aligned}$$

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,

$$\|\mathrm{int}\,(y)-\mathrm{int}\,(\tilde{y})\|_\infty\le|y-\tilde{y}|+hL\|A\|_\infty\|\mathrm{int}\,(y)-\mathrm{int}\,(\tilde{y})\|_\infty,$$

par le même calcul que plus haut, et donc

$$\|\mathrm{int}\,(y)-\mathrm{int}\,(\tilde{y})\|_\infty\le\frac{1}{1-hL\|A\|_\infty}|y-\tilde{y}|$$

dès que $h<\frac{1}{L\|A\|_\infty}$. On en déduit donc que

$$|F(t,y,h)-F(t,\tilde{y},h)|\le\frac{L\|b\|_1}{1-h_0L\|A\|_\infty}|y-\tilde{y}|,$$

22. Cette norme est donnée par $\|A\|_\infty=\max_i\sum_j|a_{ij}|$.

23. 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.

dès que $h \leq h_0 < \frac{1}{L\|A\|_\infty}$, d'où la stabilité. [24] ◊

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 4.3.9 page 143. 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 11, page 75, à ce propos. Une fois les valeurs intermédiaires calculées, il n'y a plus d'équation à résoudre pour obtenir le pas suivant.

Remarque 4.3.1 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.

24. 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.

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.

Proposition 4.3.8 *Sous les mêmes hypothèses, un schéma de Runge-Kutta est consistant.*

Démonstration. On reprend la formule (4.3.17) 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 [25] 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 (4.3.9), on a $b_1 = 1$ et il y a un seul instant intermédiaire $t_{n,1}$. Le schéma s'écrit

$$\begin{cases} y_{n,1} = y_n + a_{11} h f(t_{n,1}, y_{n,1}), \\ y_{n+1} = y_n + h f(t_{n,1}, y_{n,1}). \end{cases}$$

Si l'on souhaite imposer (4.3.12), 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

(RK1)
$$\begin{cases} y_{n,1} = y_n, \\ y_{n+1} = y_n + h f(t_n, y_n). \end{cases} \qquad \begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

25. Si on ne sait pas, on le montre.

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 (4.3.9), on impose $b_1 + b_2 = 1$. D'après (4.3.12), on prend aussi $a_{21} = c_2$, ce qui n'est pas essentiel, mais diminue le nombre de paramètres. Encore pour avoir (4.3.12), 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

$$\begin{aligned} \varepsilon_n &= y(t_{n+1}) - y(t_n) - bhf(t_n, y(t_n)) - (1-b)hf\big(t_n + ch, y(t_n) + chf\big(t_n, y(t_n)\big)\big) \\ &= y(t_{n+1}) - y(t_n) - bhy'(t_n) - (1-b)hf\big(t_n + ch, y(t_n) + chy'(t_n)\big), \end{aligned} \tag{4.3.18}$$

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 [26] 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),

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

26. avec un reste en O un peu vague...

Rappelons que

$$y''(t) = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t))y'(t). \tag{4.3.19}$$

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 (4.3.18), 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.

Pour $c = 1$, on obtient le *schéma de Heun* [27], 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 $d = c = \frac{1}{2}$.

27. Karl Heun, 1859–1929.

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,

$$\text{(RK4)} \quad \begin{cases} y_{n,1} = y_n, \\ y_{n,2} = y_n + \dfrac{h}{2} f(t_n, y_{n,1}), \\ y_{n,3} = y_n + \dfrac{h}{2} f\left(t_n + \dfrac{h}{2}, y_{n,2}\right), \\ y_{n,4} = y_n + h f\left(t_n + \dfrac{h}{2}, y_{n,3}\right), \\ y_{n+1} = y_n + \dfrac{h}{6}\left[f(t_n, y_{n,1}) + 2f\left(t_n + \dfrac{h}{2}, y_{n,2}\right) + 2f\left(t_n + \dfrac{h}{2}, y_{n,3}\right) + f(t_{n+1}, y_{n,4}) \right]. \end{cases}$$

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.

La formulation alternative sans les valeurs intermédiaires du schéma RK4 s'écrit

$$\begin{cases} f_{n,1} = f(t_n, y_n), \\ f_{n,2} = f\left(t_n + \dfrac{h}{2}, y_n + \dfrac{h}{2} f_{n,1}\right), \\ f_{n,3} = f\left(t_n + \dfrac{h}{2}, y_n + \dfrac{h}{2} f_{n,2}\right), \\ f_{n,4} = f(t_n + h, y_n + h f_{n,3}), \\ y_{n+1} = y_n + \dfrac{h}{6}\left(f_{n,1} + 2f_{n,2} + 2f_{n,3} + f_{n,4}\right). \end{cases}$$

Dans le cas où $f(t, y) = g(t)$ de dépend pas de y, le schéma RK4 redonne la méthode de Simpson [28], dont il est bien connu qu'il s'agit d'une méthode d'intégration numérique d'ordre 4. En effet, dans ce cas $y_n = h\left(\frac{1}{6} g(t_0) + \frac{2}{3} \sum_{k=0}^{n-1} g(t_k + \frac{h}{2}) + \frac{1}{3} \sum_{k=1}^{n-1} g(t_k) + \frac{1}{6} g(t_n)\right)$.

28. Thomas Simpson, 1710–1761. La formule était apparemment déjà utilisée par Kepler un bon siècle auparavant.

Juste pour le fun, écrivons le schéma RK4 sous la forme standard explicite à un pas $y_{n+1} = y_n + hF(t_n, y_n, h)$. Il vient

$$\begin{aligned} y_{n+1} = y_n + \frac{h}{6}\Big[& f(t_n, y_n) + 2f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f(t_n, y_n)\right) \\ & + 2f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f(t_n, y_n)\right)\right) \\ & + f\left(t_n + h, y_n + hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f(t_n, y_n)\right)\right)\right)\Big], \end{aligned}$$

soit

$$\begin{aligned} F(t, y, h) = \frac{1}{6}\Big[& f(t, y) + 2f\left(t + \frac{h}{2}, y + \frac{h}{2}f(t, y)\right) \\ & + 2f\left(t + \frac{h}{2}, y + \frac{h}{2}f\left(t + \frac{h}{2}, y + \frac{h}{2}f(t, y)\right)\right) \\ & + f\left(t + h, y + hf\left(t + \frac{h}{2}, y + \frac{h}{2}f\left(t + \frac{h}{2}, y + \frac{h}{2}f(t, y)\right)\right)\right)\Big], \end{aligned}$$

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 et juste pour le fun que le schéma RK4 est bien d'ordre 4. On peut penser à utiliser la Proposition 2.1.14, 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 extraordinairement pénible. Procédons plutôt par développements de Taylor usuels. [29] 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 $f(t_n, e_{n,1})$, $f\left(t_n + \frac{h}{2}, e_{n,2}\right)$, $f\left(t_n + \frac{h}{2}, e_{n,3}\right)$ 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 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.

$$\begin{aligned} y' &= f, \\ y'' &= \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}y', \\ y''' &= \frac{\partial^2 f}{\partial t^2} + 2\frac{\partial^2 f}{\partial t \partial y}y' + \frac{\partial^2 f}{\partial y^2}(y')^2 + \frac{\partial f}{\partial y}y'', \\ y^{(4)} &= \frac{\partial^3 f}{\partial t^3} + 3\frac{\partial^3 f}{\partial t^2 \partial y}y' + 3\frac{\partial^3 f}{\partial t \partial y^2}(y')^2 + \frac{\partial^3 f}{\partial y^3}(y')^3 + 3\frac{\partial^2 f}{\partial t \partial y}y'' + 3\frac{\partial^2 f}{\partial y^2}y'y'' + \frac{\partial f}{\partial y}y'''. \end{aligned} \tag{4.3.20}$$

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. Il suffit juste de prendre une bonne inspiration et de plonger.

29. Cela revient en fait plus ou moins au même.

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

$$f(t+h\alpha, y+h\beta) = f + h\left[\frac{\partial f}{\partial t}\alpha + \frac{\partial f}{\partial y}\beta\right] + \frac{h^2}{2}\left[\frac{\partial^2 f}{\partial t^2}\alpha^2 + 2\frac{\partial^2 f}{\partial t\partial y}\alpha\beta + \frac{\partial^2 f}{\partial y^2}\beta^2\right] + \frac{h^3}{6}\left[\frac{\partial^3 f}{\partial t^3}\alpha^3 + 3\frac{\partial^3 f}{\partial t^2\partial y}\alpha^2\beta + 3\frac{\partial^3 f}{\partial t\partial y^2}\alpha\beta^2 + \frac{\partial^3 f}{\partial y^3}\beta^3\right] + O(h^4). \tag{4.3.21}$$

On a donc d'abord $e_{n,1} = y$. Là c'est facile avec $\alpha = \beta = 0$,

$$f(t, e_{n,1}) = f = y' \tag{4.3.22}$$

pour le premier terme à développer. Ensuite vient par définition du schéma

$$e_{n,2} = y + \frac{h}{2}y', \tag{4.3.23}$$

d'où, prenant $\alpha = \frac{1}{2}$, $\beta = \frac{y'}{2}$,

$$\begin{aligned} f\left(t+\frac{h}{2}, e_{n,2}\right) &= f + \frac{h}{2}\left[\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}y'\right] + \frac{h^2}{8}\left[\frac{\partial^2 f}{\partial t^2} + 2\frac{\partial^2 f}{\partial t\partial y}y' + \frac{\partial^2 f}{\partial y^2}(y')^2\right] \\ &\quad + \frac{h^3}{48}\left[\frac{\partial^3 f}{\partial t^3} + 3\frac{\partial^3 f}{\partial t^2\partial y}y' + 3\frac{\partial^3 f}{\partial t\partial y^2}(y')^2 + \frac{\partial^3 f}{\partial y^3}(y')^3\right] + O(h^4) \\ &= y' + \frac{h}{2}y'' + \frac{h^2}{8}\left[y''' - \frac{\partial f}{\partial y}y''\right] \\ &\quad + \frac{h^3}{48}\left[y^{(4)} - 3\frac{\partial^2 f}{\partial t\partial y}y'' - 3\frac{\partial^2 f}{\partial y^2}y'y'' - \frac{\partial f}{\partial y}y'''\right] + O(h^4). \end{aligned} \tag{4.3.24}$$

Par conséquent,

$$e_{n,3} = y + \frac{h}{2}y' + \frac{h^2}{4}y'' + \frac{h^3}{16}\left[y''' - \frac{\partial f}{\partial y}y''\right] + O(h^4), \tag{4.3.25}$$

(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 $\alpha = \frac{1}{2}$, $\beta = \frac{1}{2}\left(y' + \frac{h}{2}y'' + \frac{h^2}{8}\left[y''' - \frac{\partial f}{\partial y}y''\right]\right) + O(h^3)$,

$$\begin{aligned} f\left(t+\frac{h}{2}, e_{n,3}\right) &= f + \frac{h}{2}\left[\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}\left(y' + \frac{h}{2}y'' + \frac{h^2}{8}\left(y''' - \frac{\partial f}{\partial y}y''\right)\right)\right] \\ &\quad + \frac{h^2}{8}\left[\frac{\partial^2 f}{\partial t^2} + 2\frac{\partial^2 f}{\partial t\partial y}\left(y' + \frac{h}{2}y''\right) + \frac{\partial^2 f}{\partial y^2}\left((y')^2 + hy'y''\right)\right] \end{aligned}$$

(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)

(4.3.26)

Réarrangeons les puissances de h dans cette dernière expression. Il vient

$$\begin{aligned} f\left(t+\frac{h}{2}, e_{n,3}\right) &= f + \frac{h}{2}\left[\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}y'\right] \\ &\quad + \frac{h^2}{8}\left[\frac{\partial^2 f}{\partial t^2} + 2\frac{\partial^2 f}{\partial t \partial y}y' + \frac{\partial^2 f}{\partial y^2}(y')^2 + 2\frac{\partial f}{\partial y}y''\right] \\ &\quad + \frac{h^3}{48}\left[\frac{\partial^3 f}{\partial t^3} + 3\frac{\partial^3 f}{\partial t^2 \partial y}y' + 3\frac{\partial^3 f}{\partial t \partial y^2}(y')^2 + \frac{\partial^3 f}{\partial y^3}(y')^3\right. \\ &\quad \left. + 3\frac{\partial f}{\partial y}\left(y''' - \frac{\partial f}{\partial y}y''\right) + 6\frac{\partial^2 f}{\partial t \partial y}y'' + 6\frac{\partial^2 f}{\partial y^2}y'y''\right] + O(h^4). \end{aligned} \tag{4.3.27}$$

Relisant alors les formules (4.3.20), on en déduit

$$\begin{aligned} f\left(t+\frac{h}{2}, e_{n,3}\right) &= y' + \frac{h}{2}y'' + \frac{h^2}{8}\left[y''' + \frac{\partial f}{\partial y}y''\right] \\ &\quad + \frac{h^3}{48}\left[y^{(4)} + 3\frac{\partial^2 f}{\partial t \partial y}y'' + 3\frac{\partial^2 f}{\partial y^2}y'y'' + 2\frac{\partial f}{\partial y}y''' - 3\left(\frac{\partial f}{\partial y}\right)^2 y''\right] + O(h^4). \end{aligned} \tag{4.3.28}$$

On refait le ménage dans les puissances de h,

$$e_{n,4} = y + hy' + \frac{h^2}{2}y'' + \frac{h^3}{8}\left[y''' + \frac{\partial f}{\partial y}y''\right] + O(h^4). \tag{4.3.29}$$

Il reste un dernier développement de f à faire avec $\alpha = 1$ et $\beta = y' + \frac{h}{2}y'' + \frac{h^2}{8}\left[y''' + \frac{\partial f}{\partial y}y''\right] + O(h^3)$

$$\begin{aligned} f\left(t+h, e_{n,4}\right) &= f + h\left[\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}\left(y' + \frac{h}{2}y'' + \frac{h^2}{8}\left[y''' + \frac{\partial f}{\partial y}y''\right]\right)\right] \\ &\quad + \frac{h^2}{2}\left[\frac{\partial^2 f}{\partial t^2} + 2\frac{\partial^2 f}{\partial t \partial y}\left(y' + \frac{h}{2}y''\right) + \frac{\partial^2 f}{\partial y^2}\left((y')^2 + hy'y''\right)\right] \\ &\quad + \frac{h^3}{6}\left[\frac{\partial^3 f}{\partial t^3} + 3\frac{\partial^3 f}{\partial t^2 \partial y}y' + 3\frac{\partial^3 f}{\partial t \partial y^2}(y')^2 + \frac{\partial^3 f}{\partial y^3}(y')^3\right] + O(h^4) \end{aligned} \tag{4.3.30}$$

Réarrangeons,

$$\begin{aligned} f\left(t+h, e_{n,4}\right) &= f + h\left[\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}y'\right] \\ &\quad + \frac{h^2}{2}\left[\frac{\partial^2 f}{\partial t^2} + 2\frac{\partial^2 f}{\partial t \partial y}y' + \frac{\partial^2 f}{\partial y^2}(y')^2 + \frac{\partial f}{\partial y}y''\right] \\ &\quad + \frac{h^3}{6}\left[\frac{\partial^3 f}{\partial t^3} + 3\frac{\partial^3 f}{\partial t^2 \partial y}y' + 3\frac{\partial^3 f}{\partial t \partial y^2}(y')^2 + \frac{\partial^3 f}{\partial y^3}(y')^3\right. \\ &\quad \left. + \frac{3}{4}\frac{\partial f}{\partial y}\left(y''' + \frac{\partial f}{\partial y}y''\right) + 3\left(\frac{\partial^2 f}{\partial t \partial y}y'' + \frac{\partial^2 f}{\partial y^2}y'y''\right)\right] + O(h^4). \end{aligned} \tag{4.3.31}$$

Relisons la liste des dérivées de y,

$$f\left(t+h, e_{n,4}\right) = y' + hy'' + \frac{h^2}{2}y''' + \frac{h^3}{6}\left[y^{(4)} - \frac{1}{4}\frac{\partial f}{\partial y}y''' + \frac{3}{4}\left(\frac{\partial f}{\partial y}\right)^2 y''\right] + O(h^4). \tag{4.3.32}$$

Nous pouvons maintenant combiner les développements (4.3.22), (4.3.24), (4.3.28) et (4.3.32) pour obtenir

$$\begin{aligned}
&f(t,e_{n,1})+2f\left(t+\frac{h}{2},e_{n,2}\right)+2f\left(t+\frac{h}{2},e_{n,3}\right)+f(t+h,e_{n,4})=\\
&\quad y'+2\left(y'+\frac{h}{2}y''+\frac{h^2}{8}\left[y'''-\frac{\partial f}{\partial y}y''\right]+\frac{h^3}{48}\left[y^{(4)}-3\frac{\partial^2 f}{\partial t\partial y}y''-3\frac{\partial^2 f}{\partial y^2}y'y''-\frac{\partial f}{\partial y}y'''\right]\right)\\
&+2\left(y'+\frac{h}{2}y''+\frac{h^2}{8}\left[y'''+\frac{\partial f}{\partial y}y''\right]+\frac{h^3}{48}\left[y^{(4)}+3\frac{\partial^2 f}{\partial t\partial y}y''+3\frac{\partial^2 f}{\partial y^2}y'y''+2\frac{\partial f}{\partial y}y'''-3\left(\frac{\partial f}{\partial y}\right)^2y''\right]\right)\\
&\quad +y'+hy''+\frac{h^2}{2}y'''+\frac{h^3}{6}\left[y^{(4)}-\frac{1}{4}\frac{\partial f}{\partial y}y'''+\frac{3}{4}\left(\frac{\partial f}{\partial y}\right)^2y''\right]+O(h^4)\\
&\quad =6y'+3hy''+h^2y'''+\frac{h^3}{4}y^{(4)}+O(h^4). \qquad (4.3.33)
\end{aligned}$$

On obtient donc l'erreur de consistance

$$\varepsilon_n=y(t_{n+1})-y(t_n)-hy'(t_n)-\frac{h^2}{2}y''(t_n)-\frac{h^3}{6}y'''(t_n)-\frac{h^4}{24}y^{(4)}(t_n)+O(h^5)=O(h^5), \qquad (4.3.34)$$

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 4.3.2 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). 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 4.3.9 *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 4.3.7, puisque $\rho(|A|) \leq \|A\|_\infty$ pour tout matrice A.

Proposition 4.3.10 *Une condition nécessaire et suffisante pour qu'une méthode de Runge-Kutta soit au moins d'ordre* 1 *(ou consistante) s'écrit* $b^T e = 1$.

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.

On a vu que les méthodes de Runge-Kutta sont à un pas, d'ordre élevé mais coûteuses, alors que les méthodes d'Adams sont à pas multiples, d'ordre élevé et économiques. Une stratégie possible lorsque l'on envisage des calculs de grande envergure, comme des calculs astronomiques 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.

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 (4.2.2). [30] Pour le schéma de RK4, on a (le vérifier)

$$y_{n+1} = \left[1 + \lambda h + \frac{(\lambda h)^2}{2} + \frac{(\lambda h)^3}{6} + \frac{(\lambda h)^4}{24}\right] y_n,$$

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 4.10.

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.

30. 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.

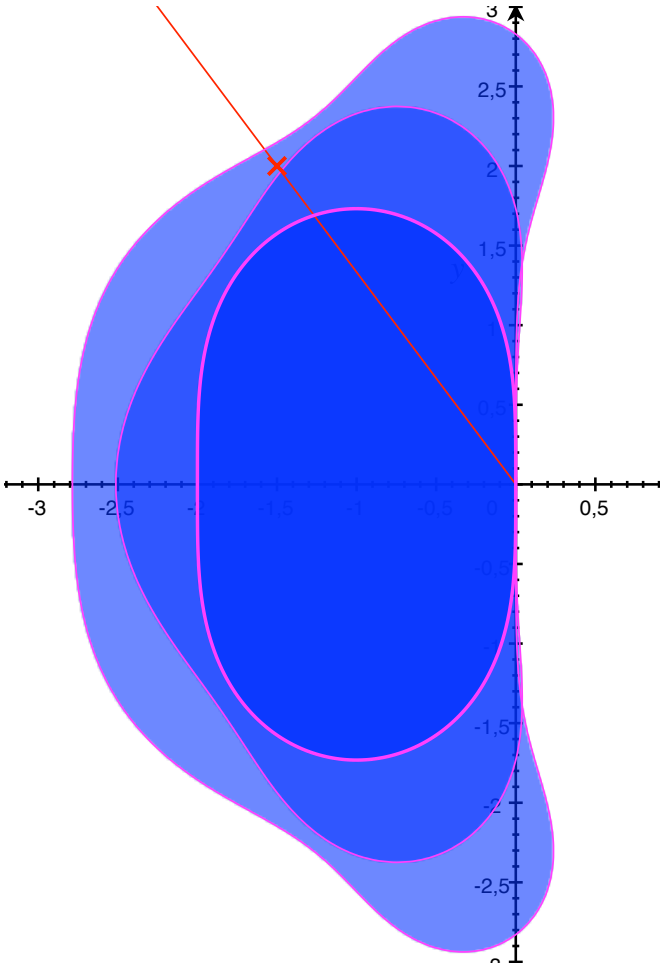


FIGURE 4.10 – Domaines de stabilité absolue des schémas RK2, RK3 et RK4. Seules les intersections avec le demi-plan $\Re(z) < 0$ sont pertinentes.

4.4 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 1.2.4 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 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 [31]

$$J = \begin{pmatrix} 0_d & I_d \\ -I_d & 0_d \end{pmatrix},$$

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 4.4.1 *Une matrice $A \in M_{2d}(\mathbb{R})$ est dite* symplectique *si elle vérifie la relation*

$$A^T J A = J.$$

Proposition 4.4.2 *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

$$\begin{cases} A_{11}^T A_{21} = A_{21}^T A_{11} \\ A_{12}^T A_{22} = A_{22}^T A_{12} \\ A_{11}^T A_{22} - A_{21}^T A_{12} = I_d \end{cases} \tag{4.4.1}$$

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 (4.4.1) sont évidemment satisfaites. La troisième n'est autre dans ce cas que $\det A = 1$. ◊

Remarquons que les deux premières relations de (4.4.1) sont équivalentes à dire que les matrices $A_{11}^T A_{21}$ et $A_{12}^T A_{22}$ doivent être symétriques.

Définition 4.4.3 *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 4.4.4 *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

$$(AB)^T J A B = B^T (A^T J A) B = B^T J B = J,$$

trivialement. ◊

31. On l'a déjà rencontrée sous forme d'opérateur page 105.

Remarque 4.4.1 Remarquons que le déterminant d'une matrice symplectique n'est pas nul. En effet, $\det J = 1$, donc $(\det A)^2 = 1$. [32] 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

$$\begin{cases} \begin{pmatrix} \dot{q}(t) \\ \dot{p}(t) \end{pmatrix} = \begin{pmatrix} \nabla_p T(p(t)) \\ -\nabla_q V(q(t)) \end{pmatrix}, \\ \begin{pmatrix} q(0) \\ p(0) \end{pmatrix} = \begin{pmatrix} q_0 \\ p_0 \end{pmatrix} \in \mathbb{R}^{2d}, \end{cases} \tag{4.4.2}$$

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.

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 (4.4.2). 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 4.4.5 *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, $\varphi_0 = id$). [33]

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

$$\begin{aligned} \frac{d}{dt}\big(\nabla\varphi_t^T J\nabla\varphi_t\big) &= \frac{d}{dt}\big(\nabla\varphi_t\big)^T J\nabla\varphi_t + \nabla\varphi_t^T J\frac{d}{dt}\big(\nabla\varphi_t\big) \\ &= \big(\nabla f(y(t))\nabla\varphi_t\big)^T J\nabla\varphi_t + \nabla\varphi_t^T J\big(\nabla f(y(t))\nabla\varphi_t\big) \\ &= \nabla\varphi_t^T\big(\nabla f(y(t))^T J + J\nabla f(y(t))\big)\nabla\varphi_t \end{aligned}$$

32. 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.

33. Voir une démonstration en annexe à titre culturel.

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 $\frac{d}{dt}\left(\nabla\varphi_t^T J\nabla\varphi_t\right) = 0$, 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

$$\begin{cases} q_{n+1} = q_n + h\nabla_p T(p_n), \\ p_{n+1} = p_n - h\nabla_q V(q_{n+1}), \end{cases} \tag{4.4.3}$$

que l'on note

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \Phi_h \begin{pmatrix} q_n \\ p_n \end{pmatrix},$$

avec

$$\Phi_h \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} q + h\nabla_p T(p) \\ p - h\nabla_q V(q + h\nabla_p T(p)) \end{pmatrix}. \tag{4.4.4}$$

Le schéma de Störmer-Verlet [34] défini par

$$\begin{cases} p_{n+\frac{1}{2}} = p_n - \dfrac{h}{2}\nabla_q V(q_n), \\ q_{n+1} = q_n + h\nabla_p T(p_{n+\frac{1}{2}}), \\ p_{n+1} = p_{n+\frac{1}{2}} - \dfrac{h}{2}\nabla_q V(q_{n+1}), \end{cases} \tag{4.4.5}$$

soit

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \Psi_h \begin{pmatrix} q_n \\ p_n \end{pmatrix},$$

avec

$$\Psi_h \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} q + h\nabla_p T\left(p - \frac{h}{2}\nabla_q V(q)\right) \\ p - \frac{h}{2}\left[\nabla_q V(q) + \nabla_q V\left(q + h\nabla_p T\left(p - \frac{h}{2}\nabla_q V(q)\right)\right)\right] \end{pmatrix}. \tag{4.4.6}$$

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 4.4.6 *Les applications Φ_h et Ψ_h sont symplectiques.*

Démonstration. On part de (4.4.4) pour calculer la matrice jacobienne de Φ_h,

$$\nabla\Phi_h = \begin{pmatrix} 1 & hT''(p) \\ -hV''(q + hT'(p)) & 1 - h^2 V''(q + hT'(p))T''(p) \end{pmatrix}.$$

34. Fredrik Carl Mülertz Størmer, 1874–1957 ; Loup Verlet, 1931–.

Or on a vu à la proposition 4.4.2 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 (4.4.6), 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 4.4.4 pour conclure. ◊

Corollaire 4.4.7 *Pour les deux schémas, l'application $(q_0, p_0) \mapsto (q_n, p_n)$ est symplectique pour tout n.*

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 4.4.4, 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. [35] Les applications Φ_h^n et Ψ_h^n sont les flots numériques des deux schémas, voir Figures 4.11 et 4.12.

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 4.4.8 *Le schéma d'Euler symplectique est d'ordre 1 et le schéma de Störmer-Verlet est d'ordre 2.*

35. 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.

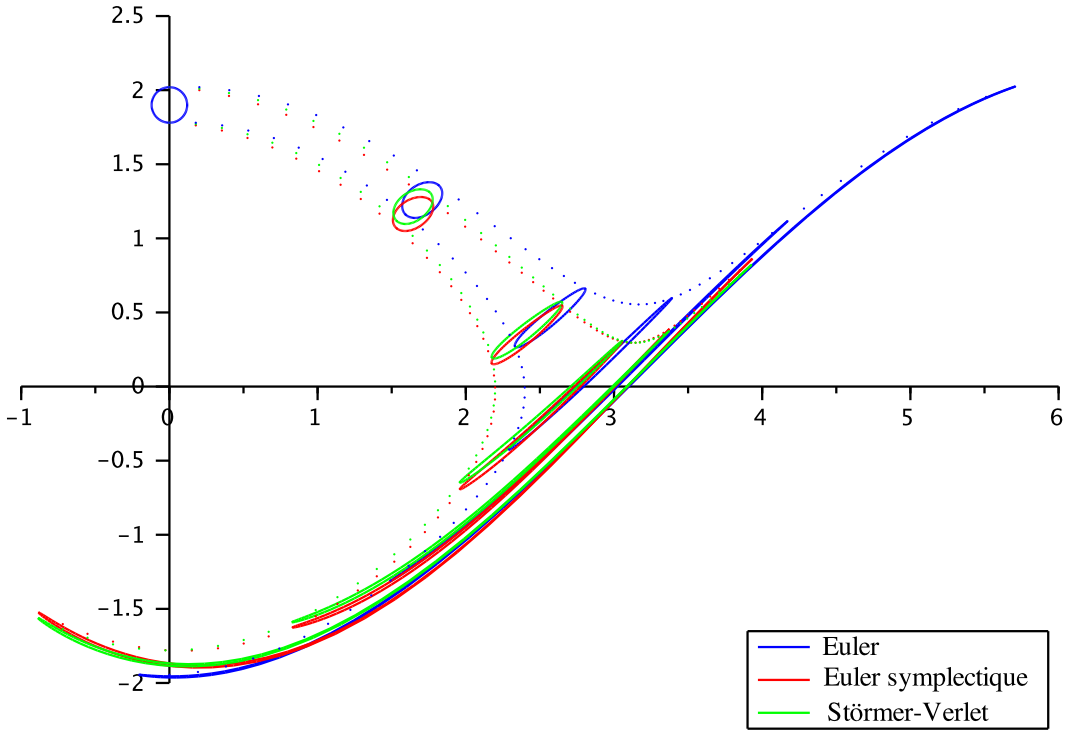


FIGURE 4.11 – Le flot approché des grandes oscillations du pendule. On a tracé les images par les flots numériques des schémas d'Euler, d'Euler symplectique et de Störmer-Verlet avec un pas de temps $h = 0.1$, du disque centré en $(0, 1.9)$ et de rayon 0.12, aux temps $t = 1, 2, 3, 4$ et 5. On a également 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.

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

$$\begin{aligned}q(t_{n+1}) - q(t_n) - hT'\Big(p(t_n) - \frac{h}{2}V'(q(t_n))\Big) &= h\dot{q}(t_n) + \frac{h^2}{2}\ddot{q}(t_n) + O(h^3) \\ &\quad - hT'(p(t_n)) + \frac{h^2}{2}T''(p(t_n))V'(q(t_n)) + O(h^3) \\ &= O(h^3),\end{aligned}$$

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

$$T'\big(\tilde{p}_{n+\frac{1}{2}}\big) = T'(p(t_n)) + O(h).$$

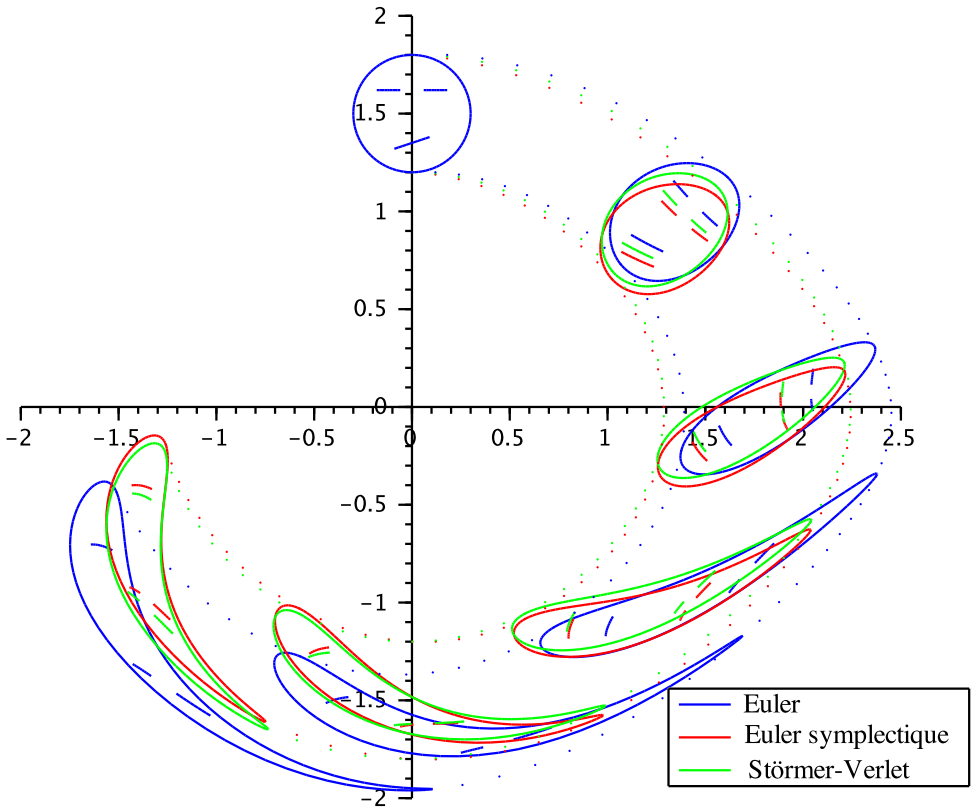


FIGURE 4.12 – Le même calcul avec le disque centré en $(0, 1.5)$ et de rayon 0.3 moins quelques traits. On voit assez clairement l'augmentation du volume causée par le schéma d'Euler.

On obtient donc

$$\begin{aligned} p(t_{n+1}) - \tilde{p}_{n+\frac{1}{2}} + \frac{h}{2}V'\big(q(t_n) + hT'\big(\tilde{p}_{n+\frac{1}{2}}\big)\big) &= h\dot{p}(t_n) + \frac{h^2}{2}\ddot{p}(t_n) + O(h^3) \\ &\quad + hV'(q(t_n)) + \frac{h^2}{2}V''(q(t_n))T'(p(t_n)) + O(h^3) \\ &= O(h^3), \end{aligned}$$

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 2.10, 2.11 et 2.12.

Annexe : différentielle du flot

On établit ici la formule donnant la différentielle du flot.[36] 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

$$\frac{d}{dt}\big(\nabla\varphi_t\big) = \nabla f(\varphi_t)\nabla\varphi_t, \quad \nabla\varphi_0 = I.$$

36. 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.

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

$$z_h(t) = \varphi_t(y_0 + h) - \varphi_t(y_0) - A(t)h,$$

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 1.4.3. 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. ◊

4.5 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 $(y_n)_{n=0,\ldots,N}$ mais des valeurs perturbées par l'erreur d'arrondi ρ_n commise à chaque pas de temps [37]

$$y_{n+1}^\star = y_n^\star + hf(t_n, y_n^\star) + \rho_n.$$

37. 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.

Posons $e_n^\star = y_n^\star - y(t_n)$. On a

$$e_{n+1}^\star = e_n^\star + h\left(f(t_n, y_n^\star) - f(t_n, y(t_n))\right) - \varepsilon_n + \rho_n$$

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 $|\rho_n| \leq \rho$,

$$|e_{n+1}^\star| \leq (1 + hL)|e_n^\star| + \frac{h^2}{2} \max |y''| + \rho,$$

d'où l'on déduit, en utilisant le lemme de Grönwall discret, que

$$|e_n^\star| \leq e^{nhL}|e_0^\star| + \frac{h}{2}(nhL)e^{nhL} \max |y''| + nLe^{nhL}\rho.$$

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 4.13 illustre ce comportement pour le schéma de Taylor d'ordre 2 et le schéma de Runge-Kutta d'ordre 4. [38] 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

38. 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.

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

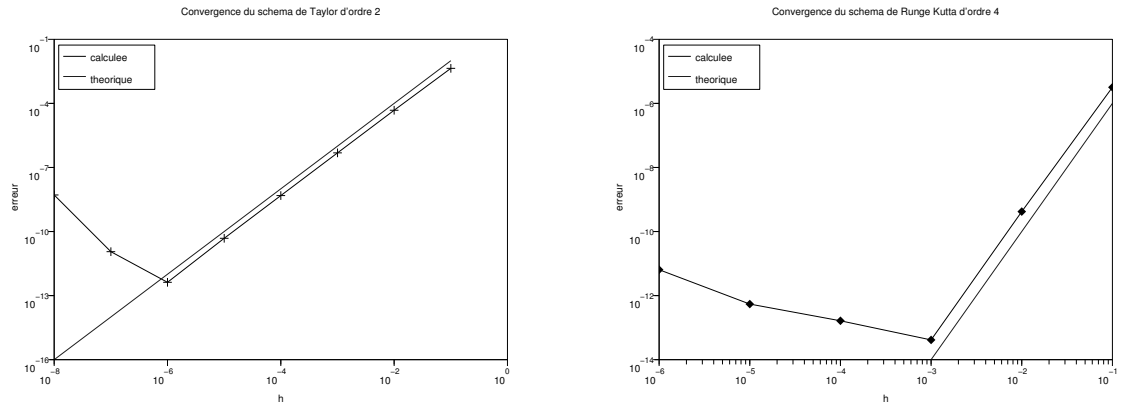


FIGURE 4.13 – Convergence de l'erreur pour les schémas de Taylor d'ordre 2 et de Runge-Kutta d'ordre 4

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.

4.5.1 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

$$\|y_{n+1}^1 - y(t_{n+1})\| \leq C_1 h^{p_1}$$
$$\|y_{n+1}^2 - y(t_{n+1})\| \leq C_2 h^{p_2}.$$

On a donc

$$y_{n+1}^1 - y(t_{n+1}) = y_{n+1}^1 - y_{n+1}^2 + y_{n+1}^2 - y(t_{n+1}),$$

d'où l'on déduit en supposant h petit et la constante C_2 pas trop méchante que

$$\|y_{n+1}^1 - y(t_{n+1})\| \approx \|y_{n+1}^1 - y_{n+1}^2\|.$$

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_n^1 et y_n^2 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_n^1 et y_n^2 avec le pas h_*.
Si $\|y_*^1 - y_*^2\| \leq Tol$, on prend $h_n = h_*$, $y_{n+1}^1 = y_*^1$ et $y_{n+1}^2 = y_*^2$
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).

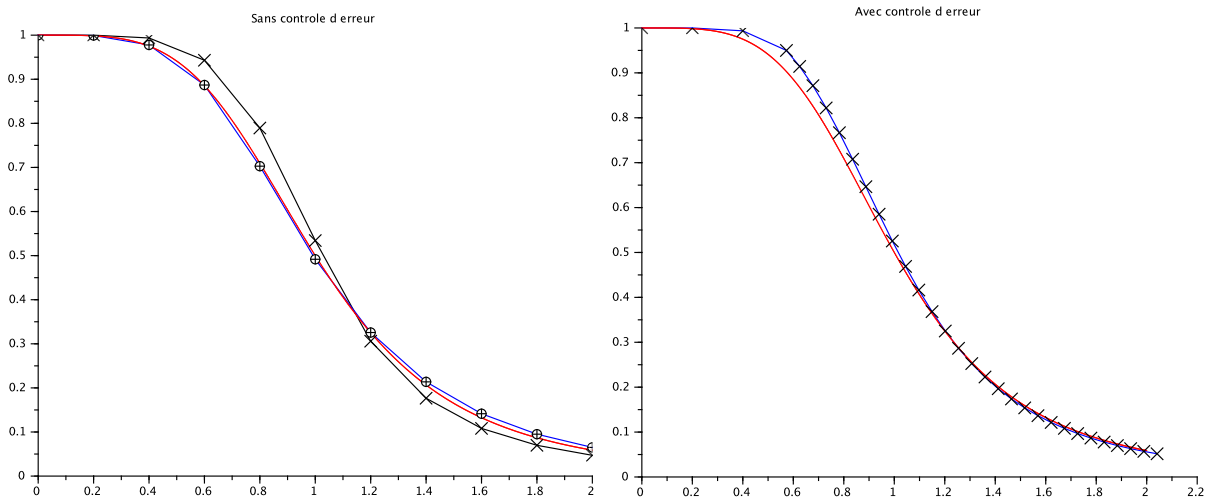


FIGURE 4.14 – Algorithme de contrôle naïf. À gauche pas constant sans contrôle : noir × Euler, bleu ⊕ Euler modifié, rouge solution exacte. À droite pas variable avec contrôle : noir × Euler, rouge solution exacte.

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 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

$$\begin{aligned}\varepsilon_n - \varepsilon_n^* &= h_n\big(\Phi^*(h_n,t_n,y(t_n)) - \Phi(h_n,t_n,y(t_n))\big),\\ &= \tilde{\varepsilon}_n + h_n\frac{\partial(\Phi^*-\Phi)}{\partial y}(h_n,t_n,\zeta_n)(y(t_n)-y_n),\end{aligned}$$

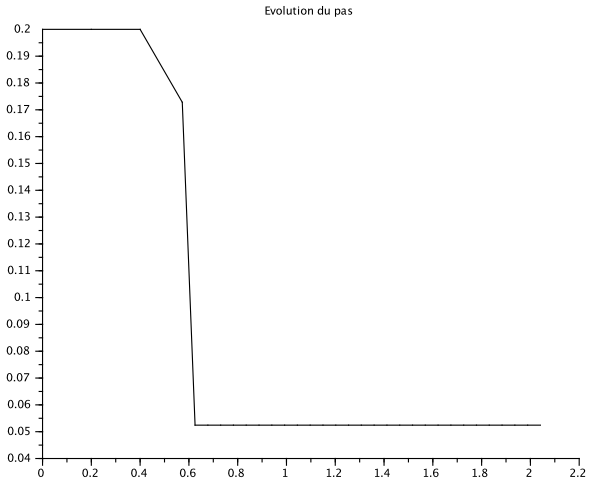


FIGURE 4.15 – Algorithme de contrôle naïf, variation du pas en fonction du temps.

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

$$\begin{cases} k_1 = f(t_n, y_n), \\ k_2 = f(t_{n+1}, y_n + h_n k_1), \\ y_{n+1}^{\mathrm{rk2}} = y_n + \dfrac{h_n}{2}(k_1 + k_2), \end{cases} \tag{4.5.1}$$

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*)

$$\begin{cases} k_1 = f(t_n, y_n), \\ k_2 = f(t_{n+1}, y_n + h_n k_1), \\ k_3 = f\left(t_n + \dfrac{h_n}{2}, y_n + \dfrac{h_n}{4}(k_1 + k_2)\right), \\ y_{n+1}^{\mathrm{rk3}} = y_n + \dfrac{h_n}{6}(k_1 + k_2 + 4k_3). \end{cases} \tag{4.5.2}$$

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 4.16 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 4.17. 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.

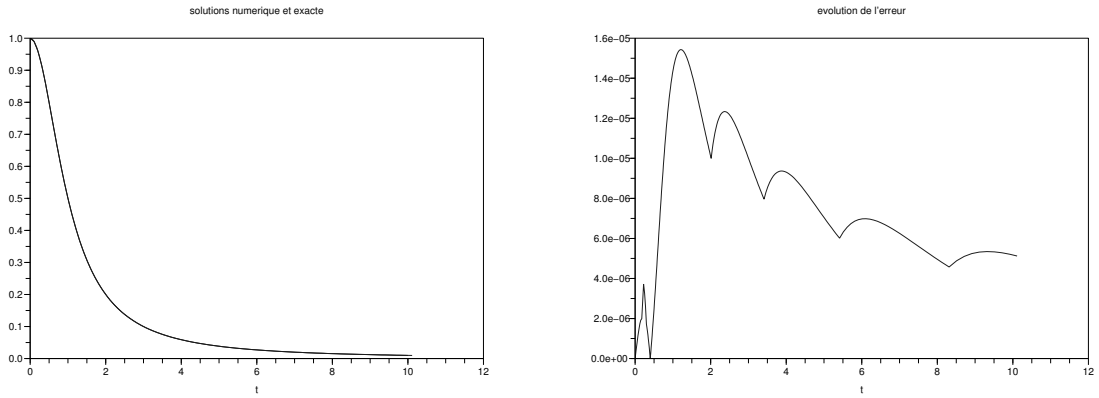


FIGURE 4.16 – Schéma RK23 avec contrôle adaptatif du pas, solution à gauche et erreur entre les solutions numériques et exactes à droite

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.

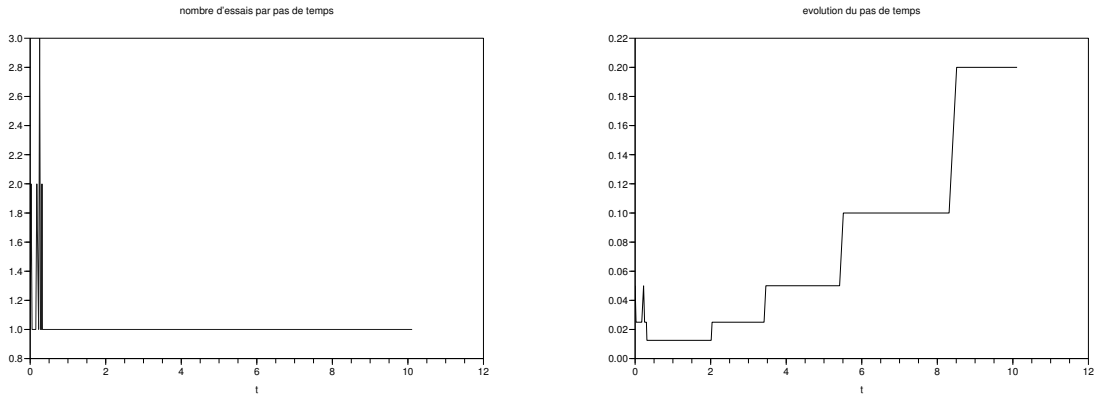


FIGURE 4.17 – Schéma RK23 avec contrôle adaptatif du pas, nombre d'essais par pas de temps à gauche et pas de temps à droite

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