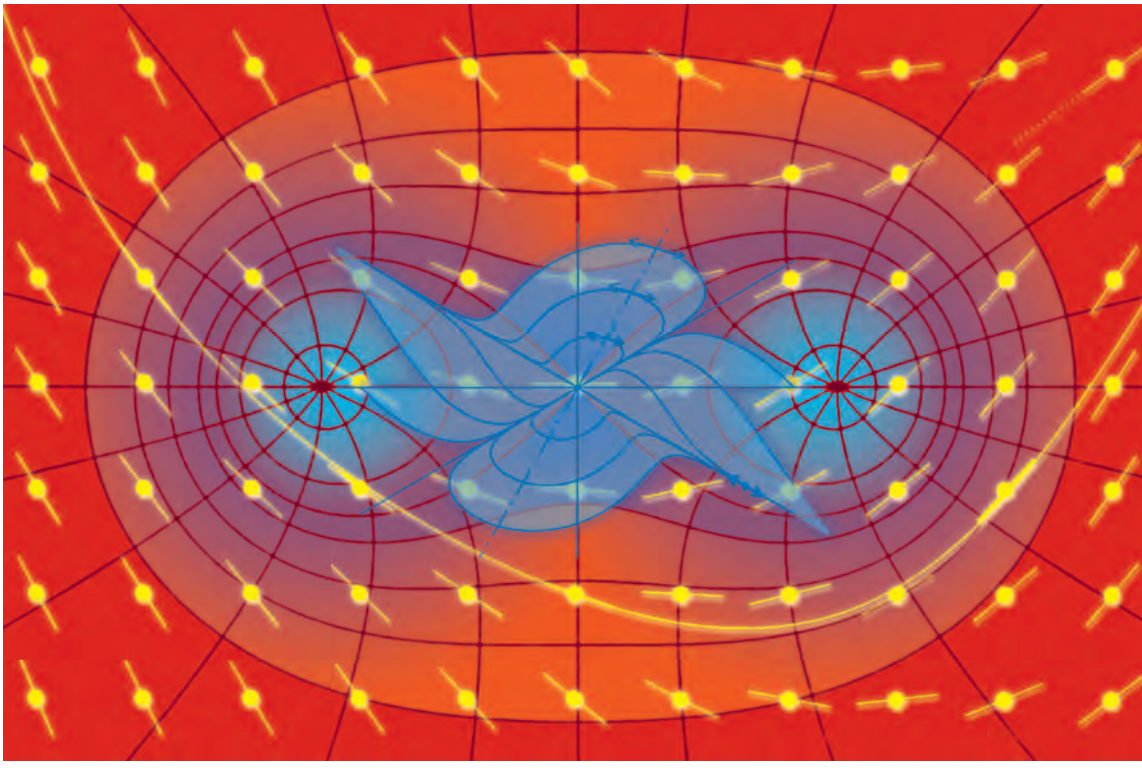
COLLECTION GRENOBLE SCIENCES
DIRIGÉE PAR JEAN BORNAREL

ANALYSE NUMÉRIQUE ET ÉQUATIONS DIFFÉRENTIELLES

Nouvelle édition

Jean-Pierre DEMAILLY



EDP SCIENCES

ANALYSE NUMÉRIQUE
ET
ÉQUATIONS DIFFÉRENTIELLES

Grenoble Sciences

Grenoble Sciences poursuit un triple objectif :

- réaliser des ouvrages correspondant à un projet clairement défini, sans contrainte de mode ou de programme,
- garantir les qualités scientifique et pédagogique des ouvrages retenus,
- proposer des ouvrages à un prix accessible au public le plus large possible.

Chaque projet est sélectionné au niveau de Grenoble Sciences avec le concours de referees anonymes. Puis les auteurs travaillent pendant une année (en moyenne) avec les membres d'un comité de lecture interactif, dont les noms apparaissent au début de l'ouvrage. Celui-ci est ensuite publié chez l'éditeur le plus adapté.

(Contact : Tél. : (33)4 76 51 46 95 - E-mail : Grenoble.Sciences@ujf-grenoble.fr)

Deux collections existent chez EDP Sciences :

- la ***Collection Grenoble Sciences***, connue pour son originalité de projets et sa qualité
- ***Grenoble Sciences - Rencontres Scientifiques***, collection présentant des thèmes de recherche d'actualité, traités par des scientifiques de premier plan issus de disciplines différentes.

Grenoble Sciences bénéficie du soutien du **Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche** et de la **Région Rhône-Alpes**.
Grenoble Sciences est rattaché à l'**Université Joseph Fourier de Grenoble**.

Illustration de couverture : Alice GIRAUD

ISBN 2-86883-891-X

ANALYSE NUMÉRIQUE
ET
ÉQUATIONS DIFFÉRENTIELLES

Jean-Pierre DEMAILLY

17, avenue du Hoggar
Parc d'Activité de Courtabœuf - BP 112
91944 Les Ulis Cedex A - France

Ouvrages Grenoble Sciences édités par EDP Sciences

Collection Grenoble Sciences

Chimie. Le minimum à savoir *(J. Le Coarer)* • Electrochimie des solides *(C. Déportes et al.)* • Thermodynamique chimique *(M. Oturan & M. Robert)* • CD de Thermodynamique chimique *(J.P. Damon & M. Vincens)* • Chimie organométallique *(D. Astruc)* • De l'atome à la réaction chimique *(sous la direction de R. Barlet)*

Introduction à la mécanique statistique *(E. Belorizky & W. Gorecki)* • Mécanique statistique. Exercices et problèmes corrigés *(E. Belorizky & W. Gorecki)* • La cavitation. Mécanismes physiques et aspects industriels *(J.P. Franc et al.)* • La turbulence *(M. Lesieur)* • Magnétisme : I Fondements, II Matériaux et applications *(sous la direction d'E. du Trémolet de Lacheisserie)* • Du Soleil à la Terre. Aéronomie et météorologie de l'espace *(J. Lilensten & P.L. Blelly)* • Sous les feux du Soleil. Vers une météorologie de l'espace *(J. Lilensten & J. Bornarel)* • Mécanique. De la formulation lagrangienne au chaos hamiltonien *(C. Gignoux & B. Silvestre-Brac)* • Problèmes corrigés de mécanique et résumés de cours. De Lagrange à Hamilton *(C. Gignoux & B. Silvestre-Brac)* • La mécanique quantique. Problèmes résolus, T. 1 et 2 *(V.M. Galitsky, B.M. Karnakov & V.I. Kogan)* • Description de la symétrie. Des groupes de symétrie aux structures fractales *(J. Sivardière)* • Symétrie et propriétés physiques. Du principe de Curie aux brisures de symétrie *(J. Sivardière)*

Exercices corrigés d'analyse, T. 1 et 2 *(D. Alibert)* • Introduction aux variétés différentielles *(J. Lafontaine)* Mathématiques pour les sciences de la vie, de la nature et de la santé *(F. & J.P. Bertrandias)* • Approximation hilbertienne. Splines, ondelettes, fractales *(M. Attéia & J. Gaches)* • Mathématiques pour l'étudiant scientifique, T. 1 et 2 *(Ph.J. Haug)* • Analyse statistique des données expérimentales *(K. Protassov)* • Nombres et algèbre *(J.Y. Mérindol)*

Bactéries et environnement. Adaptations physiologiques *(J. Pelmont)* • Enzymes. Catalyseurs du monde vivant *(J. Pelmont)* • Endocrinologie et communications cellulaires *(S. Idelman & J. Verdetti)* • Eléments de biologie à l'usage d'autres disciplines *(P. Tracqui & J. Demongeot)* • Bioénergétique *(B. Guérin)* • Cinétique enzymatique *(A. Cornish-Bowden, M. Jamin & V. Saks)* • Biodégradations et métabolismes. Les bactéries pour les technologies de l'environnement *(J. Pelmont)* • Enzymologie moléculaire et cellulaire, T. 1 et 2 *(J. Yon-Kahn & G. Hervé)*

La plongée sous-marine à l'air. L'adaptation de l'organisme et ses limites *(Ph. Foster)* • L'Asie, source de sciences et de techniques *(M. Soutif)* • La biologie, des origines à nos jours *(P. Vignais)* • Naissance de la physique. De la Sicile à la Chine *(M. Soutif)* • Le régime oméga 3. Le programme alimentaire pour sauver notre santé *(A. Simopoulos, J. Robinson, M. de Lorgeril & P. Salen)* • Gestes et mouvements justes. Guide de l'ergomotricité pour tous *(M. Gendrier)* • Science expérimentale et connaissance du vivant. La méthode et les concepts *(P. Vignais, avec la collaboration de P. Vignais)*

Listening Comprehension for Scientific English *(J. Upjohn)* • Speaking Skills in Scientific English *(J. Upjohn, M.H. Fries & D. Amadis)* • Minimum Competence in Scientific English *(S. Blattes, V. Jans & J. Upjohn)*

Grenoble Sciences - Rencontres Scientifiques

Radiopharmaceutiques. Chimie des radiotraceurs et applications biologiques *(sous la direction de M. Comet & M. Vidal)* • Turbulence et déterminisme *(sous la direction de M. Lesieur)* • Méthodes et techniques de la chimie organique *(sous la direction de D. Astruc)* • L'énergie de demain. Techniques, environnement, économie *(sous la direction de J.L. Bobin, E. Huffer & H. Nifenecker)* • Physique et biologie. Une interdisciplinarité complexe *(sous la direction de B. Jacrot)*

Introduction

Le présent ouvrage reprend avec beaucoup de compléments un cours de "Licence de Mathématiques" – ex troisième année d'Université – donné à l'Université de Grenoble I pendant les années 1985-88. Le but de ce cours était de présenter aux étudiants quelques notions théoriques de base concernant les équations et systèmes d'équations différentielles ordinaires, tout en explicitant des méthodes numériques permettant de résoudre effectivement de telles équations. C'est pour cette raison qu'une part importante du cours est consacrée à la mise en place d'un certain nombre de techniques fondamentales de l'Analyse Numérique : interpolation polynomiale, intégration numérique, méthode de Newton à une et plusieurs variables.

L'originalité de cet ouvrage ne réside pas tant dans le contenu, pour lequel l'auteur s'est inspiré sans vergogne de la littérature existante – en particulier du livre de Crouzeix-Mignot pour ce qui concerne les méthodes numériques, et des livres classiques de H. Cartan et J. Dieudonné pour la théorie des équations différentielles – mais plutôt dans le choix des thèmes et dans la présentation. S'il est relativement facile de trouver des ouvrages spécialisés consacrés soit aux aspects théoriques fondamentaux de la théorie des équations différentielles et ses applications (Arnold, Coddington-Levinson) soit aux techniques de l'Analyse Numérique (Henrici, Hildebrand), il y a relativement peu d'ouvrages qui couvrent simultanément ces différents aspects et qui se situent à un niveau accessible pour l'«honnête» étudiant de second cycle. Nous avons en particulier consacré deux chapitres entiers à l'étude des méthodes élémentaires de résolution par intégration explicite et à l'étude des équations différentielles linéaires à coefficients constants, ces questions étant généralement omises dans les ouvrages de niveau plus avancé. Par ailleurs, un effort particulier a été fait pour illustrer les principaux résultats par des exemples variés.

La plupart des méthodes numériques exposées avaient pu être effectivement mises en œuvre par les étudiants au moyen de programmes écrits en Turbo Pascal – à une époque remontant maintenant à la préhistoire de l'informatique. Aujourd'hui, les environnements disponibles sont beaucoup plus nombreux, mais nous recommandons certainement encore aux étudiants d'essayer d'implémenter les algorithmes proposés dans ce livre sous forme de programmes écrits dans des langages de base

comme C ou C++, et particulièrement dans un environnement de programmation libre comme GCC sous GNU/Linux. Bien entendu, il existe des logiciels libres spécialisés dans le calcul numérique qui implémentent les principaux algorithmes utiles sous forme de librairies toutes prêtes – Scilab est l'un des plus connus – mais d'un point de vue pédagogique et dans un premier temps au moins, il est bien plus formateur pour les étudiants de mettre vraiment "la main dans le cambouis" en programmant eux-mêmes les algorithmes. Nous ne citerons pas d'environnements ni de logiciels propriétaires équivalents, parce que ces logiciels dont le fonctionnement intime est inaccessible à l'utilisateur sont contraires à notre éthique scientifique ou éducative, et nous ne souhaitons donc pas en encourager l'usage.

L'ensemble des sujets abordés dans le présent ouvrage dépasse sans aucun doute le volume pouvant être traité en une seule année de cours – même si jadis nous avions pu en enseigner l'essentiel au cours de la seule année de Licence. Dans les conditions actuelles, il nous paraît plus judicieux d'envisager une répartition du contenu sur l'ensemble des deux années du second cycle universitaire. Ce texte est probablement utilisable aussi pour les élèves d'écoles d'ingénieurs, ou comme ouvrage de synthèse au niveau de l'agrégation de mathématiques. Pour guider le lecteur dans sa sélection, les sous-sections de chapitres les plus difficiles ainsi que les démonstrations les plus délicates sont marquées d'un astérisque. Le lecteur pourra trouver de nombreux exemples de tracés graphiques de solutions d'équations différentielles dans le livre d'Artigue-Gautheron : on y trouvera en particulier des illustrations variées des phénomènes qualitatifs étudiés au chapitre X, concernant les points singuliers des champs de vecteurs.

Je voudrais ici remercier mes collègues grenoblois pour les remarques et améliorations constantes suggérées tout au long de notre collaboration pendant les trois années qu'a duré ce cours. Mes plus vifs remerciements s'adressent également à Michèle Artigue, Alain Dufresnoy, Jean-René Joly et Marc Rogalski, qui ont bien voulu prendre de leur temps pour relire le manuscrit original de manière très détaillée. Leurs critiques et suggestions ont beaucoup contribué à la mise en forme définitive de cet ouvrage.

Saint-Martin d'Hères, le 5 novembre 1990

La seconde édition de cet ouvrage a bénéficié d'un bon nombre de remarques et de suggestions proposées par Marc Rogalski. Les modifications apportées concernent notamment le début du chapitre VIII, où la notion délicate d'erreur de consistance a été plus clairement explicitée, et les exemples des chapitres VI et XI traitant du mouvement du pendule simple. L'auteur tient à remercier de nouveau Marc Rogalski pour sa précieuse contribution.

Saint-Martin d'Hères, le 26 septembre 1996

La troisième édition de cet ouvrage a été enrichie d'un certain nombre de compléments théoriques et pratiques : comportement géométrique des suites itératives en dimension 1, théorème des fonctions implicites et ses variantes géométriques dans le chapitre IV ; critère de maximalité des solutions dans le chapitre V ; calcul de géodésiques dans le chapitre VI ; quelques exemples et exercices additionnels dans les chapitres suivants ; notions élémentaires sur les flots de champs de vecteurs dans le chapitre XI.

Saint-Martin d'Hères, le 28 février 2006

CHAPITRE I

CALCULS NUMÉRIQUES APPROCHÉS

L'objet de ce chapitre est de mettre en évidence les principales difficultés liées à la pratique des calculs numériques sur ordinateur. Dans beaucoup de situations, il existe des méthodes spécifiques permettant d'accroître à la fois l'efficacité et la précision des calculs.

1. CUMULATION DES ERREURS D'ARRONDI

1.1. REPRÉSENTATION DÉCIMALE APPROCHÉE DES NOMBRES RÉELS

La capacité mémoire d'un ordinateur est par construction finie. Il est donc nécessaire de représenter les nombres réels sous forme approchée. La notation la plus utilisée à l'heure actuelle est la représentation avec virgule flottante : un nombre réel x est codé sous la forme

$$x \simeq \pm\, m \cdot b^p$$

où b est la *base de numération*, m la *mantisse*, et p l'exposant. Les calculs internes sont généralement effectués en base $b = 2$, même si les résultats affichés sont finalement traduits en base 10.

La mantisse m est un nombre écrit avec virgule fixe et possédant un nombre maximum N de chiffres significatifs (imposé par le choix de la taille des emplacements mémoires alloués au type *réel*) : suivant les machines, m s'écrira

- $m = 0, a_1 a_2 \dots a_N = \displaystyle\sum_{k=1}^{N} a_k b^{-k}, \qquad b^{-1} \leq m < 1$;
- $m = a_0, a_1 a_2 \dots a_{N-1} = \displaystyle\sum_{0 \leq k < N} a_k b^{-k}, \quad 1 \leq m < b.$

Ceci entraîne que la précision dans l'approximation d'un nombre réel est toujours une *précision relative* :

$$\frac{\Delta x}{x} = \frac{\Delta m}{m} \leq \frac{b^{-N}}{b^{-1}} = b^{1-N}.$$

On notera $\varepsilon = b^{1-N}$ cette précision relative.

En Langage C standard (ANSI C), les réels peuvent occuper

– pour le type « float », 4 octets de mémoire, soit 1 bit de signe, 23 bits de mantisse et 8 bits d'exposant (dont un pour le signe de l'exposant). Ceci permet de représenter les réels avec une mantisse de 6 à 7 chiffres significatifs après la virgule, dans une échelle allant de 2^{-128} à 2^{127} soit environ de $10^{-38} = 1\,\mathrm{E}-38$ à $10^{38} = 1\,\mathrm{E}+38$. La précision relative est de l'ordre de 10^{-7}.

– pour le type « double », 8 octets de mémoire, soit 1 bit de signe, 51 bits de mantisse et 12 bits d'exposant (dont un pour le signe de l'exposant). Ceci permet de représenter les réels avec une mantisse de 15 chiffres significatifs après la virgule, dans une échelle allant de 2^{-2048} à 2^{2047} soit environ de $10^{-616} = 1\,\mathrm{E}-616$ à $10^{616} = 1\,\mathrm{E}+616$. La précision relative est de l'ordre de 10^{-15}.

1.2. Non-associativité des opérations arithmétiques

Supposons par exemple que les réels soient calculées avec 3 chiffres significatifs et arrondis à la décimale la plus proche. Soit à calculer la somme $x+y+z$ avec

$$x = 8,22, \quad y = 0,00317, \quad z = 0,00432$$

$$\begin{aligned}
(x+y)+z \text{ donne } : \quad & x+y = 8,22317 \simeq 8,22 \\
& (x+y)+z \simeq 8,22432 \simeq 8,22 \\
x+(y+z) \text{ donne } : \quad & y+z = 0,00749 \\
& x+(y+z) = 8,22749 \simeq 8,23.
\end{aligned}$$

L'addition est donc non associative par suite des erreurs d'arrondi !

1.3. Erreur d'arrondi sur une somme

On se propose d'étudier quelques méthodes permettant de *majorer* les erreurs d'arrondi dues aux opérations arithmétiques.

Soient x, y des nombres réels supposés représentés sans erreur avec N chiffres significatifs :

$$x = 0,a_1a_2\ldots a_N \cdot b^p, \quad b^{-1+p} \le x < b^p$$
$$y = 0,a'_1a'_2\ldots a'_N \cdot b^q, \quad b^{-1+q} \le y < b^q$$

Notons $\Delta(x+y)$ l'erreur d'arrondi commise sur le calcul de $x+y$. Supposons par exemple $p \ge q$. S'il n'y a pas débordement, c'est-à-dire si $x+y < b^p$, le calcul de $x+y$ s'accompagne d'une perte des $p-q$ derniers chiffres de y correspondant aux puissances $b^{-k+q} < b^{-N+p}$; donc $\Delta(x+y) \le b^{-N+p}$, alors que $x+y \ge x \ge b^{-1+p}$. En cas de débordement $x+y \ge b^p$ (ce qui se produit par exemple si $p=q$ et $a_1+a'_1 \ge b$), la décimale correspondant à la puissance b^{-N+p} est elle aussi perdue, d'où $\Delta(x+y) \le b^{1-N+p}$. Dans les deux cas :

$$\Delta(x+y) \le \varepsilon(|x|+|y|),$$

où $\varepsilon = b^{1-N}$ est la précision relative décrite au §1.1. Ceci reste vrai quel que soit le signe des nombres x et y.

En général, les réels x, y ne sont eux-mêmes connus que par des valeurs approchées x', y' avec des erreurs respectives $\Delta x = |x' - x|$, $\Delta y = |y' - y|$. A ces erreurs s'ajoute l'erreur d'arrondi

$$\Delta(x' + y') \leq \varepsilon(|x'| + |y'|) \leq \varepsilon(|x| + |y| + \Delta x + \Delta y).$$

Les erreurs Δx, Δy sont elles-mêmes le plus souvent d'ordre ε par rapport à $|x|$ et $|y|$, de sorte que l'on pourra négliger les termes $\varepsilon\Delta x$ et $\varepsilon\Delta y$. On aura donc :

$$\Delta(x + y) \leq \Delta x + \Delta y + \varepsilon(|x| + |y|).$$

Soit plus généralement à calculer une somme $\sum_{k=1}^{n} u_k$ de réels *positifs*. Les sommes partielles $s_k = u_1 + u_2 + \ldots + u_k$ vont se calculer de proche en proche par les formules de récurrence

$$\begin{cases} s_0 = 0 \\ s_k = s_{k-1} + u_k, \qquad k \geq 1. \end{cases}$$

Si les réels u_k sont connus exactement, on aura sur les sommes s_k des erreurs Δs_k telles que $\Delta s_1 = 0$ et

$$\Delta s_k \leq \Delta s_{k-1} + \varepsilon(s_{k-1} + u_k) = \Delta s_{k-1} + \varepsilon s_k.$$

L'erreur globale sur s_n vérifie donc

$$\Delta s_n \leq \varepsilon(s_2 + s_3 + \ldots + s_n),$$

soit

$$\Delta s_n \leq \varepsilon(u_n + 2u_{n-1} + 3u_{n-2} + \ldots + (n-1)u_2 + (n-1)u_1).$$

Comme ce sont les premiers termes sommés qui sont affectés des plus gros coefficients dans l'erreur Δs_n, on en déduit la règle générale suivante (cf. exemple 1.2).

Règle générale – Dans une sommation de réels, l'erreur a tendance à être minimisée lorsqu'on somme en premier les termes ayant la plus petite valeur absolue.

1.4. Erreur d'arrondi sur un produit

Le produit de deux mantisses de N chiffres donne une mantisse de $2N$ ou $2N - 1$ chiffres dont les N ou $N - 1$ derniers vont être perdus. Dans le calcul d'un produit xy (où x, y sont supposés représentés sans erreur) il y aura donc une erreur d'arrondi

$$\Delta(xy) \leq \varepsilon|xy|, \quad \text{où} \quad \varepsilon = b^{1-N}.$$

Si x et y ne sont eux-mêmes connus que par des valeurs approchées x', y' et si $\Delta x = |x' - x|$, $\Delta y = |y' - y|$, on a une erreur initiale

$$\begin{aligned} |x'y' - xy| = |x(y' - y) + (x' - x)y'| &\leq |x|\Delta y + \Delta x|y'| \\ &\leq |x|\Delta y + \Delta x|y| + \Delta x\Delta y. \end{aligned}$$

A cette erreur s'ajoute une erreur d'arrondi

$$\Delta(x'y') \leq \varepsilon|x'y'| \leq \varepsilon(|x| + \Delta x)(|y| + \Delta y).$$

En négligeant les termes $\Delta x \Delta y$, $\varepsilon \Delta x$, $\varepsilon \Delta y$, on obtient la formule approximative

$$\Delta(xy) \leq |x|\Delta y + \Delta x|y| + \varepsilon|xy|. \qquad (*)$$

Soit plus généralement des réels $x_1, \ldots, x_k$, supposés représentés sans erreur. La formule $(*)$ entraîne

$$\Delta(x_1 x_2 \ldots x_k) \leq \Delta(x_1 \ldots x_{k-1})|x_k| + \varepsilon|x_1 \ldots x_{k-1} \cdot x_k|,$$

d'où par une récurrence aisée :

$$\Delta(x_1 x_2 \ldots x_k) \leq (k-1)\varepsilon|x_1 x_2 \ldots x_k|.$$

L'erreur sur un quotient est donnée de même par $\Delta(x/y) \leq \varepsilon|x/y|$. On en déduit pour tous exposants $\alpha_i \in \mathbb{Z}$ la formule générale

$$\Delta(x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_k^{\alpha_k}) \leq (|\alpha_1| + \ldots + |\alpha_k| - 1)\varepsilon|x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_k^{\alpha_k}| \ ;$$

on observera que $|\alpha_1| + \ldots + |\alpha_k| - 1$ est exactement le nombre d'opérations requises pour calculer $x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_k^{\alpha_k}$ par multiplications ou divisions successives des x_i.
Contrairement au cas de l'addition, la majoration de l'erreur d'un produit *ne dépend pas de l'ordre des facteurs.*

1.5. Règle de Hörner

On s'intéresse ici au problème de l'évaluation d'un polynôme

$$P(x) = \sum_{k=0}^{n} a_k x^k.$$

La méthode la plus « naïve » qui vient à l'esprit consiste à poser $x^0 = 1$, $s_0 = a_0$, puis à calculer par récurrence

$$\begin{cases} x^k = x^{k-1} \cdot x \\ u_k = a_k \cdot x^k \\ s_k = s_{k-1} + u_k \end{cases} \qquad \text{pour} \quad k \geq 1.$$

Pour chaque valeur de k, deux multiplications et une addition sont donc nécessaires. Il existe en fait une méthode plus efficace :

Règle de Hörner – On factorise $P(x)$ sous la forme :

$$P(x) = a_0 + x(a_1 + x(a_2 + \ldots + x(a_{n-1} + xa_n)\ldots)).$$

Si l'on pose

$$p_k = a_k + a_{k+1}x + \ldots + a_n x^{n-k},$$

cette méthode revient à calculer $P(x) = p_0$ par récurrence descendante :

$$\begin{cases} p_n = a_n \\ p_{k-1} = a_{k-1} + xp_k, \qquad 1 \leq k \leq n. \end{cases}$$

On effectue ainsi seulement une multiplication et une addition à chaque étape, ce qui économise une multiplication et donc une fraction substantielle du temps d'exécution.

Comparons maintenant les erreurs d'arrondi dans chacune des deux méthodes, en supposant que les réels $x, a_0, a_1, \ldots, a_n$ sont représentés sans erreur.

• **Méthode « naïve ».** On a ici $P(x) = s_n$ avec

$$\begin{aligned} \Delta(a_k \cdot x^k) &\leq k\varepsilon|a_k||x|^k, \\ \Delta s_k &\leq \Delta s_{k-1} + k\varepsilon|a_k||x|^k + \varepsilon(|s_{k-1}| + |u_k|) \\ &\leq \Delta s_{k-1} + k\varepsilon|a_k||x|^k + \varepsilon(|a_0| + |a_1||x| + \ldots + |a_k||x|^k). \end{aligned}$$

Comme $\Delta s_0 = 0$, il vient après sommation sur k :

$$\begin{aligned} \Delta s_n &\leq \sum_{k=1}^{n} k\varepsilon|a_k||x|^k + \varepsilon\sum_{k=1}^{n}(|a_0| + |a_1||x| + \ldots + |a_k||x|^k) \\ &\leq \sum_{k=1}^{n} k\varepsilon|a_k||x|^k + \varepsilon\sum_{k=0}^{n}(n+1-k)|a_k||x|^k. \end{aligned}$$

On obtient par conséquent

$$\Delta P(x) \leq (n+1)\varepsilon\sum_{k=0}^{n}|a_k||x|^k.$$

• **Règle de Hörner.** Dans ce cas, on a

$$\begin{aligned} \Delta p_{k-1} &\leq \Delta(xp_k) + \varepsilon(|a_{k-1}| + |xp_k|) \\ &\leq (|x|\Delta p_k + \varepsilon|xp_k|) + \varepsilon(|a_{k-1}| + |xp_k|) \\ &= \varepsilon(|a_{k-1}| + 2|x||p_k|) + |x|\Delta p_k. \end{aligned}$$

En développant $\Delta P(x) = \Delta p_0$, il vient

$$\Delta p_0 \leq \varepsilon(|a_0| + 2|x||p_1|) + |x|\Big(\varepsilon|a_1| + 2|x||p_2| + |x|\Big(\varepsilon|a_2| + \ldots\Big)\Big)$$

d'où

$$\begin{aligned} \Delta P(x) &\leq \varepsilon\sum_{k=0}^{n}|a_k||x|^k + 2\varepsilon\sum_{k=1}^{n}|x|^k|p_k|, \\ \Delta P(x) &\leq \varepsilon\sum_{k=0}^{n}|a_k||x|^k + 2\varepsilon\sum_{k=1}^{n}(|a_k||x|^k + \ldots + |a_n||x|^n), \\ \Delta P(x) &\leq \varepsilon\sum_{k=0}^{n}(2k+1)|a_k||x|^k. \end{aligned}$$

On voit que la somme des coefficients d'erreur affectés aux termes $|a_k||x|^k$, soit $\varepsilon\sum_{k=0}^{n}(2k+1) = \varepsilon(n+1)^2$, est la même que pour la méthode naïve ; comme $2k+1 \leq 2(n+1)$, l'erreur commise sera dans le pire des cas égale à 2 fois celle de la méthode naïve. Néanmoins, les petits coefficients portent sur les premiers termes calculés, de sorte que la précision de la méthode de Hörner sera nettement meilleure si le terme $|a_k||x|^k$ décroît rapidement : c'est le cas par exemple si $P(x)$ est le début d'une série convergente.

Exercice – *Evaluer dans les deux cas l'erreur commise sur les sommes partielles de la série exponentielle*

$$\sum_{k=0}^{n}\frac{x^k}{k!}, \quad x \geq 0$$

en tenant compte du fait qu'on a une certaine erreur d'arrondi sur $a_k = \frac{1}{k!}$.

Réponse. *On trouve* $\Delta P(x) \leq \varepsilon(1+(n+x)e^x)$ *pour la méthode naïve, tandis que la factorisation*

$$P(x) = 1 + x\Big(1 + \frac{x}{2}\Big(1 + \frac{x}{3}\Big(1 + \ldots\Big(1 + \frac{x}{n-1}\Big(1 + \frac{x}{n}\Big)\Big)\ldots\Big)\Big)\Big)$$

donne $\Delta P(x) \leq \varepsilon(1+3xe^x)$, *ce qui est nettement meilleur en pratique puisque* n *doit être choisi assez grand.*

1.6. CUMULATION D'ERREURS D'ARRONDI ALÉATOIRES

Les majorations d'erreurs que nous avons données plus haut pêchent en général par excès de pessimisme, car nous n'avons tenu compte que de la valeur absolue des erreurs, alors qu'en pratique elles sont souvent de signe aléatoire et se compensent donc partiellement entre elles.

Supposons par exemple qu'on cherche à calculer une somme s_n de rang élevé d'une série convergente $S = \sum_{k=0}^{+\infty} u_k$, les u_k étant des réels ≥ 0 supposés représentés sans erreur. On pose donc

$$s_k = s_{k-1} + u_k, \quad s_0 = u_0,$$

et les erreurs Δs_k vérifient

$$\begin{aligned} &\Delta s_k = \Delta s_{k-1} + \alpha_k \\ \text{avec} \quad &\Delta s_0 = 0 \quad \text{et} \quad |\alpha_k| \leq \varepsilon(s_{k-1} + u_k) = \varepsilon s_k \leq \varepsilon S. \end{aligned}$$

On en déduit donc

$$\Delta s_n = \alpha_1 + \alpha_2 + \ldots + \alpha_n$$

et en particulier $|\Delta s_n| \leq n\varepsilon S$. Dans le pire des cas, l'erreur est donc proportionnelle à n. On va voir qu'on peut en fait espérer beaucoup mieux sous des hypothèses raisonnables.

Hypothèses

(1) Les erreurs α_k sont des variables aléatoires globalement indépendantes les unes des autres (lorsque les u_k sont choisis aléatoirement).

(2) L'espérance mathématique $E(\alpha_k)$ est nulle, ce qui signifie que les erreurs d'arrondi n'ont aucune tendance à se faire par excès ou par défaut.

L'hypothèse (2) entraîne $E(\Delta s_n) = 0$ tandis que l'hypothèse (1) donne

$$\mathrm{var}(\Delta s_n) = \mathrm{var}(\alpha_1) + \ldots + \mathrm{var}(\alpha_n).$$

Comme $E(\alpha_k) = 0$ et $|\alpha_k| \leq \varepsilon S$, on a $\mathrm{var}(\alpha_k) \leq \varepsilon^2 S^2$, d'où

$$\sigma(\Delta s_n) = \sqrt{\mathrm{var}(\Delta s_n)} \leq \sqrt{n}\varepsilon S$$

L'erreur quadratique moyenne croît seulement dans ce cas comme $\sqrt{n}$. D'après l'inégalité de Bienaymé-Tchebychev on a :

$$P(|\Delta s_n| \geq \alpha\, \sigma(\Delta s_n)) \leq \alpha^{-2}.$$

La probabilité que l'erreur dépasse $10\sqrt{n}\varepsilon S$ est donc inférieure à 1%.

2. Phénomènes de compensation

Les phénomènes de compensation se produisent lorsqu'on tente d'effectuer des soustractions de valeurs très voisines. Ils peuvent conduire à des pertes importantes de précision.

Les exemples suivants illustrent les difficultés pouvant se présenter et les remèdes à apporter dans chaque cas.

2.1. Exemple : résolution de l'équation $x^2 - 1634x + 2 = 0$

Supposons que les calculs soient effectués avec 10 chiffres significatifs. Les formules habituelles donnent alors

$$\begin{aligned}
&\Delta' = 667\,487, \qquad \sqrt{\Delta'} \simeq 816,9987760\\
&x_1 = 817 + \sqrt{\Delta'} \simeq 1633,998776,\\
&x_2 = 817 - \sqrt{\Delta'} \simeq 0,0012240.
\end{aligned}$$

On voit donc qu'on a une perte de 5 chiffres significatifs sur x_2 si l'on effectue la soustraction telle qu'elle se présente naturellement ! Ici, le remède est simple : il suffit d'observer que $x_1 x_2 = 2$, d'où

$$x_2 = \frac{2}{x_1} \simeq 1,223991125 \cdot 10^{-3}.$$

C'est donc l'*algorithme numérique utilisé qui doit être modifié.*

2.2. EXEMPLE : CALCUL APPROCHÉ DE e^{-10}

Supposons qu'on utilise pour cela la série

$$e^{-10} \simeq \sum_{k=0}^{n} (-1)^k \frac{10^k}{k!},$$

les calculs étant toujours effectués avec 10 chiffres significatifs. Le terme général $|u_k| = 10^k/k!$ est tel que

$$\frac{|u_k|}{|u_{k-1}|} = \frac{10}{k} \geq 1 \quad \text{dès que} \quad k \leq 10.$$

On a donc 2 termes de valeur absolue maximale

$$|u_9| = |u_{10}| = \frac{10^{10}}{10!} \simeq 2,755 \cdot 10^3$$

tandis que $e^{-10} \simeq 4,5 \cdot 10^{-5}$. Comparons u_{10} et e^{-10} :

u_{10} :	2	7	5	5,	·	·	·	·	·	·
e^{-10} :				0,	0	0	0	0	4	5

Ceci signifie qu'au moins 8 chiffres significatifs vont être perdus par compensation des termes u_k de signes opposés. Un remède simple consiste à utiliser la relation

$$e^{-10} = 1/e^{10} \quad \text{avec} \quad e^{10} \simeq \sum_{k=0}^{n} \frac{10^k}{k!}.$$

On essaiera dans la mesure du possible d'*éviter les sommations dans lesquelles des termes de signes opposés se compensent.*

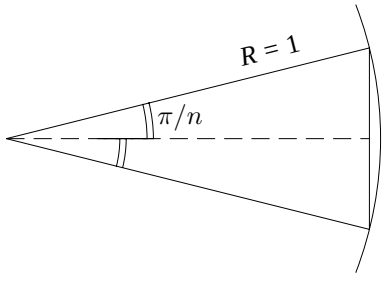
2.3. EXEMPLE : CALCUL APPROCHÉ DE π PAR LES POLYGÔNES INSCRITS

Soit P_n le demi-périmètre du polygone régulier à n côtés inscrit dans un cercle de rayon 1.

Le côté de ce polygone vaut $2 \cdot R \sin \pi/n = 2 \sin \pi/n$, d'où

$$P_n = n \sin \frac{\pi}{n},$$
$$P_n = \pi - \frac{\pi^3}{6n^2} + o\Big(\frac{1}{n^3}\Big).$$

On obtient donc une approximation de π avec une précision de l'ordre de $6/n^2$.



Pour évaluer P_n, on utilise une méthode de dichotomie permettant de calculer par récurrence

$$x_k = P_{2^k} = 2^k \sin \frac{\pi}{2^k}.$$

Si α est un angle compris entre 0 et $\frac{\pi}{2}$ on a

$$\sin \frac{\alpha}{2} = \sqrt{\frac{1}{2}(1 - \cos \alpha)} = \sqrt{\frac{1}{2}(1 - \sqrt{1 - \sin^2 \alpha})}. \qquad (*)$$

En substituant $\alpha = \pi/2^k$, on en déduit les formules

$$\begin{cases} x_{k+1} = 2^k \sqrt{2(1 - \sqrt{1 - (x_k/2^k)^2})} \\ x_1 = 2, \end{cases}$$

et d'après ce qu'on a dit plus haut $\lim\limits_{k \to +\infty} x_k = \pi$.

Ce n'est pourtant pas du tout ce qu'on va observer sur machine ! Dès que $(x_k/2^k)^2$ sera inférieur à la précision relative des calculs, l'ordinateur va donner

$$\sqrt{1 - (x_k/2^k)^2} = 1 \quad \text{d'où} \quad x_{k+1} = 0.$$

Pour éviter cette difficulté, il suffit de remplacer $(*)$ par

$$\sin \frac{\alpha}{2} = \sqrt{\frac{1}{2} \, \frac{1 - \cos^2 \alpha}{1 + \cos \alpha}} = \frac{\sin \alpha}{\sqrt{2(1 + \cos \alpha)}} \qquad (**)$$

d'où

$$\sin \frac{\alpha}{2} = \frac{\sin \alpha}{\sqrt{2(1 + \sqrt{1 - \sin^2 \alpha})}}.$$

On obtient alors la formule de récurrence

$$x_{k+1} = \frac{2x_k}{\sqrt{2(1 + \sqrt{1 - (x_k/2^k)^2})}}$$

qui évite le phénomène de compensation précédent, de sorte que le calcul des x_k peut être poussé beaucoup plus loin.

On obtiendra une méthode plus efficace encore en observant qu'on peut évaluer $\cos\alpha$ dans $(**)$ par la formule $\cos\alpha = \frac{\sin 2\alpha}{2\sin\alpha}$. Ceci donne

$$\sin\frac{\alpha}{2} = \sin\alpha\sqrt{\frac{\sin\alpha}{2\sin\alpha+\sin 2\alpha}}, \quad \text{d'où}$$
$$x_{k+1} = x_k\sqrt{\frac{2x_k}{x_k+x_{k-1}}}.$$

Deux valeurs d'initialisation sont alors requises pour démarrer, par exemple $x_1 = 2$ et $x_2 = 2\sqrt{2}$.

3. PHÉNOMÈNES D'INSTABILITÉ NUMÉRIQUE

Il s'agit de phénomènes d'amplification des erreurs d'arrondi. Une telle amplification se produit assez fréquemment dans le cas de calculs récurrents ou itératifs.

3.1. CAS D'UN CALCUL RÉCURRENT

Supposons à titre d'exemple qu'on cherche à évaluer numériquement l'intégrale

$$I_n = \int_0^1 \frac{x^n}{10+x}, \qquad n\in\mathbb{N}.$$

Un calcul immédiat donne

$$I_0 = \int_0^1 \frac{dx}{10+x} = \Big[\ln(10+x)\Big]_0^1 = \ln\frac{11}{10},$$
$$I_n = \int_0^1 \frac{x}{10+x}\cdot x^{n-1}dx = \int_0^1 \Big(1-\frac{10}{10+x}\Big)x^{n-1}dx$$
$$= \int_0^1 x^{n-1}dx - 10\int_0^1 \frac{x^{n-1}}{10+x}\,dx = \frac{1}{n} - 10\,I_{n-1}.$$

Ceci permet de calculer I_n par récurrence avec

$$\begin{cases} I_0 = \ln\frac{11}{10} \\ I_n = \frac{1}{n} - 10\,I_{n-1}. \end{cases}$$

Ce problème apparemment bien posé mathématiquement conduit numériquement à des résultats catastrophiques. On a en effet

$$\Delta I_n \simeq 10\,\Delta I_{n-1},$$

même si on néglige l'erreur d'arrondi sur $1/n$. L'erreur sur I_n explose donc exponentiellement, l'erreur initiale sur I_0 étant multipliée par 10^n à l'étape n. Comment faire alors pour calculer par exemple I_{36} ? La suite x^n étant décroissante

pour $x \in [0,1]$, on voit que la suite I_n est elle-même décroissante. Comme $10 \leq 10+x \leq 11$, on a de plus

$$\frac{1}{11(n+1)} \leq I_n \leq \frac{1}{10(n+1)}.$$

L'approximation $I_n \simeq \frac{1}{11(n+1)}$ donne une erreur absolue $\leq \frac{1}{110(n+1)}$ et donc une erreur relative $\frac{\Delta I_n}{I_n} \leq \frac{1}{10}$. Ceci donne l'ordre de grandeur mais n'est pas très satisfaisant. L'idée est alors de *renverser la récurrence* en posant

$$I_{n-1} = \frac{1}{10}\Big(\frac{1}{n} - I_n\Big).$$

En négligeant l'erreur sur $\frac{1}{n}$, on a donc cette fois $\Delta I_{n-1} \simeq \frac{1}{10}\,\Delta I_n$, estimation qui va dans le bon sens. Si l'on part de $I_{46} \simeq \frac{1}{11.47}$, on obtiendra pour I_{36} une erreur relative sans doute meilleure que 10^{-10}.

Exercice – *Montrer que* $0 \leq I_n - I_{n+1} \leq \frac{1}{10(n+1)(n+2)}$, *et en déduire à partir de la formule exprimant* I_n *en fonction de* I_{n+1} *que l'on a en fait l'estimation*

$$\frac{1}{11(n+1)} \leq I_n \leq \frac{1}{11(n+1)} + \frac{1}{110(n+1)(n+2)},$$

donc $\Delta I_n \simeq \frac{1}{11(n+1)}$ *avec erreur relative* $\leq \frac{1}{10(n+2)}$.

On voit donc le rôle fondamental joué par le coefficient d'amplification de l'erreur, 10 dans le premier cas, 1/10 dans le second. En général, si on a un coefficient d'amplification $A > 1$, il est impératif de limiter le nombre n d'étapes en sorte que $A^n\varepsilon$ reste très inférieur à 1, si ε est la précision relative des calculs.

3.2. Calculs itératifs

Soit à calculer une suite (u_n) définie par sa valeur initiale u_0 et par la relation de récurrence

$$u_{n+1} = f(u_n),$$

où f est une fonction donnée. On a donc $u_n = f^n(u_0)$ où $f^n = f \circ f \circ \ldots \circ f$ est la n-ième itérée de f. On considère par exemple la suite (u_n) telle que

$$u_0 = 2, \qquad u_{n+1} = |\ln(u_n)|,$$

dont on cherche à évaluer le terme u_{30}. Un calcul effectué à la précision 10^{-9} sur un ordinateur nous a donné $u_{30} \simeq 0,880833175$.

A la lumière de l'exemple précédent, il est néanmoins légitime de se demander si ce calcul est bien significatif, compte tenu de la présence des erreurs d'arrondi. En partant de valeurs de u_0 très voisines de 2, on obtient en fait les résultats suivants

(arrondis à 10^{-9} près, sur la même implémentation de calcul que ci-dessus) :

u_0	2,000000000	2,000000001	1,999999999	$5 \cdot 10^{-10}$
u_5	5,595485181	5,595484655	5,595485710	$9 \cdot 10^{-8}$
u_{10}	0,703934587	0,703934920	0,703934252	$5 \cdot 10^{-7}$
u_{15}	1,126698502	1,126689382	1,126707697	$8 \cdot 10^{-6}$
u_{20}	1,266106839	1,266256924	1,265955552	10^{-4}
u_{24}	1,000976376	1,001923276	1,000022532	10^{-3}
u_{25}	0,000975900	0,001921429	0,000022532	100%
u_{26}	6,932150628	6,254686211	10,700574400	50%
u_{30}	0,880833175	0,691841353	1,915129896	100%

La dernière colonne donne l'ordre de grandeur de l'écart relatif $\Delta u_n/u_n$ observé entre la deuxième ou troisième colonne et la première colonne. On voit que cet écart augmente constamment pour atteindre environ 10^{-3} sur u_{24}. Pour le calcul de u_{25}, il se produit une véritable catastrophe numérique : l'écart relatif devient voisin de 100% ! Il en résulte que toutes les valeurs calculées à partir de u_{25} sont certainement non significatives pour une précision des calculs de 10^{-9}.

Pour comprendre ce phénomène, il suffit d'observer qu'une erreur Δx sur la variable x entraîne une erreur $\Delta f(x)$ sur $f(x)$, approximativement donnée par

$$\Delta f(x) = |f'(x)|\,\Delta x.$$

Ceci se voit bien sûr en approximant $f(x+\Delta x) - f(x)$ par sa différentielle $f'(x)\Delta x$, lorsque f est dérivable au point x. Le coefficient d'amplification de l'erreur absolue est donc donné par la valeur absolue de la dérivée $|f'(x)|$; ce coefficient peut être parfois assez grand. Souvent dans les calculs numériques (et ici en particulier), il est plus pertinent de considérer les erreurs relatives. La formule

$$\frac{\Delta f(x)}{|f(x)|} = \frac{|f'(x)||x|}{|f(x)|}\,\frac{\Delta x}{|x|}$$

montre que le coefficient d'amplification de l'erreur relative est $|f'(x)||x|/|f(x)|$. Dans le cas $f(x) = \ln(x)$ qui nous intéresse, ce coefficient vaut $1/|\ln x|$; il devient très grand lorsque x est proche de 1, comme c'est le cas par exemple pour u_{24}.

4. PROBLÈMES

4.1. Soit $x \geq 0$; on note $F(x) = \dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^x e^{-t^2}\,dt$.

(a) Encadrer $F(x)$ par deux entiers consécutifs.

(b) En remplaçant e^{-t^2} par un développement en série entière de x, exprimer $F(x)$ comme somme d'une série. On choisit $x = 3$; calculer les 10 premiers termes

de la série. En déduire que pour $x \geq 3$ on a un phénomène de compensation dans le calcul de la somme des premiers termes de la série.

(c) On définit $g(x)$ par $F(x) = e^{-x^2} g(x)$.

Montrer que g est solution d'une équation différentielle.

Exprimer $g(x)$ comme somme d'une série entière en x.

(d) En déduire l'expression $F(x) = \sum_{n=0}^{+\infty} a_n(x)$ où les $a_n(x)$ sont tous positifs. Déterminer $a_0(x)$ et donner la solution de récurrence entre $a_n(x)$ et $a_{n-1}(x)$. Montrer l'inégalité

$$\sum_{n=N+1}^{+\infty} a_n(x) \leq a_N \, \frac{x^2}{N - x^2} \quad (\text{pour } N > x^2)$$

(e) En utilisant les résultats précédents, écrire un programme en langage informatique qui, à la lecture de x et d'un entier $k \leq 1$ calcule une valeur approchée de $F(x)$ à 10^{-k} près.

4.2. Soit $(I_n)_{n\in\mathbb{N}}$ la suite des intégrales

$$I_n = \int_0^1 \frac{x^n}{6 + x - x^2} \, dx.$$

(a) Montrer que I_n vérifie une relation de récurrence de la forme

$$I_{n+1} = \alpha I_n + \beta I_{n-1} + c_n \qquad (*)$$

où α, β sont des constantes et (c_n) une suite numérique explicite.

(b) On envisage le calcul récurrent de I_n à partir de I_0 et I_1 par la formule $(*)$.

On suppose que les valeurs de I_0 et I_1 sont affectées d'erreurs d'arrondis ε_0 et ε_1, et on note ε_n l'erreur qui en résulte sur I_n (on néglige ici l'erreur sur le calcul de c_n et les erreurs d'arrondi pouvant intervenir dans l'application de la formule $(*)$).

(α) Déterminer ε_n en fonction de ε_0 et ε_1.

(β) Est-il possible de calculer I_{50} par ce procédé avec un ordinateur donnant une précision relative de 10^{-10} ?

4.3. Etant donné une suite x_k, $k = 1, \ldots, n$ de réels, on note $\mu_n = \frac{1}{n} \sum_{k=1}^{n} x_k$ la moyenne et $\sigma_n = \sqrt{\sigma_n^2}$ l'écart type avec $\sigma_n^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \mu_n)^2$.

(a) Soit $q_n = \sum_{k=1}^{n} x_k^2$. Exprimer σ_n^2 en fonction de q_n et de μ_n.

(b) Ecrire un programme qui calcule les moyennes et l'écart type d'un nombre indéterminé de réels. Les données réelles sont entrées au clavier ; après chaque entrée on affichera la moyenne et l'écart type de la suite des nombres déjà entrés.

(c) On suppose que pour $k = 1, \ldots, n$ on a $x_k = \mu + \varepsilon_k$ avec $|\varepsilon_k| < \varepsilon$ où ε est petit devant μ. Montrer que l'on a l'inégalité $\left|\frac{q_n}{n} - \mu^2\right| \leq 3\mu\varepsilon$.

En déduire que la méthode de calcul de σ_n utilisant la formule du (a) est inadaptée pour une telle suite.

(d) On veut obtenir un algorithme de calcul de σ_n plus stable.

Etablir les égalités :

$$(n+1)\sigma_{n+1}^2 = n\sigma_n^2 + n(\mu_{n+1} - \mu_n)^2 + (x_{n+1} - \mu_{n+1})^2,$$
$$(n+1)\mu_{n+1} = n\mu_n + x_{n+1}.$$

En déduire $\sigma_{n+1}^2 = \frac{n}{n+1}\ \sigma_n^2 + \frac{1}{n}\ (x_{n+1} - \mu_{n+1})^2$.

(e) Reprendre la question (b) avec le nouvel algorithme.

(f) On considère une suite de réels $x_k = 1 + \varepsilon\ \frac{2k-n-1}{n-1}$, $k = 1, \ldots, n$. Déterminer sa moyenne et son écart type.

(g) Même question pour la suite des 2^n réels x_k, $k = 1, \ldots, 2^n$ telle que pour $p = 0, \ldots, n$ on ait C_n^p termes égaux à $\mu + \frac{2p-n}{\sqrt{n}}$ (On pourra remarquer que $p^2 C_n^p = pn C_{n-1}^{p-1}$).

CHAPITRE II

APPROXIMATION POLYNOMIALE DES FONCTIONS NUMÉRIQUES

Les fonctions les plus faciles à évaluer numériquement sont les fonctions polynômes. Il est donc important de savoir approximer une fonction arbitraire par des polynômes. Dans ce cadre, l'un des outils de base est la méthode d'interpolation de Lagrange.

Notations – Dans toute la suite, on désignera par $\mathcal{P}_n$ l'espace vectoriel des fonctions polynômes sur $\mathbb{R}$ à coefficients réels, de degré inférieur ou égal à n. On a donc $\dim \mathcal{P}_n = n+1$.

Par ailleurs, si f est une fonction définie sur un intervalle $[a,b] \subset \mathbb{R}$ à valeurs dans $\mathbb{R}$ ou $\mathbb{C}$, la *norme uniforme* de f sur $[a,b]$ sera notée

$$\|f\|_{[a,b]} = \sup_{x\in[a,b]} |f(x)|$$

où même simplement $\|f\|$ s'il n'y a pas d'ambiguïté. Enfin $\mathcal{C}([a,b])$ désignera l'espace des fonctions continues sur $[a,b]$ à valeurs dans $\mathbb{R}$.

1. MÉTHODE D'INTERPOLATION DE LAGRANGE

1.1. EXISTENCE ET UNICITÉ DU POLYNÔME D'INTERPOLATION

Soit $f : [a,b] \to \mathbb{R}$ une fonction continue. On se donne $n+1$ points $x_0, x_1, \ldots, x_n$ dans $[a,b]$, deux à deux distincts, non nécessairement rangés par ordre croissant.

Problème – *Existe-t-il un polynôme $p_n \in \mathcal{P}_n$ tel que $p_n(x_i) = f(x_i)$, $\forall i = 0, 1, \ldots, n$?*

Un tel polynôme sera appelé *polynôme d'interpolation (de Lagrange) de f aux points $x_0, x_1, \ldots, x_n$*. Posons

$$l_i(x) = \prod_{j\neq i} \frac{(x - x_j)}{(x_i - x_j)}, \quad 0 \leq i \leq n,$$

où le produit est effectué sur les indices j tels que $0 \leq j \leq n$, $j \neq i$. Il est clair que $l_i \in \mathcal{P}_n$ et que

$$l_i(x_j) = 0 \quad \text{si} \quad j \neq i,$$
$$l_i(x_i) = 1.$$

Le problème ci-dessus admet donc au moins une solution

$$p_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x), \quad p_n \in \mathcal{P}_n. \tag{$*$}$$

Théorème – *Le problème d'interpolation $p_n(x_i) = f(x_i)$, $0 \leq i \leq n$, admet une solution et une seule, donnée par la formule $(*)$.*

Il reste à prouver l'*unicité*. Supposons que $q_n \in \mathcal{P}_n$ soit une autre solution du problème. Alors $p_n(x_i) = q_n(x_i) = f(x_i)$, donc x_i est racine de $q_n - p_n$. Par suite le polynôme

$$\pi_{n+1}(x) = \prod_{j=0}^{n} (x - x_j)$$

divise $q_n - p_n$. Comme $\deg \pi_n = n + 1$ et $q_n - p_n \in \mathcal{P}_n$, la seule possibilité est que $q_n - p_n = 0$. ∎

Remarque 1 – On a $\pi_{n+1}(x) = (x - x_i) \cdot \prod_{j \neq i} (x - x_j)$, d'où

$$\pi'_{n+1}(x_i) = \prod_{j \neq i} (x_i - x_j).$$

Ceci donne la formule

$$l_i(x) = \frac{\pi_{n+1}(x)}{(x - x_i)\pi'_{n+1}(x_i)}.$$

Remarque 2 – Pour démontrer le théorème, on peut également poser $p_n(x) = \sum_{j=0}^{n} a_j x^j$ et résoudre un système linéaire de $n + 1$ équations

$$\begin{cases} \displaystyle\sum_{j=0}^{n} a_j x_i^j = f(x_i), \quad 0 \leq i \leq n, \end{cases}$$

en les $n+1$ inconnues $a_0, a_1, \ldots, a_n$. Le déterminant du système est un déterminant dit de Van der Monde :

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & & & \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix}.$$

Il s'agit de montrer que $\Delta \neq 0$ si les x_i sont distincts. Or Δ est un polynôme de degré total $1 + 2 + \ldots + n = \frac{n(n+1)}{2}$ en les variables $x_0, x_1, \ldots, x_n$. Il est clair que $\Delta = 0$ chaque fois que $x_i = x_j$ pour un couple (i, j) tel que $0 \leq j < i \leq n$. Δ est donc divisible par le polynôme $\prod_{0 \leq j < i \leq n}(x_i - x_j)$, qui est lui aussi de degré total $\frac{n(n+1)}{2}$. Le quotient est donc une constante, donnée par exemple par le coefficient de $x_1 x_2^2 \ldots x_n^n$ dans Δ, qui vaut 1. Par suite

$$\Delta = \prod_{0 \leq j < i \leq n} (x_i - x_j). \qquad \blacksquare$$

Il n'est pas recommandé de résoudre numériquement le système précédent pour obtenir p_n. Nous verrons plus loin une méthode beaucoup plus efficace (cf. § 1.3).

Exercice – *On se propose de donner deux autres démonstrations des résultats ci-dessus, grâce à des arguments d'algèbre linéaire.*

(a) *Montrer que l'application $\phi_n : \mathcal{P}_n \to \mathbb{R}^{n+1}$, $p \mapsto (p(x_i))_{0 \leq i \leq n}$ est linéaire. En déduire que ϕ_n est injective si et seulement si elle est surjective. Traduire ces résultats en terme d'existence et d'unicité du polynôme d'interpolation.*

(b) *Montrer par récurrence sur n que ϕ_n est surjective* [Indication : *si le résultat est vrai pour $n-1$, ajuster la valeur $p(x_n)$ à l'aide du polynôme de degré n $(x - x_0) \ldots (x - x_{n-1})$]. Conclure.*

(c) *Montrer directement que les polynômes $(l_i)_{0 \leq i \leq n}$ forment une famille libre.*

En déduire que c'est une base de $\mathcal{P}_n$ et que ϕ_n est un isomorphisme.

1.2. Formule d'erreur

L'erreur d'interpolation est donnée par la formule théorique suivante.

Théorème – *On suppose que f est $n+1$ fois dérivable sur $[a, b]$. Alors pour tout $x \in [a, b]$, il existe un point $\xi_x \in]\min(x, x_i),\ \max(x, x_i)[$ tel que*

$$f(x) - p_n(x) = \frac{1}{(n+1)!}\, \pi_{n+1}(x) f^{(n+1)}(\xi_x).$$

On a besoin du lemme suivant, qui découle du théorème de Rolle.

Lemme – *Soit g une fonction p fois dérivable sur $[a, b]$. On suppose qu'il existe $p+1$ points $c_0 < c_1 < \ldots < c_p$ de $[a, b]$ tels que $g(c_i) = 0$. Alors il existe $\xi \in]c_0, c_p[$ tel que $g^{(p)}(\xi) = 0$.*

Le lemme se démontre par récurrence sur p. Pour $p = 1$, c'est le théorème de Rolle. Supposons le lemme démontré pour $p - 1$. Le théorème de Rolle donne des points $\gamma_0 \in]c_0, c_1[, \ldots, \gamma_{p-1} \in]c_{p-1}, c_p[$ tels que $g'(\gamma_i) = 0$. Par hypothèse de récurrence, il existe donc $\xi \in]\gamma_0, \gamma_{p-1}[\subset]c_0, c_p[$ tel que $(g')^{(p-1)}(\xi) = g^{(p)}(\xi) = 0$. $\blacksquare$

Démonstration du théorème

• Si $x = x_i$, on a $\pi_{n+1}(x_i) = 0$, tout point ξ_x convient.

• Supposons maintenant x distinct des points x_i.

Soit $p_{n+1}(t)$ le polynôme d'interpolation de $f(t)$ aux points $x, x_0, \ldots, x_{n+1}$, de sorte que $p_{n+1} \in \mathcal{P}_{n+1}$. Par construction $f(x) - p_n(x) = p_{n+1}(x) - p_n(x)$. Or le polynôme $p_{n+1} - p_n$ est de degré $\leq n+1$ et s'annule aux $n+1$ points $x_0, x_1, \ldots, x_n$. On a donc

$$p_{n+1}(t) - p_n(t) = c \cdot \pi_{n+1}(t), \quad c \in \mathbb{R}.$$

Considérons la fonction

$$g(t) = f(t) - p_{n+1}(t) = f(t) - p_n(t) - c\pi_{n+1}(t).$$

Cette fonction s'annule en les $n+2$ points $x, x_0, x_1, \ldots, x_n$ donc d'après le lemme il existe $\xi_x \in \,]\min(x, x_i), \max(x, x_i)[$ tel que $g^{(n+1)}(\xi_x) = 0$. Or

$$p_n^{(n+1)} = 0, \quad \pi_{n+1}^{(n+1)} = (n+1)!$$

On a par conséquent $g^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - c \cdot (n+1)! = 0$, d'où

$$f(x) - p_n(x) = p_{n+1}(x) - p_n(x) = c\pi_{n+1}(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\ \pi_{n+1}(x). \qquad \blacksquare$$

En prenant la borne supérieure de la valeur absolue des deux membres dans la formule d'erreur, on obtient en particulier :

Corollaire – $\|f - p_n\| \ \leq \dfrac{1}{(n+1)!}\ \|\pi_{n+1}\|\ \|f^{(n+1)}\|.$

Ces formules montrent que la taille de l'erreur d'interpolation $f(x) - p_n(x)$ dépend à la fois de la quantité $\|f^{(n+1)}\|$, qui peut être grande si f oscille trop vite, et de la quantité $\|\pi_{n+1}\|$, qui est liée à la répartition des points x_i dans l'intervalle $[a, b]$.

1.3. Méthode des différences divisées

On va décrire ici une méthode simple et efficace permettant de calculer les polynômes d'interpolation de f. Soit p_k le polynôme d'interpolation de f aux points $x_0, x_1, \ldots, x_k$.

Notation – *On désigne par $f[x_0, x_1, \ldots, x_k]$ le coefficient directeur du polynôme p_k (= coefficient de t^k dans $p_k(t)$).*

Alors $p_k - p_{k-1}$ est un polynôme de degré $\leq k$, s'annulant aux points $x_0, x_1, \ldots, x_{k-1}$, et admettant $f[x_0, x_1, \ldots, x_k]$ pour coefficient directeur. Par suite :

$$p_k(x) - p_{k-1}(x) = f[x_0, x_1, \ldots, x_k](x - x_0) \ldots (x - x_{k-1}).$$

Comme $p_0(x) = f(x_0)$, on en déduit la formule fondamentale

$$p_n(x) = f(x_0) + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k](x - x_0)\ldots(x - x_{k-1}). \qquad (**)$$

Pour pouvoir exploiter cette formule, il reste bien entendu à évaluer les coefficients $f[x_0, x_1, \ldots, x_k]$. On utilise à cette fin une récurrence sur le nombre k de points x_i, en observant que $f[x_0] = f(x_0)$.

Formule de récurrence – Pour $k \geq 1$, on a

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}. \qquad (***)$$

A cause de cette formule, la quantité $f[x_0, x_1, \ldots, x_k]$ est appelée *différence divisée d'ordre k* de f aux points $x_0, \ldots, x_k$.

Vérification de $(***)$. Désignons par $q_{k-1} \in \mathcal{P}_{k-1}$ le polynôme de f aux points $x_1, x_2, \ldots, x_k$. Posons

$$\widetilde{p}_k(x) = \frac{(x - x_0)q_{k-1}(x) - (x - x_k)p_{k-1}(x)}{x_k - x_0}.$$

Alors $\widetilde{p}_k \in \mathcal{P}_k$, $\widetilde{p}_k(x_0) = p_{k-1}(x_0) = f(x_0)$, $\widetilde{p}_k(x_k) = q_{k-1}(x_k) = f(x_k)$ et pour $0 < i < k$ on a

$$\widetilde{p}_k(x_i) = \frac{(x_i - x_0)f(x_i) - (x_i - x_k)f(x_i)}{x_k - x_0} = f(x_i).$$

Par conséquent $\widetilde{p}_k = p_k$. Comme le coefficient directeur de q_{k-1} est $f[x_1, \ldots, x_k]$, on obtient la formule $(***)$ cherchée en égalant les coefficients de x^k dans l'identité

$$p_k(x) = \frac{(x - x_0)q_{k-1}(x) - (x - x_k)p_{k-1}(x)}{x_k - x_0}.$$

Algorithme pratique – On range les valeurs $f(x_i)$ dans un tableau TAB, puis on modifie ce tableau en n étapes successives, en procédant par indices décroissants :

Tableau	Etape 0	Etape 1	Etape 2	...	Etape n
TAB $[n]$	$f(x_n)$	$f[x_{n-1}, x_n]$	$f[x_{n-2}, x_{n-1}, x_n]$	...	$f[x_0, \ldots, x_n]$
TAB $[n-1]$	$f(x_{n-1})$	$f[x_{n-2}, x_{n-1}]$			
TAB $[n-2]$	$f(x_{n-2})$				
$\vdots$	$\vdots$	$\vdots$	$\vdots$		
TAB $[2]$	$f(x_2)$	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$		
TAB $[1]$	$f(x_1)$	$f[x_0, x_1]$			
TAB $[0]$	$f(x_0)$				

A l'issue de la n-ième étape, la case mémoire TAB[k] contient le coefficient $f[x_0,\dots,x_k]$ cherché, et on peut alors utiliser la formule $(**)$. Il est commode d'appliquer ici la règle de Hörner :

$$p_n(x) = \text{TAB }[0]+(x-x_0)(\text{TAB }[1]+(x-x_1)(\text{TAB }[2]+\ldots+(x-x_{n-1})\text{TAB }[n])))$$

On effectue donc une récurrence descendante

$$\begin{cases} u_n = \text{TAB }[n] \\ u_k = \text{TAB }[k] + (x-x_k)u_{k+1}, \quad 0 \le k < n, \end{cases}$$

qui aboutit à $u_0 = p_n(x)$.

Remarque – D'après l'égalité précédant $(**)$ on a

$$p_k(x_k) - p_{k-1}(x_k) = f[x_0,x_1,\dots,x_k](x_k-x_0)\dots(x_k-x_{k-1}).$$

Or la formule d'erreur 1.2 donne

$$p_k(x_k) - p_{k-1}(x_k) = f(x_k) - p_{k-1}(x_k) = \frac{1}{k!}\,\pi_k(x_k)f^{(k)}(\xi)$$

avec $\pi_k(x) = (x-x_0)\dots(x-x_{k-1})$ et $\xi \in \,]\min(x_0,\dots,x_k),\max(x_0,\dots,x_k)[$. En comparant les deux égalités il vient

$$f[x_0,x_1,\dots,x_k] = \frac{1}{k!}\,f^{(k)}(\xi), \quad \xi \in \,]\min(x_i),\max(x_i)[.$$

Si l'on suppose que $f, f', \dots, f^{(n)}$ existent et sont continues, on voit que les différences divisées $f[x_0,\dots,x_k]$ sont bornées indépendamment du choix des x_i, même si certains de ces points sont très voisins.

1.4. Cas où les points d'interpolation sont équidistants

On considère la subdivision de l'intervalle $[a,b]$ de pas constant $h = \frac{b-a}{n}$. Les points d'interpolation sont donc

$$x_i = a + ih = a + i\,\frac{b-a}{n}, \quad 0 \le i \le n.$$

On note $f_i = f(x_i)$ les valeurs de f correspondantes, et on introduit un opérateur noté Δ, appelée *opérateur aux différences finies*, défini par

$$\Delta : (f_0,f_1,\dots,f_n) \mapsto (\Delta f_0, \Delta f_1, \dots, \Delta f_{n-1})$$

avec

$$\Delta f_i = f_{i+1} - f_i, \quad 0 \le i \le n-1.$$

Lorsqu'on itère l'opération Δ, on obtient des réels $\Delta^k f_i$, $0 \le i \le n-k$, définis par la formule de récurrence

$$\Delta^k f_i = \Delta^{k-1} f_{i+1} - \Delta^{k-1} f_i, \quad k \ge 1, \quad 0 \le i \le n-k,$$

où l'on convient que $\Delta^0 f_i = f_i$, $0 \leq i \leq n$.

Exercice – *Vérifier que* $\Delta^k f_i = \sum_{j=0}^{k} (-1)^j C_k^j f_{i+j}$,

Il est alors facile de montrer par récurrence que les différences divisées sont données par

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{\Delta^k f_i}{k! h^k}.$$

Récrivons avec ces notations la formule fondamentale $(**)$. Pour $x \in [a, b]$, effectuons le changement de variable

$$x = a + sh, \quad s \in [0, n].$$

On a alors

$$\begin{aligned}(x - x_0) \ldots (x - x_{k-1}) &= sh(sh - h) \ldots (sh - (k-1)h) \\ &= h^k s(s-1) \ldots (s - k + 1).\end{aligned}$$

On obtient la *formule de Newton* suivante, dans laquelle $s = (x - a)/h$:

$$\begin{aligned}p_n(x) &= \sum_{k=0}^{n} \Delta^k f_0 \cdot \frac{s(s-1) \ldots (s-k+1)}{k!} \\ &= f_0 + \frac{s}{1} \left(\Delta^1 f_0 + \frac{s-1}{2} \left(\Delta^2 f_0 + \ldots + \frac{s-n+1}{n} \Delta^n f_0 \right) \ldots . \right)\end{aligned}$$

Les coefficients $\Delta^k f_0$ se calculent suivant le schéma décrit au § 1.3 :

$$\begin{array}{ccccccc} f_n & & \Delta f_{n-1} & & \Delta^2 f_{n-2} & \ldots & \Delta^{n-1} f_1 \quad \searrow \; \Delta^n f_0 \\ f_{n-1} & \nearrow\searrow & \Delta f_{n-2} & \nearrow & & & \Delta^{n-1} f_0 \\ f_{n-2} & \nearrow & & & & & \end{array}$$

$$\vdots$$

$$\begin{array}{ccccc} f_2 & & \Delta f_1 & & \Delta^2 f_0 \\ f_1 & \nearrow\searrow & \Delta f_0 & \nearrow & \\ f_0 & \nearrow & & & \end{array}$$

A l'issue de la n-ième étape, le tableau contient les coefficients $\Delta^k f_0$ cherchés.

Estimation de l'erreur d'interpolation – On a

$$\pi_{n+1}(x) = (x - x_0) \ldots (x - x_n) = h^{n+1} s(s-1) \ldots (s-n),$$

d'où

$$f(x) - p_n(x) = \frac{s(s-1) \ldots (s-n)}{(n+1)!} \; h^{n+1} f^{(n+1)}(\xi_x).$$

La fonction $\varphi(s) = |s(s-1)\dots(s-n)|$, $s \in [0,n]$, vérifie $\varphi(n-s) = \varphi(s)$, donc elle atteint son maximum dans $\left[0, \frac{n}{2}\right]$. Comme $\varphi(s-1)/\varphi(s) = (n+1-s)/s > 1$ pour $1 \leq s \leq \frac{n}{2}$, on voit que φ atteint en fait son maximum dans $[0,1]$, d'où

$$\max_{[0,n]} \varphi = \max_{s\in[0,1]} \varphi(s) \leq n!$$

Il en résulte

$$|f(x) - p_n(x)| \leq h^{n+1} \cdot \frac{1}{n+1} \max_{[x_0,\dots,x_n]} |f^{(n+1)}|.$$

Une application typique de ces formules est le calcul d'une valeur approchée de l'image $f(x)$ au moyen d'une table numérique donnant les valeurs successives $f(x_i)$ avec un pas constant h. Supposons par exemple que $h = 10^{-2}$ et que l'on cherche à évaluer $f(x)$ à 10^{-8} près. Une interpolation linéaire (cas $n = 1$) donnerait une erreur en $h^2 = 10^{-4}$ beaucoup trop grande. On doit ici aller jusqu'au degré $n = 3$, ce qui permet d'obtenir un erreur $\leq h^4 = 10^{-8}$ pourvu que $\max |f^{(4)}| \leq 4$.

Remarque – D'après l'expression de π_{n+1} donnée plus haut, on voit que

$$\|\pi_{n+1}\|_{[a,b]} = h^{n+1} \max_{s\in[0,n]} \varphi(s) \leq h^{n+1} n! = \frac{n!}{n^{n+1}} (b-a)^{n+1}$$

et la formule de Stirling $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ montre que l'ordre de grandeur de $\|\pi_{n+1}\|$ est

$$\|\pi_{n+1}\| = O\left(\frac{b-a}{e}\right)^{n+1} \quad \text{quand} \quad n \to +\infty.$$

Comme

$$\varphi\left(\frac{1}{2}\right) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{2} \dots \left(n - \frac{1}{2}\right) \geq \frac{1}{4}\, 1 \cdot 2 \dots (n-1)$$
$$\geq \frac{1}{4n}\, n! \geq \frac{1}{4n} \sqrt{6n}\, \frac{n^n}{e^n} \geq \frac{n^{n-\frac{1}{2}}}{e^{n+1}}$$

pour n grand, on voit en fait que

$$\|\pi_{n+1}\| \geq h^{n+1} \frac{n^{n-\frac{1}{2}}}{e^{n+1}} = \frac{1}{n^{3/2}} \left(\frac{b-a}{e}\right)^{n+1}.$$

Nous obtiendrons au § 2.2 des estimations beaucoup plus précises. L'exercice suivant montre l'intérêt de la formule de Newton en arithmétique.

Exercice

(a) *Montrer que les polynômes de Newton*

$$N_k(s) = \frac{s(s-1)\dots(s-k+1)}{k!}, \quad 0 \leq k \leq n$$

forment une base de $\mathcal{P}_n$ et que pour tout $s \in \mathbb{Z}$ on a $N_k(s) \in \mathbb{Z}$
[Indication : *utiliser les C_n^k ou une récurrence à partir de la relation $N_k(s) - N_k(s-1) = N_{k-1}(s)$*].

(b) *Montrer qu'un polynôme $p \in \mathcal{P}_n$ est tel que $p(s) \in \mathbb{Z}$ pour tout $s \in \mathbb{Z}$ si et seulement si p est combinaison linéaire à coefficients dans $\mathbb{Z}$ de $N_0, \dots, N_n$.*

1.5. Interpolation aux points de Tchebychev

On définit les *polynômes de Tchebychev* par

$$t_n(x) = \cos(n \operatorname{Arccos} x), \quad x \in [-1, 1]$$

Il n'est pas évident *a priori* que t_n est un polynôme ! Pour le voir, on procède comme suit. Posons $\theta = \operatorname{Arc}\cos x$, c'est-à-dire $x = \cos\theta$ avec $\theta \in [0, \pi]$. Il vient alors

$$\begin{aligned} t_n(x) &= \cos n\theta, \\ t_{n+1}(x) + t_{n-1}(x) &= \cos((n+1)\theta) + \cos((n-1)\theta) \\ &= 2\cos n\theta \cos\theta = 2x t_n(x). \end{aligned}$$

La fonction t_n se calcule donc par les formules de récurrence

$$\begin{cases} t_0(x) = 1, \quad t_1(x) = x \\ t_{n+1}(x) = 2x\, t_n(x) - t_{n-1}(x). \end{cases}$$

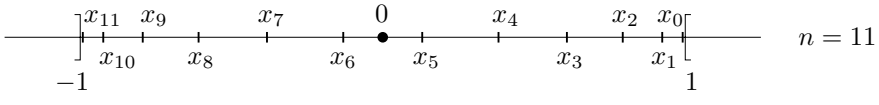
Il en résulte que t_n est un polynôme de degré n, dont le coefficient directeur est 2^{n-1} si $n \geq 1$. Déterminons les racines de t_n. Si $x = \cos\theta \in [-1, 1]$ avec $\theta \in [0, \pi]$, on a $t_n(x) = \cos n\theta = 0$ si et seulement si $n\theta = \frac{\pi}{2} + i\pi$, soit $\theta = \frac{2i+1}{2n}\,\pi$ avec $0 \leq i \leq n-1$. Le polynôme t_n admet donc exactement n racines distinctes :

$$\cos\frac{2i+1}{2n}\,\pi \in\,]-1, 1[, \quad 0 \leq i \leq n-1.$$

Comme t_n est de degré n, il ne peut avoir d'autres racines.

Définition – *Les points d'interpolation de Tchebychev d'ordre n sont les points $x_i = \cos\frac{2i+1}{2n+2}\,\pi$, $0 \leq i \leq n$, racines du polynôme t_{n+1}.*

Les points x_i sont répartis symétriquement autour de 0 (avec $x_{n-i} = -x_i$), de façon plus dense au voisinage de 1 et -1 :



Puisque le coefficient directeur de t_{n+1} est 2^n, il vient

$$t_{n+1}(x) = 2^n \prod_{i=0}^{n} (x - x_i) = 2^n \pi_{n+1}(x).$$

Pour se ramener à un intervalle $[a, b]$ quelconque au lieu de $[-1, 1]$, on utilisera la bijection linéaire

$$\begin{aligned} [-1, 1] &\longrightarrow [a, b] \\ u &\longmapsto x = \frac{a+b}{2} + \frac{b-a}{2}\, u \end{aligned}$$

qui envoie -1 sur a et 1 sur b. Les images des points d'interpolation de Tchebychev $u_i \in]-1,1[$ sont donnés par

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{2i+1}{2n+2}\pi, \quad 0 \leq i \leq n.$$

Ces points sont encore appelés points d'interpolation de Tchebychev d'ordre n de l'intervalle $[a,b]$. Dans ce cas, on a $x - x_i = \frac{b-a}{2}(u - u_i)$, donc le polynôme π_{n+1} est donné par

$$\pi_{n+1}(x) = \prod_{i=0}^{n}(x - x_i) = \left(\frac{b-a}{2}\right)^{n+1} \prod_{i=0}^{n}(u - u_i)$$

où $\prod_{i=0}^{n}(u-u_i) = \frac{1}{2^n} t_{n+1}(u)$ est le polynôme $\pi_{n+1}(u)$ correspondant à $[-1,1]$. On obtient donc

$$\pi_{n+1}(x) = \frac{(b-a)^{n+1}}{2^{2n+1}}\, t_{n+1}(u) = \frac{(b-a)^{n+1}}{2^{2n+1}}\, t_{n+1}\left(\frac{2}{b-a}\left(x - \frac{a+b}{2}\right)\right).$$

Par définition des polynômes de Tchebychev on a $\|t_{n+1}\| = 1$, donc

$$\|\pi_{n+1}\| = 2\left(\frac{b-a}{4}\right)^{n+1}.$$

Cette valeur est beaucoup plus petite que l'estimation $(b-a/e)^{n+1}$ obtenue pour $\|\pi_{n+1}\|$ avec des points x_i équidistants, surtout lorsque n est assez grand : pour $n = 30$ par exemple, on a $(e/4)^{n+1} < 7 \cdot 10^{-6}$.

Il en résulte que l'interpolation aux points de Tchebychev est en général considérablement plus précise que l'interpolation en des points équidistants, d'où son intérêt pratique. Nous reviendrons sur ces questions au § 3.

2. CONVERGENCE DES POLYNÔMES D'INTERPOLATION p_n QUAND n TEND VERS $+\infty$

Soit $f : [a,b] \to \mathbb{R}$ une fonction continue. Pour chaque entier $n \in \mathbb{N}$, on se donne une suite de $n+1$ points $x_{i,n} \in [a,b]$, $0 \leq i \leq n$, deux à deux distincts et on considère le polynôme d'interpolation p_n de f aux points $x_{0,n}, x_{1,n}, \ldots, x_{n,n}$.

Problème – *A quelle(s) condition(s) (portant sur la nature de la fonction f et/ou le choix des points $x_{i,n}$) pourra-t-on être sûr que p_n converge uniformément vers f quand $n \to +\infty$?*

Si l'on ne dispose d'aucune information sur la répartition des points $x_{i,n}$, la meilleure majoration de $\pi_{n+1}(x)$ dont on dispose *a priori* est

$$|\pi_{n+1}(x)| = \prod_{i=0}^{n} |x - x_{i,n}| \leq (b-a)^{n+1}, \quad \forall x \in [a,b],$$

en majorant $|x - x_{i,n}|$ par $b - a$. Dans le cas où les points x_i sont équidistants ou sont les points de Tchebychev, on a bien entendu une meilleure estimation

$$\|\pi_{n+1}\| \leq \Big(\frac{b-a}{e}\Big)^{n+1}, \quad \text{resp. } \|\pi_{n+1}\| \leq 2\Big(\frac{b-a}{4}\Big)^{n+1}.$$

Comme l'erreur d'interpolation dépend par ailleurs de $\|f^{(n+1)}\|$ d'après le § 1.2, on est amené à chercher une majoration des dérivées successives de f.

2.1. Cas d'une fonction analytique

Une fonction *analytique* est par définition une fonction qui est *somme d'une série entière au voisinage de tout point où elle est définie.*

Supposons $f(x) = \sum_{k=0}^{+\infty} a_k x^k$, où la série a un rayon de convergence $R > 0$. La fonction f est donc définie sur $]-R, R[$ au moins. Pour tout $r < R$, la série $\sum a_k r^k$ est convergente, donc la suite $a_k r^k$ est bornée (et tend vers 0), c'est-à-dire qu'il existe une constante $C(r) \geq 0$ telle que

$$|a_k| \leq \frac{C(r)}{r^k}, \quad \forall k \in \mathbb{N}.$$

On peut alors dériver terme à terme $f(x)$ sur $]-r, r[\subset]-R, R[$, ce qui donne

$$\begin{aligned}
f^{(n)}(x) &= \sum_{k=0}^{+\infty} a_k \, \frac{d^n}{dx^n}\,(x^k), \\
|f^{(n)}(x)| &\leq C(r) \sum_{k=0}^{+\infty} \frac{1}{r^k} \, \frac{d^n}{dx^n}\,(x^k) \quad \text{si} \quad x \geq 0 \\
&= C(r) \, \frac{d^n}{dx^n} \left[\sum_{k=0}^{+\infty} \Big(\frac{x}{r}\Big)^k\right] \\
&= C(r) \, \frac{d^n}{dx^n} \left(\frac{1}{1-\frac{x}{r}}\right) = C(r) \, \frac{d^n}{dx^n}\Big(\frac{r}{r-x}\Big) \\
&= \frac{n! r C(r)}{(r-x)^{n+1}}.
\end{aligned}$$

Sur tout intervalle $[-\alpha, \alpha]$ avec $\alpha < r < R$, on a donc

$$\frac{1}{n!} \, \|f^{(n)}\|_{[-\alpha,\alpha]} \leq \frac{rC(r)}{(r-\alpha)^{n+1}}.$$

Supposons maintenant que $f : [a, b] \to \mathbb{R}$ soit somme d'une série entière de centre $c = \frac{a+b}{2}$ et de rayon $R > \alpha = \frac{b-a}{2}$. Pour tout r tel que $\frac{b-a}{2} < r < R$ et tout $n \in \mathbb{N}$ on a alors d'après ce qui précède

$$\frac{1}{n!} \, \|f^{(n)}\|_{[a,b]} \leq \frac{rC(r)}{\Big(r - \frac{b-a}{2}\Big)^{n+1}}.$$

L'erreur d'interpolation admet donc la majoration

$$\|f-p_n\| \leq \frac{1}{(n+1)!}\,\|\pi_{n+1}\|\,\|f^{(n+1)}\| \;\leq 2\Big(\frac{b-a}{\lambda}\Big)^{n+1}\frac{rC(r)}{\Big(r-\frac{b-a}{2}\Big)^{n+2}}$$

$$\leq \frac{2rC(r)}{r-\frac{b-a}{2}}\;\left(\frac{\frac{b-a}{\lambda}}{r-\frac{b-a}{2}}\right)^{n+1}$$

avec respectivement $\lambda = 1$ si les points $x_{i,n}$ sont quelconques, $\lambda = e$ s'ils sont équidistants, $\lambda = 4$ si ce sont les points de Tchebychev. L'erreur va converger vers 0 si l'on peut choisir r tel que $(b-a)/\lambda < r-(b-a)/2$, soit $r > \big(\frac{1}{\lambda}+\frac{1}{2}\big)(b-a)$. Ceci est possible dès que le rayon de convergence R vérifie lui-même cette minoration. On peut donc énoncer :

Théorème – *Soit $f : [a,b] \to \mathbb{R}$ une fonction analytique donnée par une série entière de rayon de convergence R centrée au point $c = \frac{a+b}{2}$. Alors pour des points d'interpolation $x_{i,n}$ quelconques et $\lambda = 1$ (respectivement, équidistants et $\lambda = e$, de Tchebychev et $\lambda = 4$), les polynômes d'interpolation p_n aux points $x_{i,n}$ convergent uniformément vers f pourvu que $R > \big(\frac{1}{\lambda}+\frac{1}{2}\big)(b-a)$.*

Exercice – *Si les points $x_{i,n}$ sont répartis systématiquement par rapport à $c = \frac{a+b}{2}$, montrer que l'on peut prendre $\lambda = 2$.*

Indication : *en supposant $c = 0$ pour simplifier, utiliser le fait que*

$$|(x-x_{i,n})(x+x_{i,n})| \leq \frac{1}{4}\,(b-a)^2$$

pour tout $x \in [a,b]$ et tout $i = 0, 1, \ldots, n$.

Ces résultats sont en fait un peu grossiers, car ils fournissent des conditions suffisantes de convergence qui sont en général très loin d'être nécessaires. Par ailleurs, ce sont des résultats purement théoriques qui ne tiennent aucun compte des erreurs d'arrondi. Nous allons maintenant faire des calculs plus fins sur des exemples, en estimant de façon précise le produit π_{n+1} pour des points équidistants.

2.2.* ESTIMATION DE $\pi_n(z)$, $z \in \mathbb{C}$, POUR DES POINTS D'INTERPOLATION ÉQUIDISTANTS

Posons $h = \frac{b-a}{n}$, $x_j = a_j + jh$, $0 \leq j \leq n$, et soit $z \in \mathbb{C}$,

$$|\pi_{n+1}(z)| = |z-x_i| \cdot \prod_{j\neq 1} |z-x_j|,$$

$$\ln\,|\pi_{n+1}(z)| = \ln\,\delta_n(z) + \sum_{j\neq i} \ln\,|z-x_j|$$

où $\delta_n(z) = |z-x_i|$ est la distance de z au plus proche point x_i. La dernière sommation apparaît comme une somme de Riemann de la fonction $x \mapsto \ln\,|z-x|$. On va donc comparer cette sommation à l'intégrale correspondante.

Lemme 1 – *Pour tout $a \in \mathbb{C}$, on pose $\phi(a) = \int_0^1 \ln|1-at|\,dt$.*
Alors l'intégrale converge et la fonction ϕ est continue sur $\mathbb{C}$. De plus :

(i) $\dfrac{1}{h}\displaystyle\int_{x_j}^{x_{j+1}} \ln|z-x|dx - \ln|z-x_j| = \phi\Big(\frac{h}{z-x_j}\Big)$, $0 \le j \le i-1$;

(ii) $\dfrac{1}{h}\displaystyle\int_{x_j}^{x_{j+1}} \ln|z-x|dx - \ln|z-x_{j+1}| = \phi\Big(-\frac{h}{z-x_{j+1}}\Big)$, $i \le j \le n-1$.

Démonstration. Si $a \notin [1,+\infty]$, la fonction $t \mapsto \ln|1-at|$ est définie et continue sur $[0,1]$. Soit Log la détermination principale du logarithme complexe, définie sur $\mathbb{C} \setminus]-\infty, 0]$. Comme $\ln|z| = \mathrm{Re}(\mathrm{Log}\ z)$, on en déduit aisément

$$\phi(0) = 0, \quad \phi(a) = \mathrm{Re}\Big[\Big(1-\frac{1}{a}\Big)\ \mathrm{Log}\,(1-a)\Big] - 1 \quad \text{si} \quad a \notin \{0\} \cup [1,+\infty[$$

grâce à une intégration par parties. Si $a \in [1,+\infty[$, un calcul analogue donne $\phi(1) = -1$ et $\phi(a) = \big(1-\frac{1}{a}\big)\ln(a-1) - 1$ pour $a > 1$. La continuité de ϕ se vérifie sur ces formules (exercice !)

Égalité (i) : on effectue le changement de variable

$$x = x_j + ht, \quad dx = h\,dt, \quad t \in [0,1].$$

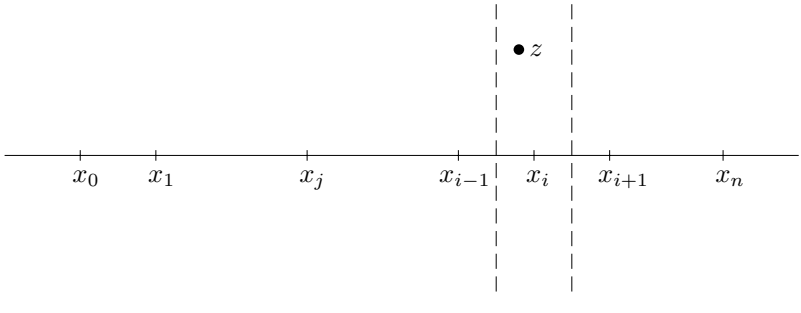
Il vient :

$$\begin{aligned}\frac{1}{h}\int_{x_j}^{x_{j+1}} \ln|z-x|\,dx &= \int_0^1 \ln|z-x_j-ht|\,dt \\ &= \int_0^1 \ln\Big(|z-x_j|\cdot\Big|1-\frac{h}{z-x_j}t\Big|\Big)dt = \ln|z-x_j| + \phi\Big(\frac{h}{z-x_j}\Big).\end{aligned}$$

Égalité (ii) : s'obtient de même en posant $x = x_{j+1} - ht$. ■

En sommant les différentes égalités (i) et (ii), on obtient

$$\frac{1}{h}\int_a^b \ln|z-x|dx - \sum_{j\neq i}\ln|z-x_j| = \sum_{j=0}^{i-1}\phi\Big(\frac{h}{z-x_j}\Big) + \sum_{j=i+1}^{n}\phi\Big(-\frac{h}{z-x_j}\Big). \quad (*)$$



Si $0 \le j < i$, le fait que $|z - x_i| = \min |z - x_k|$ implique $\mathrm{Re}\, z \ge x_i - \frac{h}{2}$ d'où

$$|z - x_j| \ge \mathrm{Re}(z - x_j) \ge x_i - \frac{h}{2} - x_j = \Big(i - j - \frac{1}{2}\Big)h \ge \frac{1}{2}(i-j)h$$

car $\frac{1}{2} \le \frac{1}{2}(i-j)$. Comme $\mathrm{Re}\, w > 0$ implique $\mathrm{Re}\,(1/w) > 0$, on en déduit donc :

$$\mathrm{Re}\Big(\frac{h}{z - x_j}\Big) > 0, \quad \Big|\frac{h}{z - x_j}\Big| \le \frac{2}{i-j} \le 2.$$

Si $i < j \le n$, on obtient de même $\mathrm{Re}\, z \le x_i + \frac{h}{2}$ et

$$|z - x_j| \ge \mathrm{Re}(x_j - z) \ge x_j - x_i - \frac{h}{2} = \Big(j - i - \frac{1}{2}\Big)h \ge \frac{1}{2}(j-i)h,$$
$$\mathrm{Re}\Big(-\frac{h}{z - x_j}\Big) > 0, \quad \Big|\frac{h}{z - x_j}\Big| \le \frac{2}{j-i} \le 2.$$

Lemme 2 – *Pour $a \in \mathbb{C}$ tel que* $\mathrm{Re}\, a \ge 0$ *et* $|a| \le 2$ *on a*

$$\phi(a) = -\frac{1}{2} \ln |1 + a| + O(|a|^2).$$

Démonstration. Les deux membres étant continus sur $\mathrm{Re}\, a \ge 0$, il suffit de montrer l'estimation lorsque a est voisin de 0. On sait que $\mathrm{Log}\,(1+z) = z + O(|z|^2)$ d'où

$$\ln |1 + z| = \mathrm{Re\ Log}(1+z) = \mathrm{Re\ Log}(1+z) = \mathrm{Re}\ z + O(|z|^2),$$
$$\phi(a) = \int_0^1 (-\mathrm{Re}\ a \cdot t + O(|a|^2 t^2))dt = -\frac{1}{2}\ \mathrm{Re}\ a + O(|a|^2),$$

tandis que $\ln |1 + a| = \mathrm{Re}\ a + O(|a|^2)$. Le lemme s'ensuit. ∎

L'égalité $(*)$ ci-dessus et le lemme 2 appliqué avec $a = \pm\frac{h}{z - x_j} = O\big(\frac{1}{j-i}\big)$ impliquent alors

$$\sum_{j \ne i} \ln |z - x_j| - \frac{1}{h}\int_a^b \ln |z - x|dx = \frac{1}{2}\sum_{j=0}^{i-1} \ln \Big|1 + \frac{h}{z - x_j}\Big| + \frac{1}{2}\sum_{j=i+1}^{n} \ln \Big|1 - \frac{h}{z - x_j}\Big|$$
$$+ O\left(\frac{1}{i^2} + \frac{1}{(i-1)^2} + \ldots + \frac{1}{1^2}\right) + O\left(\frac{1}{1^2} + \ldots + \frac{1}{(n-i)^2}\right).$$

Comme la série $\sum_{n=1}^{+\infty} \frac{1}{n^2}$ est convergente, les termes complémentaires sont bornés, c'est-à-dire $O(1)$. De plus

$$1 + \frac{h}{z - x_j} = \frac{z - x_j + h}{z - x_j} = \frac{z - x_{j-1}}{z - x_j},$$
$$1 - \frac{h}{z - x_j} = \frac{z - x_j - h}{z - x_j} = \frac{z - x_{j+1}}{z - x_j}.$$

Dans les deux sommations, les logarithmes se simplifient alors mutuellement, ce qui donne

$$\sum_{j\neq i} \ln|z-x_j| - \frac{1}{h}\int_a^b \ln|z-x|dx = \frac{1}{2}\ln\left|\frac{z-x_{-1}}{z-x_{i-1}}\cdot\frac{z-x_{n+1}}{z-x_{i+1}}\right| + O(1)$$

avec $x_{-1} = a-h$ et $x_{n+1} = b+h$. En prenant l'exponentielle et en multipliant par $|z-x_i|$, on obtient

$$\prod|z-x_j| = |z-x_i|\,\exp\Big(O(1)+\frac{1}{h}\int_a^b \ln|z-x|dx\Big)\cdot\sqrt{\frac{|z-x_1|}{|z-x_{i-1}|}\cdot\frac{|z-x_{n+1}|}{|z-x_{i+1}|}}. \quad (**)$$

La quantité sous la racine est comprise entre 1 et $(1+2n)^2$. En effet on a

$$1 \leq \frac{|z-x_{-1}|}{|z-x_{i-1}|} \leq \frac{|z-x_{i-1}|+ih}{|z-x_{i-1}|} \leq 1+2i,$$

car $|z-x_{i-1}| \geq \mathrm{Re}(z-x_{i-1}) \geq \frac{h}{2}$ si $i \neq 0$, le premier quotient étant égal à 1 si $i=0$. Le deuxième quotient est majoré de même par $1+2(n-i)$. Comme $\exp(O(1))$ est encadré par deux constantes positives et $\frac{1}{h} = \frac{n}{b-a}$, on obtient l'estimation suivante.

Estimation de π_{n+1} – *On pose* $A(z) = \exp\Big(\dfrac{1}{b-a}\displaystyle\int_a^b \ln|z-x|dx\Big)$. *Alors il existe des constantes* $C_1, C_2 > 0$ *telles que*

$$C_1\delta_n(z)A(z)^n \leq |\pi_{n+1}(z)| \leq C_2 n\delta_n(z)A(z)^n.$$

On voit donc que le terme dominant du comportement de $|\pi_{n+1}(z)|$ est le facteur exponentiel $A(z)^n$. Pour évaluer $\|\pi_{n+1}\|_{[a,b]}$, il suffit de calculer $A(x)$ lorsque $x \in [a,b]$:

$$A(x) = \exp\Big(\frac{1}{b-a}\int_a^b \ln|x-t|dt\Big).$$

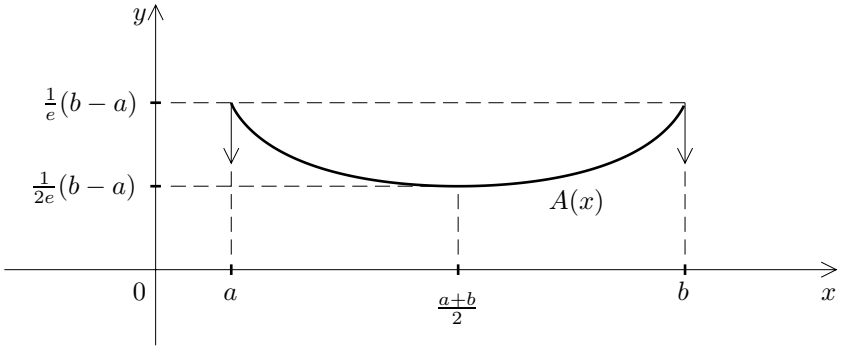
La fonction $t \mapsto \ln|t-x|$ est discontinue en $t=x$, mais le lecteur pourra s'assurer que l'intégration par parties suivante est légitime :

$$\begin{aligned}\int_a^b \ln|t-x|dt &= [(t-x)\ln|t-x|]_a^b - \int_a^b (t-x)\frac{dt}{t-x}\\ &= (b-x)\ln(b-x) + (x-a)\ln(x-a) - (b-a),\end{aligned}$$

car la fonction $t \mapsto (t-x)\ln|t-x|$ est continue sur $[a,b]$ et on peut passer à la limite sur chacun des intervalles $[a, x-\varepsilon]$ et $[x+\varepsilon, b]$. Il en résulte

$$\begin{cases} A(x) = \dfrac{1}{e}(x-a)^{\frac{x-a}{b-a}}(b-x)^{\frac{b-x}{b-a}} & \text{si} \quad x \in\,]a,b[,\\ A(a) = A(b) = \dfrac{1}{e}(b-a). \end{cases}$$

Une étude de $x \mapsto A(x)$ donne l'allure du graphe :



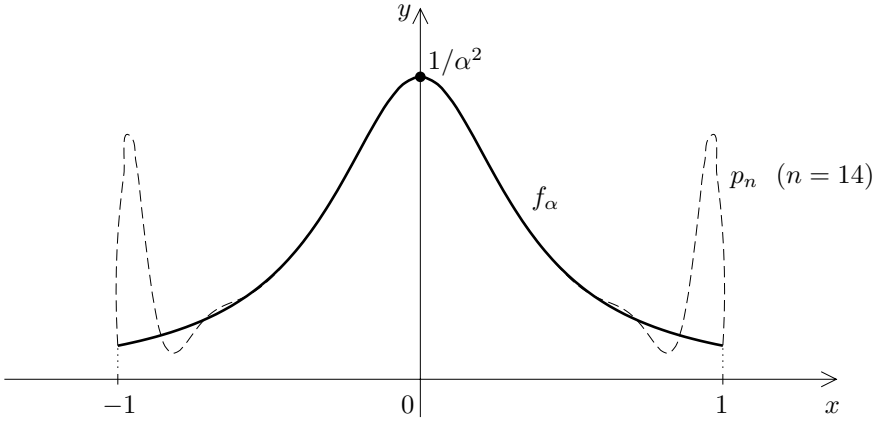
La fonction A atteint donc son maximum $\|A\| = \frac{1}{e}(b-a)$ en $x = a$ ou b et son minimum $\frac{1}{2e}(b-a)$ en $x = \frac{a+b}{2}$. D'un point de vue pratique, il va en résulter que la convergence de $p_n(x)$ est en général beaucoup moins bonne au voisinage des extrémités a, b qu'au centre de l'intervalle.

2.3. Phénomène de Runge*

L'objet de ce paragraphe est de donner un exemple concret de fonction analytique f pour laquelle les polynômes d'interpolation ne forment pas une suite convergente. Nous considérons pour cela la fonction

$$f_\alpha(x) = \frac{1}{x^2 + \alpha^2}, \quad x \in [-1, 1],$$

où $\alpha > 0$ est un paramètre.



Soit p_n le polynôme d'interpolation de f aux points $x_j = -1 + j\frac{1}{n}$, $0 \le j \le n$.

On a ici

$$f_\alpha(x) = \frac{1}{\alpha^2}\ \frac{1}{1+\frac{x^2}{\alpha^2}} = \frac{1}{\alpha^2}\sum_{k=0}^{+\infty}(-1)^k\Big(\frac{x^2}{\alpha^2}\Big)^k$$

avec rayon de convergence $R=\alpha$. D'après le § 2.1, on voit donc que p_n converge uniformément vers f_α dès que $\alpha > 2\big(\frac{1}{2}+\frac{1}{e}\big) \simeq 1,74$. Qu'en est-il si α est petit ?

Calcul de $p_n(x)$ – L'erreur d'interpolation est donnée ici par

$$f_\alpha(x) - p_n(x) = \frac{1}{x^2+\alpha^2} - p_n(x) = \frac{1-(x^2+\alpha^2)p_n(x)}{x^2+\alpha^2}.$$

Le polynôme $1-(x^2+\alpha^2)p_n(x)$ est de degré $\le n+2$, nul aux points $x_0,\dots,x_n$ (puisque p_n interpole f) et égal à 1 aux points $\pm i\alpha$. En particulier $1-(x^2+\alpha^2)p_n(x)$ est divisible par $\pi_{n+1}(x)=\prod(x-x_j)$, le quotient étant de degré 0 ou 1. Examinons la parité de ce quotient.

Comme les points x_j sont répartis symétriquement par rapport à 0, le polynôme p_n est toujours pair, tandis que π_{n+1} est pair si n est impair et vice-versa. Le quotient est un binôme c_0+c_1x, pair si n est impair, impair si n est pair. Par conséquent

$$1-(x^2+\alpha^2)p_n(x) = \begin{cases} c_0\cdot\pi_{n+1}(x) & \text{si } n \text{ est impair,}\\ c_1x\cdot\pi_{n+1}(x) & \text{si } n \text{ est pair.}\end{cases}$$

En substituant $x=i\alpha$, on trouve $c_0 = 1/\pi_{n+1}(i\alpha)$ et $c_1 = 1/i\alpha\pi_{n+1}(i\alpha)$, d'où

$$f_\alpha(x)-p_n(x) = \begin{cases} \dfrac{1}{x^2+\alpha^2}\ \dfrac{\pi_{n+1}(x)}{\pi_{n+1}(i\alpha)} & \text{si } n \text{ est impair,}\\[2ex] \dfrac{x}{i\alpha(x^2+\alpha^2)}\ \dfrac{\pi_{n+1}(x)}{\pi_{n+1}(i\alpha)} & \text{si } n \text{ est pair.}\end{cases}$$

On va maintenant étudier très précisément la convergence ponctuelle de $p_n(x)$, en utilisant les estimations du § 2.2.

Étude de la convergence ponctuelle de la suite $p_n(x)$

Si $x=\pm1$, $p_n(x)=p_n(\pm1)=f_\alpha(\pm1)=\frac{1}{1+\alpha^2}$ est une suite constante. On suppose donc dans la suite que x est un point fixé dans $]-1,1[$ et on cherche à obtenir une estimation de $|\pi_{n+1}(x)/\pi_{n+1}(i\alpha)$. Pour $x=i\alpha$, la formule $(**)$ du § 2.2 montre qu'il existe des constantes $C_3, C_4>0$ telles que

$$C_3A(i\alpha)^n \le |\pi_{n+1}(i\alpha)| \le C_4A(i\alpha)^n$$

car $\alpha \le |i\alpha - x_j| \le \sqrt{\alpha^2+4}$ pour tout $j\in\{-1,\dots,n+1\}$. De même pour $z=x\in\,]-1,1[$, les quantités $|z-x_{i-1}|$ et $|z-x_{i+1}|$ sont du même ordre de grandeur que $h=\frac{2}{n}$, tandis que $|z-x_{-1}|$ et $|z-x_{n+1}|$ tendent respectivement vers $1+x$ et $1-x$. On a donc des constantes positives $C_5, C_6,\dots$ telles que

$$C_5n\delta_n(x)A(x)^n \le |\pi_{n+1}(x)| \le C_6n\delta_n(x)A(x)^n,$$

$$C_7n\delta_n(x)\left(\frac{A(x)}{A(i\alpha)}\right)^n \le |f_\alpha(x)-p_n(x)| \le C_8n\delta_n(x)\left(\frac{A(x)}{A(i\alpha)}\right)^n.$$

Les calculs du § 2.2 donnent

$$A(x) = \frac{1}{e}\,(1+x)^{\frac{1+x}{2}}(1-x)^{\frac{1-x}{2}},$$
$$A(i\alpha) = \exp\left(\frac{1}{2}\int_{-1}^{1} \ln\,|i\alpha - x|dx\right) = \exp\left(\frac{1}{4}\int_{-1}^{1} \ln\,(x^2+\alpha^2)dx\right)$$
$$= \exp\left(\frac{1}{2}\int_{0}^{1} \ln(x^2+\alpha^2)dx\right) = \frac{1}{e}\sqrt{1+\alpha^2}\,\exp\left(\alpha\,\mathrm{Arctg}\,\frac{1}{\alpha}\right),$$

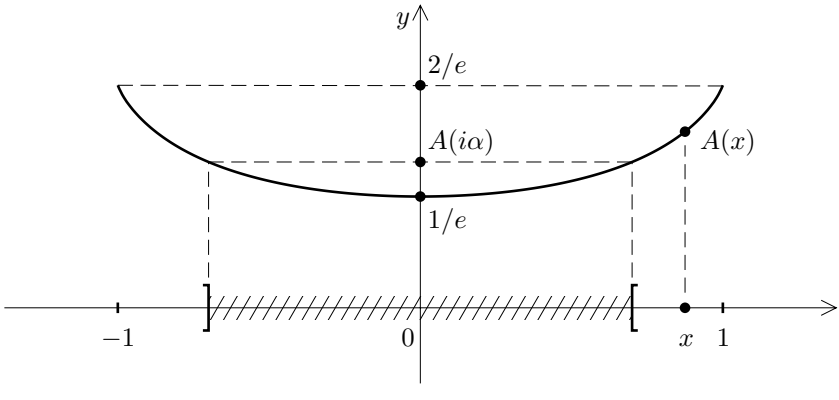
$\ln\,(x^2+\alpha^2)$ ayant pour primitive $x\,\ln\,(x^2+\alpha^2) - 2x + \alpha\ \mathrm{Arctg}\,x/\alpha$. La fonction $\alpha \mapsto A(i\alpha)$ est strictement croissante sur $]0,+\infty[$, avec

$$\lim_{\alpha\to 0} A(i\alpha) = \frac{1}{e}, \quad \lim_{\alpha\to+\infty} A(i\alpha) = +\infty.$$

La valeur critique est la valeur α_0 telle que

$$A(i\alpha_0) = \sup_{x\in[-1,1]} A(x) = \frac{2}{e},$$

soit $\alpha_0 \simeq 0,526$. Pour $\alpha > \alpha_0$, la suite (p_n) converge ponctuellement (et même uniformément) vers f_α sur $[-1,1]$. Pour $\alpha < \alpha_0$, on a le schéma suivant :



• Si $A(x) < A(i\alpha)$ (intervalle ouvert hachuré), $p_n(x)$ converge vers $f_\alpha(x)$.

• Si $x \in\,]-1,1[$ et $A(x) \geq A(i\alpha)$, la suite $(p_n(x))$ diverge comme on le voit à l'aide du lemme suivant.

Lemme – *Pour tout $n \in \mathbb{N}^*$,*

$$\max\,(n\delta_n(x), (n+1)\delta_{n+1}(x)) \geq \frac{1}{2}\,\min(1+x, 1-x) > 0.$$

Il existe en effet des indices j, k tels que

$$\delta_n(x) = |x - x_{j,n}| = \Big|x - \Big(-1 + j \cdot \frac{2}{n}\Big)\Big|,$$
$$\delta_{n+1}(x) = |x - x_{k,n}| = \Big|x - \Big(-1 + k \cdot \frac{2}{n+1}\Big)\Big|.$$

On obtient donc

$$\max\left(n\delta_n(x), (n+1)\delta_{n+1}(x)\right) \geq \frac{1}{2}(n\delta_n(x) + (n+1)\delta_{n+1}(x))$$

$$\begin{aligned}
&\geq \frac{1}{2}\Big(|n(x+1) - 2j| + |(n+1)(x+1) - 2k|\Big) \\
&\geq \frac{1}{2}\ |\text{différence}| = \frac{1}{2}\ |x + 1 - 2k + 2j| \\
&\geq \frac{1}{2}\ \text{distance}\ (x, \text{entiers impairs dans } \mathbb{Z}) \\
&= \frac{1}{2}\ \min(|x-1|, |x+1|).
\end{aligned}$$

■

Grâce au lemme, on voit que

$$\max\ \Big(|f_\alpha(x) - p_n(x)|\,,\ |f_\alpha(x) - p_{n+1}(x)|\Big) \geq C\left(\frac{A(x)}{A(i\alpha)}\right)^n,$$

donc la suite $(|f_\alpha(x) - p_n(x)|)_{n\in\mathbb{N}}$ n'est pas bornée si $A(x) > A(i\alpha)$ et ne tend pas vers 0 si $A(x) = A(i\alpha)$.

Cet exemple montre donc que, même pour une fonction f parfaitement régulière, il ne faut pas s'attendre à ce que les polynômes d'interpolation p_n aux points équidistants convergent vers f sur l'intervalle d'interpolation.

3. Meilleure approximation uniforme

3.1. Existence et unicité du polynôme de meilleure approximation

On munit l'espace vectoriel $\mathcal{C}([a,b])$ des fonctions continues $f : [a,b] \to \mathbb{R}$ de la norme uniforme

$$\|f\| = \sup_{x\in[a,b]} |f(x)|,$$

et de la distance uniforme associée $d(f,g) = \|f - g\|$. On note donc

$$d(f, \mathcal{P}_n) = \inf_{p\in\mathcal{P}_n} \|f - p\|.$$

Théorème et définition – *Pour tout $n \in \mathbb{N}$, il existe un unique polynôme $q_n \in \mathcal{P}_n$ qui réalise le minimum de la distance*

$$\|f - q_n\| = d(f, \mathcal{P}_n)$$

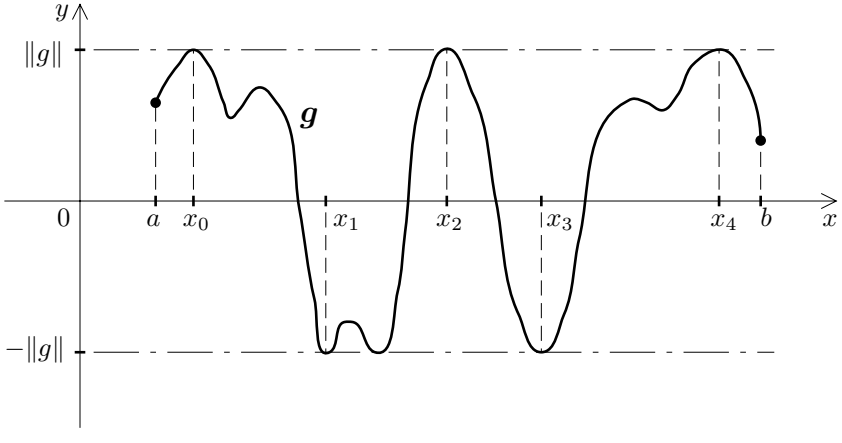
Ce polynôme est appelée polynôme de meilleure approximation uniforme de f à l'ordre n.

Démontrons d'abord l'existence de q_n. En approximant f par $p = 0$, on voit que $d(f, \mathcal{P}_n) \leq \|f\|$. L'ensemble des polynômes $p \in \mathcal{P}_n$ tels que $\|f - p\| \leq \|f\|$ est une partie fermée et bornée $K \subset \mathcal{P}_n$, non vide puisque $0 \in K$. Comme $\mathcal{P}_n$ est de dimension finie, K est une partie compacte, donc la fonction continue $p \mapsto \|f - p\|$ atteint son inf en un point $p = q_n \in K$. ■

Avant de prouver l'unicité, nous introduisons une définition commode.

Définition – *On dit qu'une fonction $g \in \mathcal{C}([a,b])$ équioscille sur $(k+1)$ points de $[a,b]$ s'il existe des points $x_0 < x_1 < \ldots < x_k$ dans $[a,b]$ tels que*

$$\forall i = 0, 1, \ldots, k, \quad |g(x_i)| = \|g\| \quad \text{et} \quad \forall i = 0, 1, \ldots, k-1, \quad g(x_{i+1}) = -g(x_i).$$



Preuve de l'unicité.* Montrons que si $p \in \mathcal{P}_n$ est un polynôme réalisant le minimum de la distance $\|f - p\|$, alors $g = f - p$ équioscille sur $n + 2$ points de $[a,b]$. Si ce n'est pas le cas, soit

$$x_0 = \inf \{x \in [a,b]\,;\ |g(x)| = \|g\|\}$$

le premier point en lequel g atteint sa valeur absolue maximum, puis x_1 le premier point $> x_0$ en lequel $g(x_1) = -g(x_0), \ldots, x_{i+1}$ le premier point $> x_i$ en lequel $g(x_{i+1}) = -g(x_i)$. Supposons que cette suite s'arrête en $i = k \leq n$. D'après le théorème des valeurs intermédiaires, g s'annule nécessairement sur chaque intervalle

$[x_{i-1}, x_i]$. Soit $c_i \in [x_{i-1}, x_i]$ le plus grand réel de cet intervalle tel que $g(c_i) = 0$, de sorte que

$$a \leq x_0 < c_1 < x_1 < c_2 < \ldots < x_{k-1} < c_k < x_k \leq b.$$

Supposons par exemple $g(x_0) > 0$ et posons

$$\pi(x) = (c_1 - x)(c_2 - x)\ldots(c_k - x), \quad \pi \in \mathcal{P}_n,$$
$$g_\varepsilon(x) = g(x) - \varepsilon\pi(x) = f(x) - (p(x) + \varepsilon\pi(x)).$$

On va montrer que $\|g_\varepsilon\| < \|g\|$ pour $\varepsilon > 0$ assez petit, ce qui contredira la minimalité de $\|f - p\|$. Par construction, on a signe$(g(x_i)) = (-1)^i$ et

$$-\|g\| < g(x) \leq \|g\| \quad \text{sur} \quad [a, x_0],$$
$$-\|g\| \leq (-1)^i g(x) < \|g\| \quad \text{sur} \quad [x_{i-1}, c_i]$$

(si on avait seulement $\leq$ au lieu de $<$, alors on aurait $x_i \leq c_i$),

$$0 \leq (-1)^i g(x) \leq \|g\| \quad \text{sur} \quad [c_i, x_i]$$

(si on avait une valeur < 0, $g(x)$ s'annulerait sur $]c_i, x_i[$),

$$-\|g\| < (-1)^k g(x) \leq \|g\| \quad \text{sur} \quad [x_k, b]$$

(si on avait seulement $\leq$ au lieu de $<$, il y aurait un point x_{k+1}).

Il existe donc une constante $A < \|g\|$ positive telle que $g(x) \geq -A$ sur $[a, x_0]$, $(-1)^i g(x) \leq A$ sur $[x_{i-1}, c_i]$ et $(-1)^k g(x) \geq -A$ sur $[x_k, b]$. En notant $M = \sup_{[a,b]} |\pi(x)|$ et en tenant compte du fait que signe$(\pi(x)) = (-1)^i$ sur $]c_i, c_{i+1}[$, on obtient donc

$$\begin{array}{lll} -A - \varepsilon M \leq g_\varepsilon(x) < \|g\| & \text{sur} & [a, x_0], \\ -\|g\| < (-1)^i g_\varepsilon(x) \leq A + \varepsilon M & \text{sur} & [x_{i-1}, c_i], \\ -\varepsilon M \leq (-1)^i g_\varepsilon(x) < \|g\| & \text{sur} & [c_i, x_i], \\ -A - \varepsilon M \leq (-1)^k g_\varepsilon(x) < \|g\| & \text{sur} & [x_k, b], \end{array}$$

ce qui implique $\|g_\varepsilon\| < \|g\|$ dès que ε est assez petit. Cette contradiction entraîne $k \geq n + 1$, ce qu'il fallait démontrer.

Pour vérifier l'unicité de p, il suffit de montrer que pour tout polynôme $q \in \mathcal{P}_n$, $q \neq p$, il existe un point x_i avec $0 \leq i \leq n + 1$ tel que

$$(-1)^i(f(x_i) - q(x_i)) > (-1)^i(f(x_i) - p(x_i)) \ ;$$

ceci entraînera en particulier $\|f - q\| > \|f - p\|$. Sinon, pour tout $i = 0, 1, \ldots, n+1$ on aurait

$$(-1)^i(p(x_i) - q(x_i)) \leq 0.$$

D'après le théorème des valeurs intermédiaires, il existerait un point $\xi_i \in [x_i, x_{i+1}]$ tel que $p(\xi_i) - q(\xi_i) = 0$ pour $i = 0, 1, \ldots, n$. Si les ξ_i sont tous distincts, alors $p - q$ aurait $n + 1$ racines, donc $p = q$ contrairement à l'hypothèse. Or, on peut choisir $\xi_{i-1} < \xi_i$, sauf si dans l'intervalle $[x_{i-1}, x_{i+1}]$ le polynôme $(-1)^i(p(x) - q(x))$ ne

s'annule qu'en $x = x_i$, auquel cas on doit prendre $\xi_{i-1} = x_i = \xi_i$. Dans ce cas $(-1)^i(p(x) - q(x))$ reste ≥ 0 sur $[x_{i-1}, x_{i+1}]$ car son signe est positif en $x = x_{i-1}$ et $x = x_{i+1}$. Ceci entraîne que $\xi_i = x_i$ est racine au moins double de $p - q$, par suite $p - q$ aurait encore $n + 1$ racines compte tenu des multiplicités, contradiction. ■

Observons en outre que d'après la démonstration précédente, le polynôme de meilleure approximation uniforme se caractérise comme suit :

Caractérisation – *Pour $f \in \mathcal{C}([a, b])$, le polynôme de meilleure approximation uniforme $q_n \in \mathcal{P}_n$ de f est l'unique polynôme de degré $\leq n$ tel que $f - q_n$ équioscille sur au moins $(n + 2)$ points de $[a, b]$.*

Exemple – Écrivons les polynômes de Tchebychev sous la forme

$$2^{-n} t_{n+1}(x) = x^{n+1} - q_n(x)$$

avec q_n de degré $\leq n$. Comme $t_{n+1}(\cos\theta) = \cos(n + 1)\theta$ équioscille sur les $n + 2$ points $\theta_i = i\,\frac{\pi}{n+1}$, $0 \leq i \leq n+1$, on en déduit que $q_n(x)$ est le polynôme de meilleure approximation uniforme à l'ordre n de x^{n+1} sur $[-1, 1]$. Autrement dit, $2^{-n} t_{n+1}$ est le polynôme unitaire de degré $n+1$ ayant la plus petite norme uniforme possible sur $[-1, 1]$: cette norme vaut 2^{-n}.

3.2. DENSITÉ DES POLYNÔMES DANS $\mathcal{C}([a, b])$

Il est malheureusement très difficile en général de déterminer le polynôme de meilleure approximation uniforme q_n. C'est pourquoi nous allons étudier ici une méthode beaucoup plus explicite d'approximation.

Définition – *Si $f \in \mathcal{C}([a, b])$, le module de continuité de f est la fonction $\omega_f : \mathbb{R}_+ \to \mathbb{R}_+$ définie par*

$$\omega_f(t) = \sup\{|f(x) - f(y)| \,;\quad x, y \in [a, b] \quad \text{avec} \quad |x - y| \leq t\}.$$

Pour tous $x, y \in [a, b]$, on a alors

$$|f(x) - f(y)| \leq \omega_f(|x - y|),$$

de sorte que ω_f mesure quantitativement la continuité de f.

Propriétés du module de continuité

(i) *$t \mapsto \omega_f(t)$ est une fonction croissante.*

(ii) $\lim\limits_{t \to 0^+} \omega_f(t) = 0$.

(iii) *Pour tous $t_1, t_2 \in \mathbb{R}_+$, $\omega_f(t_1 + t_2) \leq \omega_f(t_1) + \omega_f(t_2)$.*

(iv) *Pour tout $n \in \mathbb{N}$ et tout $t \in \mathbb{R}_+$, $\omega_f(nt) \leq n\,\omega_f(t)$.*

(v) *Pour tout $\lambda \in \mathbb{R}_+$ et tout $t \in \mathbb{R}_+$, $\omega_f(\lambda t) \leq (\lambda + 1)\omega_f(t)$.*

Démonstration. (i) est évident, (ii) résulte du fait que toute fonction continue sur $[a,b]$ y est uniformément continue.

(iii) Soient $x, y \in [a,b]$ quelconques tels que $|x-y| \le t_1+t_2$. Il existe alors $z \in [x,y]$ tel que $|x-z| \le t_1$ et $|z-y| \le t_2$, d'où

$$|f(x)-f(y)| \le |f(x)-f(z)| + |f(z)-f(y)| \le \omega_f(t_1)+\omega_f(t_2).$$

L'inégalité (iii) s'en déduit en prenant le sup sur x, y.

(iv) se déduit immédiatement de (iii) et (v) s'obtient en appliquant (iv) à $n = E(\lambda)+1$. ∎

Nous allons maintenant introduire les polynômes *dits de Jackson*, donnant une assez bonne approximation d'une fonction continue quelconque. Pour tout entier $n \ge 2$, on considère le polynôme trigonométrique $J_n \ge 0$ de degré $n-2$

$$J_n(\theta) = c_n \prod_{1\le k\le n-2} \left(1-\cos\left(\theta - \frac{2k+1}{n}\right)\pi\right)$$

où $c_n > 0$ est fixée telle que $J_n\big(\frac{\pi}{n}\big) = \frac{1}{2}$. En changeant k en $n-1-k$ on voit aussitôt que $J_n(-\theta) = J_n(\theta)$. De plus, pour tout $k \in \mathbb{Z}$, on a $J_n\big(\frac{2k+1}{n}\,\pi\big) = 0$ si $k \not\equiv 0$ ou $-1 \pmod n$, tandis que $J_n\big(\pm\frac{\pi}{n}\big) = \frac{1}{2}$. Nous avons besoin du lemme suivant.

Lemme – *Soit $P(\theta) = \sum_{|j|\le n-1} a_j e^{ij\theta}$, $j \in \mathbb{Z}$, un polynôme trigonométrique de degré au plus $n-1$. Alors*

$$(\forall \theta \in \mathbb{R}) \qquad \sum_{0\le k\le n-1} P\Big(\theta - k\,\frac{2\pi}{n}\Big) = n a_0.$$

En effet $\displaystyle\sum_{0\le k\le n-1} e^{ij(\theta-k2\pi/n)} = e^{ij\theta}\,\frac{1-e^{-ijn2\pi/n}}{1-e^{-ij2\pi/n}} = 0$ si $j \not\equiv 0 \pmod n$, et la somme vaut n si $j = 0$. ∎

Comme $J_n(\theta)$ et $J_n(\theta)(1-\cos\theta)$ sont des polynômes trigonométriques de degré $n-2$ et $n-1$ respectivement, on en déduit que

$$\sum_{0\le k\le n-1} J_n\Big(\theta - k\,\frac{2\pi}{n}\Big) = 1,$$

$$\sum_{0\le k\le n-1} J_n\Big(\theta - k\,\frac{2\pi}{n}\Big)\Big(1-\cos\Big(\theta - k\,\frac{2\pi}{n}\Big)\Big) = 1-\cos\frac{\pi}{n}\ ;$$

en effet, d'après le lemme, ces sommes sont des constantes et pour $\theta = -\frac{\pi}{n}$ la définition de $J_n(\theta)$ montre que $J_n\big(\theta - k\,\frac{2\pi}{n}\big) = 0$ sauf pour $k = 0$ et $k = n-1$, auquel cas $J_n\big(\theta - k\,\frac{2\pi}{n}\big) = \frac{1}{2}$. Observons de plus que

$$\begin{aligned}\Big|\cos\theta - \cos k\,\frac{2\pi}{n}\Big|^2 &= \Big|\operatorname{Re}\big(e^{i\theta} - e^{ik\,2\pi/n}\big)\Big|^2 \\ &\le \Big|e^{i\theta} - e^{ik\,2\pi/n}\Big|^2 = \Big|e^{i\left(\theta - k\,2\pi/n\right)} - 1\Big|^2 \\ &= 2\Big(1-\cos\Big(\theta - k\,\frac{2\pi}{n}\Big)\Big).\end{aligned}$$

En appliquant l'inégalité de Cauchy-Schwarz $|\sum a_k b_k| \le (\sum a_k^2)^{1/2}(\sum b_k^2)^{1/2}$ aux quantités

$$a_k = J_n\Big(\theta - k\,\frac{2\pi}{n}\Big)^{1/2}, \quad b_k = J_n\Big(\theta - k\,\frac{2\pi}{n}\Big)^{1/2}\Big|\cos\theta - \cos\,k\,\frac{2\pi}{n}\Big|,$$

on en déduit

$$\begin{aligned}
&\sum_{0\le k\le n-1} J_n\Big(\theta - k\,\frac{2\pi}{n}\Big)\Big|\cos\theta - \cos\,k\,\frac{2\pi}{n}\Big| \\
&\qquad \le \Big(\sum J_n\Big(\theta - k\,\frac{2\pi}{n}\Big)\Big)^{1/2}\cdot\Big(\sum J_n\Big(\theta - k\,\frac{2\pi}{n}\Big)\Big|\cos\theta - \cos\,k\,\frac{2\pi}{n}\Big|^2\Big)^{1/2} \\
&\qquad \le \Big(2\sum J_n\Big(\theta - k\,\frac{2\pi}{n}\Big)\Big(1 - \cos\Big(\theta - k\,\frac{2\pi}{n}\Big)\Big)\Big)^{1/2} \\
&\qquad \le \Big(2\Big(1 - \cos\,\frac{\pi}{n}\Big)\Big)^{1/2} = 2\,\sin\,\frac{\pi}{2n} \le \frac{\pi}{n}. \qquad (*)
\end{aligned}$$

Soit maintenant $f \in \mathcal{C}([-1,1])$ une fonction continue quelconque. On lui associe le polynôme trigonométrique de degré $\le n-2$

$$\varphi_n(\theta) = \sum_{0\le k\le n-1} f\Big(\cos\,k\,\frac{2\pi}{n}\Big)J_n\Big(\theta - k\,\frac{2\pi}{n}\Big).$$

En changeant k en $n-k$ pour $1 \le k \le n-1$, on voit que $\varphi_n(-\theta) = \varphi_n(\theta)$, par conséquent $\varphi_n(\theta)$ est combinaison linéaire des fonctions paires 1, $\cos\theta$, …, $\cos(n-2)\theta$. Comme celles-ci sont précisément données par les polynômes de Tchebychev $t_k(\cos\theta)$, on voit qu'il existe un polynôme p_{n-2} de degré $\le n-2$ tel que $\varphi_n(\theta) = p_{n-2}(\cos\theta)$. Pour des raisons similaires, il existe un polynôme $j_{n,k}(x)$ de degré $\le n-2$ tel que

$$\frac{1}{2}\,\Big(J_n\Big(\theta + k\,\frac{2\pi}{n}\Big) + J_n\Big(\theta - k\,\frac{2\pi}{n}\Big)\Big) = j_{n,k}(\cos\theta).$$

En observant que $p_{n-2}(\cos\theta) = \frac{1}{2}\big(\varphi_n(\theta) + \varphi_n(-\theta)\big)$ et en substituant $x = \cos\theta$, on peut exprimer explicitement le polynôme p_{n-2} à partir des $j_{n,k}$:

$$p_{n-2}(x) = \sum_{0\le k\le n-1} f\Big(\cos\,k\frac{2\pi}{n}\Big)j_{n,k}(x), \quad \forall x \in [-1,1].$$

Ce polynôme sera appelé polynôme d'approximation de Jackson de degré $n-2$ de f. Cela étant, nous avons le :

Théorème de Jackson – *Pour tout $f \in \mathcal{C}([a,b])$, les polynômes d'approximation de Jackson vérifient*

$$\|f - p_n\| \le 3\,\omega_f\Big(\frac{b-a}{n+2}\Big).$$

Démonstration. Le cas d'un intervalle $[a,b]$ quelconque se ramène facilement au cas où $[a,b]=[-1,1]$, en utilisant le même changement de variable qu'au § 1.5. On suppose donc $f\in\mathcal{C}([-1,1])$ et on cherche à majorer

$$\|f-p_{n-2}\|_{[-1,1]}=\sup_{\theta\in[0,\pi]}|f(\cos\theta)-\varphi_n(\theta)|.$$

La propriété $\sum J_n\big(\theta-k\,\frac{2\pi}{n}\big)=1$ permet d'écrire

$$f(\cos\theta)=\sum_{0\le k\le n-1} f(\cos\theta)J_n\Big(\theta-k\,\frac{2\pi}{n}\Big),$$

d'où

$$f(\cos\theta)-\varphi_n(\theta)=\sum_{0\le k\le n-1}\Big(f(\cos\theta)-f\Big(\cos k\frac{2\pi}{n}\Big)\Big)\,J_n\Big(\theta-k\,\frac{2\pi}{n}\Big)$$

par définition de φ_n. La propriété (v) du module de continuité avec $t=\frac{2}{n}$ et $\lambda=\frac{n}{2}\Big|\cos\theta-\cos k\frac{2\pi}{n}\Big|$ implique

$$\begin{aligned}\Big|f(\cos\theta)-f\Big(\cos k\frac{2\pi}{k}\Big)\Big|&\le\omega_f(\lambda t)\\&\le\Big(1+\frac{n}{2}\Big|\cos\theta-\cos k\frac{2\pi}{n}\Big|\Big)\omega_f\Big(\frac{2}{n}\Big),\end{aligned}$$

par conséquent l'inégalité $(*)$ donne

$$\begin{aligned}|f(\cos\theta)-\varphi_n(\theta)|&\le\left[\sum_{0\le k\le n-1}J_n\Big(\theta-k\,\frac{2\pi}{n}\Big)\Big(1+\frac{n}{2}\Big|\cos\theta-\cos k\frac{2\pi}{n}\Big|\Big)\right]\omega_f\Big(\frac{2}{n}\Big)\\&\le\Big(1+\frac{n}{2}\cdot\frac{\pi}{n}\Big)\omega_f\Big(\frac{2}{n}\Big).\end{aligned}$$

On obtient donc finalement

$$\|f-p_{n-2}\|\le\Big(1+\frac{\pi}{2}\Big)\omega_f\Big(\frac{2}{n}\Big)\le 3\,\omega_f\Big(\frac{2}{n}\Big).$$ ■

Il résulte du théorème de Jackson que (p_n) converge uniformément vers f quand n tend vers $+\infty$. Comme le polynôme de meilleure approximation q_n satisfait par définition $\|f-q_n\|\le\|f-p_n\|$, on en déduit que (q_n) converge uniformément vers f quand n tend vers $+\infty$. Ceci équivaut à l'énoncé suivant :

Théorème de Weierstrass – *L'espace $\mathcal{P}$ des polynômes est dense dans $\mathcal{C}([a,b])$ pour la norme uniforme.*

4. STABILITÉ NUMÉRIQUE DU PROCÉDÉ D'INTERPOLATION DE LAGRANGE

4.1. CONSTANTE DE LEBESGUE ASSOCIÉE AUX POINTS D'INTERPOLATION

Soient $x_0, x_1, \ldots, x_n \in [a,b]$ des points 2 à 2 distincts. On considère l'opérateur d'interpolation de Lagrange

$$\begin{aligned} L_n : \ \mathcal{C}([a,b]) &\longrightarrow \mathcal{P}_n \\ f &\longmapsto p_n. \end{aligned}$$

Dans la pratique, la fonction f à interpoler n'est pas connue exactement : on ne dispose que d'une valeur approchée $\widetilde{f} = f+g$, où g est un terme d'erreur. Au lieu de calculer $p_n = L_n(f)$, on va donc calculer $\widetilde{p}_n = L_n(\widetilde{f}) = L_n(f)+L_n(g) = p_n+L_n(g)$. Si g est l'erreur commise sur f, l'erreur sur p_n sera donc $L_n(g)$. D'un point de vue numérique, il va être très important de pouvoir estimer $\|L_n(g)\|$ en fonction de $\|g\|$. Rappelons la formule d'interpolation (*) démontrée au § 1.1 : si $L_n(g) = r_n$, alors

$$r_n(x) = \sum_{i=0}^{n} g(x_i) l_i(x).$$

On a donc

$$|r_n(x)| \leq \left(\sum_{i=0}^{n} |l_i(x)| \right) \|g\|.$$

Théorème et définition – *La norme de l'opérateur d'interpolation L_n est*

$$\Lambda_n = \sup_{x\in[a,b]} \left(\sum_{i=0} |l_i(x)| \right).$$

Le nombre Λ_n est appelé constante de Lebesgue associée à $x_0, x_1, \ldots, x_n$.

Démonstration. D'après ce qui précède, on a $\|L_n(g)\| = \|r_n\| \leq \Lambda_n \|g\|$, donc $|||L_n||| \leq \Lambda_n$. Réciproquement, la continuité des l_i entraîne qu'il existe un point $\xi \in [a,b]$ tel que $\Lambda_n = \sum_{i=0}^{n} |l_i(\xi)|$. On peut trouver une fonction $g \in \mathcal{C}([a,b])$ affine par morceaux, telle que $\|g\| = 1$ et $g(x_i) = \pm 1 = \text{signe}\,(l_i(\xi))$. Alors

$$L_n(g)(\xi) = \sum_{i=0}^{n} |l_i(\xi)| = \Lambda_n,$$

de sorte que $\|L_n(g)\| \geq \Lambda_n$ et $|||L_n||| \geq \Lambda_n$. ∎

Intuitivement, la constante Λ_n peut s'interpréter comme le facteur d'*amplification de l'erreur* dans le procédé d'interpolation de Lagrange. On va voir que Λ_n est

également liée au problème de la convergence des polynômes d'interpolation, grâce à l'inégalité suivante :

Théorème – *Pour tout $f \in \mathcal{C}([a,b])$, on a*

$$\|f - L_n(f)\| \leq (1 + \Lambda_n) d(f, \mathcal{P}_n).$$

Démonstration. Soit q_n le polynôme de meilleure approximation uniforme de f, de sorte que $\|f - q_n\| = d(f, \mathcal{P}_n)$. Puisque $q_n \in \mathcal{P}_n$, on a $L_n(q_n) = q_n$, donc

$$\begin{aligned} f - L_n(f) &= f - q_n - L_n(f - q_n) \\ \|f - L_n(f)\| &\leq \|f - q_n\| + \|L_n(f - q_n)\| \\ &\leq \|f - q_n\| + \Lambda_n \|f - q_n\| = (1 + \Lambda_n) d(f, \mathcal{P}_n). \end{aligned}$$

■

4.2. Cas où les points x_i sont équidistants

Posons $x_i = a + ih$, $0 \leq i \leq n$ et $x = a + sh$, où $s \in [0, n]$, $h = \frac{b-a}{n}$. On a alors

$$\begin{aligned} l_i(x) \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)} &= \prod_{j \neq i} \frac{s - j}{i - j} \\ &= (-1)^{n-i} \frac{s(s-1)\dots\widehat{(s-i)}\dots(s-n)}{i!(n-i)!}, \end{aligned}$$

où $\widehat{s-i}$ désigne un facteur omis. On peut démontrer à partir de là que

$$\Lambda_n \sim \frac{2^{n+1}}{en \ln(n)}.$$

Nous nous contenterons de démontrer une minoration de Λ_n. Pour $s = \frac{1}{2}$, c'est-à-dire pour $x = a + \frac{h}{2}$, il vient

$$\begin{aligned} |l_i(x)| &= \frac{\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{2} \dots \widehat{\left(i - \frac{1}{2}\right)} \dots \left(n - \frac{1}{2}\right)}{i!(n-i)!} \\ &\geq \frac{1}{4} \cdot \frac{1 \cdot 2 \dots \widehat{(i-1)} \dots (n-1)}{i!(n-i)!} \geq \frac{1}{4n^2} \, \frac{n!}{i!(n-i)!}. \end{aligned}$$

On en déduit

$$\Lambda_n \geq \sum_{i=0}^{n} |l_i(x)| \geq \frac{1}{4n^2} \sum_{i=0}^{n} C_n^i = \frac{1}{4n^2} 2^n.$$

Comme Λ_n tend vers $+\infty$ assez rapidement, on voit que l'interpolation de Lagrange en des points équidistants *n'est pas une méthode numérique très stable* : les erreurs sont fortement amplifiées lorsque n est grand. Comme Λ_n tend en général nettement plus vite vers $+\infty$ que $d(f, \mathcal{P}_n)$ ne tend vers 0 (cf. Th. de Jackson), le théorème

ci-dessus donne également une indication de la raison pour laquelle le phénomène de Runge se produit.

Exercice – *Si $x = a + sh$ avec $s \in [k, k+1]$, $0 \leq k < n$, montrer que $|l_i(x)| \leq \frac{(n-k)!(k+1)!}{i!(n-i)!} \leq \frac{n!}{i!(n-i)!}$. En déduire que $\Lambda_n \leq 2^n$.*

4.3. CAS DES POINTS D'INTERPOLATION DE TCHEBYCHEV

Il est facile de voir que la constante Λ_n reste inchangée si l'on effectue un changement affine de coordonnées $x \mapsto \alpha x + \beta$. On se placera donc pour simplifier sur l'intervalle $[-1, 1]$. Dans ce cas, comme $\pi_{n+1}(x) = 2^{-n} t_{n+1}(x)$ d'après le § 1.5, le polynôme d'interpolation d'une fonction $f \in \mathcal{C}([-1, 1])$ est donné par

$$P_n(x) = \sum_{i=0} f(x_i) l_i(x)$$

avec

$$l_i(x) = \frac{\pi_{n+1}(x)}{(x - x_i)\pi'_{n+1}(x_i)} = \frac{t_{n+1}(x)}{(x - x_i)t'_{n+1}(x_i)}.$$

Estimation de la constante de Lebesgue – On peut montrer que

$$\Lambda_n \sim \frac{2}{\pi} \ln(n) \quad \text{quand} \quad n \to +\infty.$$

Nous nous contenterons de vérifier que $\Lambda_n \leq C \ln(n)$, où C est une constante positive, et laisserons au lecteur l'initiative de raffiner la méthode pour obtenir le résultat plus précis ci-dessus. Posons

$$x = \cos\theta, \quad x_i = \cos\theta_i \quad \text{où} \quad \theta_i = \frac{2i+1}{2n+2}\pi, \quad 0 \leq i \leq n.$$

La relation $t_{n+1}(\cos\theta) = \cos(n+1)\theta$ entraîne par dérivation

$$\sin\theta\, t'_{n+1}(\cos\theta) = (n+1)\sin(n+1)\theta.$$

Comme $\sin(n+1)\,\theta_i = \sin(2i+1)\frac{\pi}{2} = (-1)^i$, il vient

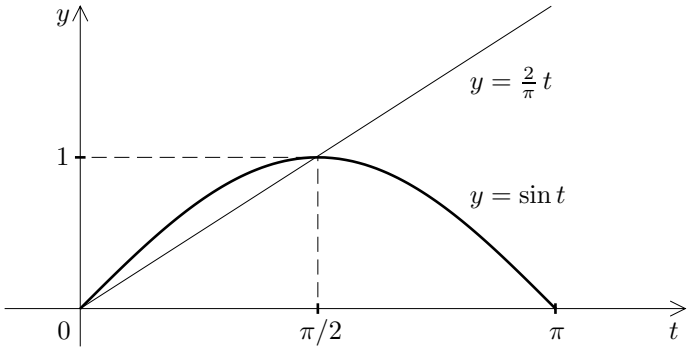
$$t'_{n+1}(x_i) = (n+1)\,\frac{(-1)^i}{\sin\theta_i},$$

$$l_i(\cos\theta) = \frac{(-1)^i \sin\theta_i \cos(n+1)\theta}{(n+1)(\cos\theta - \cos\theta_i)},$$

$$|l_i(\cos\theta)| = \frac{|\sin\theta_i \cos(n+1)\,\theta|}{(n+1)|\cos\theta - \cos\theta_i|}.$$

Minorons la quantité

$$\cos\theta - \cos\theta_i = -2\sin\frac{\theta - \theta_i}{2}\sin\frac{\theta + \theta_i}{2}.$$



Pour $t \in \left[0, \frac{\pi}{2}\right]$ on a $\sin t \geq \frac{2}{\pi}\, t$, or

$$\frac{\theta - \theta_i}{2} \in \Big[-\frac{\pi}{2}, \frac{\pi}{2}\Big], \quad \text{donc} \quad \Big|\sin \frac{\theta - \theta_i}{2}\Big| \geq \frac{2}{\pi}\ \frac{|\theta - \theta_i|}{2}.$$

Par ailleurs $\frac{\theta+\theta_i}{2} \in \left[\frac{\theta_i}{2}, \frac{\theta_i+\pi}{2}\right]$ avec $\frac{\theta_i}{2} \leq \frac{\pi}{2}$ et $\frac{\theta_i+\pi}{2} \geq \frac{\pi}{2}$, donc

$$\Big|\sin \frac{\theta + \theta_i}{2}\Big| \geq \min\Big(\sin \frac{\theta_i}{2}, \sin \frac{\theta_i + \pi}{2}\Big) = \min\Big(\sin \frac{\theta_i}{2}, \cos \frac{\theta_i}{2}\Big).$$

Comme $\sin \theta_i = 2 \sin \frac{\theta_i}{2} \cos \frac{\theta_i}{2} \leq 2 \min\Big(\sin \frac{\theta_i}{2}, \cos \frac{\theta_i}{2}\Big)$, on obtient

$$|l_i(\cos \theta)| \leq \pi \frac{|\cos (n+1)\theta|}{(n+1)|\theta - \theta_i|}. \qquad (*)$$

D'après le théorème des accroissements finis

$$\cos (n+1)\theta = \cos (n+1)\theta - \cos (n+1)\theta_i = (n+1)(\theta - \theta_i)(-\sin \xi),$$
$$|\cos (n+1)\theta| \leq (n+1)|\theta - \theta_i|,$$

donc $|l_i(\cos \theta)| \leq \pi$, $\forall \theta \in [0, \pi] \setminus \{\theta_i\}$, et ceci est encore vrai par continuité si $\theta = \theta_i$. Fixons $\theta \in [0, \pi]$ et soit θ_j le point le plus proche de θ. Si on note $h = \frac{\pi}{n+1} = \theta_{i+1} - \theta_i$, alors on a

$$|\theta - \theta_j| \leq \frac{h}{2},$$
$$|\theta - \theta_i| \geq |\theta_j - \theta_i| - |\theta - \theta_j| \geq (|j - i| - 1)h.$$

L'inégalité $(*)$ donne

$$\sum_{i=0}^{n} |l_i(\cos \theta)| \leq \frac{\pi}{(n+1)h} \sum_{j \neq i, i+1, i-1} \frac{1}{|j - i| - 1} + 3\pi,$$

d'où $\Lambda_n \leq 2\Big(1 + \frac{1}{2} + \ldots + \frac{1}{n}\Big) + 3\pi \leq C \ln (n)$. ■

Exercice – *Montrer inversement que*

$$\sum_{i=0}^{n} |l_i(1)| = \frac{1}{n+1} \sum_{i=0}^{n} \operatorname{cotan} \frac{\theta_i}{2} \geq \frac{2}{\pi} \int_{\theta_0/2}^{\pi/2} \operatorname{cotan}\, t\, dt \geq \frac{2}{\pi} \ln(n).$$

D'après le théorème du § 4.1 et le théorème de Jackson, on obtient pour tout $f \in \mathcal{C}([a,b])$:

$$\|f - L_n(f)\| \leq (1 + \Lambda_n) d(f, \mathcal{P}_n) \leq C' \ln(n) \cdot \omega_f\Big(\frac{b-a}{n+2}\Big).$$

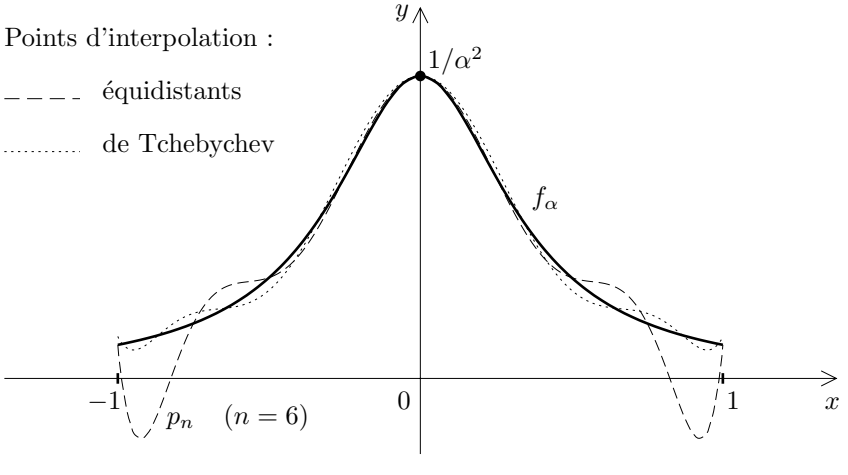
Corollaire – *On suppose que f est lipschitzienne, c'est-à-dire qu'il existe une constante $K \geq 0$ telle que $\forall x, y \in [a,b]$ on ait $|f(x) - f(y)| \leq K(x-y)$.*
Alors la suite $L_n(f)$ des polynômes d'interpolation de Tchebychev converge uniformément vers f sur $[a,b]$.

Sous ces hypothèses on a en effet $\omega_f(t) \leq Kt$, donc

$$\|f - L_n(f)\| \leq KC'(b-a)\, \frac{\ln(n)}{n+2},$$

ce qui tend vers 0 quand n tend vers $+\infty$. ■

Ces résultats montrent que *l'interpolation aux points de Tchebychev est considérablement plus fiable que l'interpolation en des points équidistants.* Le schéma ci-dessous compare à titre d'exemple les polynômes d'interpolation de degré 6 associés à la fonction $f_x(x) = 1/(x^2 + \alpha^2)$ pour $\alpha = \sqrt{8}$ (voir aussi le §2.3).



5. Polynômes orthogonaux

Soit $]a,b[$ un intervalle ouvert borné ou non dans $\mathbb{R}$. On se donne *un poids* sur $]a,b[$, c'est-à-dire une fonction $w :]a,b[\to]0,+\infty[$ continue. On suppose en outre que pour tout entier $n \in \mathbb{N}$ l'intégrale $\displaystyle\int_a^b |x|^n w(x)dx$ est convergente ; c'est le cas par exemple si $]a,b[$ est borné et si $\displaystyle\int_a^b w(x)dx$ converge. Sous ces hypothèses, on considère l'espace vectoriel E des fonctions continues sur $]a,b[$ telles que

$$\|f\|_2 = \sqrt{\int_a^b |f(x)|^2 w(x)dx} < +\infty.$$

Grâce aux hypothèses faites ci-dessus, E contient l'espace vectoriel des fonctions polynômes. L'espace E est muni d'un produit scalaire naturel

$$\langle f,g\rangle = \int_a^b f(x)g(x)w(x)dx,$$

et $\|\ \|_2$ est la norme associée à ce produit scalaire ; cette norme est appelée *norme L^2* ou *norme moyenne quadratique*. On notera $d_2(f,g) = \|f-g\|_2$ la distance associée.

Théorème 1 – *Il existe une suite de polynômes unitaires $(p_n)_{n\in\mathbb{N}}$, $\deg(p_n) = n$, orthogonaux 2 à 2 pour le produit scalaire de E. Cette suite est unique. Les polynômes p_n sont appelés* polynômes orthogonaux pour le poids w.

Démonstration. On construit p_n par récurrence à l'aide du procédé d'orthogonalisation de Schmidt. On a $p_0(x) = 1$, puisque p_0 doit être unitaire.

Supposons $p_0, p_1, \ldots, p_{n-1}$ déjà construits. Comme $\deg p_i = i$, ces polynômes forment une base de $\mathcal{P}_{n-1}$. On peut donc chercher p_n sous la forme

$$p_n(x) = x^n - \sum_{j=0}^{n-1} \lambda_{j,n} p_j(x).$$

La condition $\langle p_n, p_k\rangle = 0$ pour $k = 0,1,\ldots,n-1$ donne

$$\begin{aligned}\langle p_n,p_k\rangle = 0 &= \langle x^n,p_k\rangle - \sum_{j=0}^{n-1}\lambda_{j,n}\langle p_j,p_k\rangle \\ &= \langle x^n,p_k\rangle - \lambda_{k,n}\|p_k\|_2^2.\end{aligned}$$

On a donc un et un seul choix possible, à savoir

$$\lambda_{k,n} = \frac{\langle x^n,p_k\rangle}{\|p_k\|_2^2}.$$

Remarque – La suite p_n ainsi construite n'est pas orthonormée en général. La suite normalisée $\widetilde{p}_n = \frac{1}{\|p_n\|_2}\, p_n$ est une base orthonormée de l'espace $\mathcal{P}$ des polynômes.

Théorème 2 – *Les polynômes p_n vérifient la relation de récurrence*

$$p_n(x) = (x - \lambda_n)p_{n-1}(x) - \mu_n p_{n-2}(x), \quad n \geq 2$$

avec

$$\lambda_n = \frac{\langle xp_{n-1}, p_{n-1}\rangle}{\|p_{n-1}\|_2^2}, \quad \mu_n = \frac{\|p_{n-1}\|_2^2}{\|p_{n-2}\|_2^2}.$$

Démonstration. Le polynôme xp_{n-1} est unitaire de degré n, donc on peut écrire

$$xp_{n-1} = p_n + \sum_{k=0}^{n-1} \alpha_k p_k,$$

où $\langle xp_{n-1}, p_k\rangle = \alpha_k \|p_k\|_2^2$, $0 \leq k \leq n-1$. Par définition du produit scalaire, on a

$$\langle xp_{n-1}, p_k\rangle = \langle p_{n-1}, xp_k\rangle = \int_a^b x\, p_{n-1}(x)p_k(x)w(x)\, dx.$$

Si $k \leq n-3$, $xp_k \in \mathcal{P}_{n-2}$, donc $\langle p_{n-1}, xp_k\rangle = 0$. Il y a donc au plus deux coefficients non nuls :

$$\alpha_{n-1} = \frac{\langle xp_{n-1}, p_{n-1}\rangle}{\|p_{n-1}\|_2^2} = \lambda_n, \quad \alpha_{n-2} = \frac{\langle p_{n-1}, xp_{n-2}\rangle}{\|p_{n-2}\|_2^2}.$$

Or $xp_{n-2} = p_{n-1} + q$, $q \in \mathcal{P}_{n-2}$, donc

$$\langle p_{n-1}, xp_{n-2}\rangle = \|p_{n-1}\|_2^2 + \langle p_{n-1}, q\rangle = \|p_{n-1}\|_2^2,$$

ce qui donne $\alpha_{n-2} = \mu_n$ et

$$xp_{n-1} = p_n + \lambda_n p_{n-1} + \mu_n p_{n-2}.$$ ■

Exemples – Certains cas particuliers ont donné lieu à des études plus poussées. Mentionnons entre autres les cas suivants :

- $]a, b[=]0, +\infty[$, $w(x) = e^{-x}$, p_n = polynômes de Laguerre ;
- $]a, b[=]-\infty, +\infty[$, $w(x) = e^{-x^2}$, p_n = polynômes de Hermite ;
- $]a, b[=]-1, 1[$, $w(x) = 1$, p_n = polynômes de Legendre ;
- $]a, b[=]-1, 1[$, $w(x) = \frac{1}{\sqrt{1-x^2}}$, p_n = polynômes de Tchebychev.

Vérifions en effet que les polynômes de Tchebychev t_n sont 2 à 2 orthogonaux relativement au poids $w(x) = 1/\sqrt{1-x^2}$. Le changement de variable $x = \cos\theta$, $\theta \in [0, \pi]$ donne :

$$\int_{-1}^{1} t_n(x)t_k(x)\ \frac{1}{\sqrt{1-x^2}}\,dx = \int_0^{\pi} t_n(\cos\theta)t_k(\cos\theta)\,d\theta$$
$$= \int_0^{\pi} \cos n\theta \cdot \cos k\theta\,d\theta = \begin{cases} 0 & \text{si} \quad n \neq k \\ \frac{\pi}{2} & \text{si} \quad n = k \neq 0. \\ \pi & \text{si} \quad n = k = 0 \end{cases}$$

Comme t_n a pour coefficient directeur 2^{n-1} si $n \geq 1$, on en déduit

$$\begin{cases} p_0(x) = t_0(x) = 1 \\ p_n(x) = 2^{1-n}t_n(x) \quad \text{si} \quad n \geq 1. \end{cases}$$

On sait que t_n a n zéros distincts dans $]-1, 1[$. On va voir que c'est une propriété générale des polynômes orthogonaux.

Théorème 3 – *Pour tout poids w sur $]a, b[$, le polynôme p_n possède n zéros distincts dans l'intervalle $]a, b[$.*

Démonstration. Soient $x_1, \ldots, x_k$ les zéros distincts de p_n contenus dans $]a, b[$ et $m_1, \ldots, m_k$ leurs multiplicités respectives. On a $m_1 + \ldots + m_k \leq \deg p_n = n$. Posons $\varepsilon_i = 0$ si m_i est pair, $\varepsilon_i = 1$ si m_i est impair, et

$$q(x) = \prod_{i=1}^{k}(x - x_i)^{\varepsilon_i}, \quad \deg q \leq k \leq n.$$

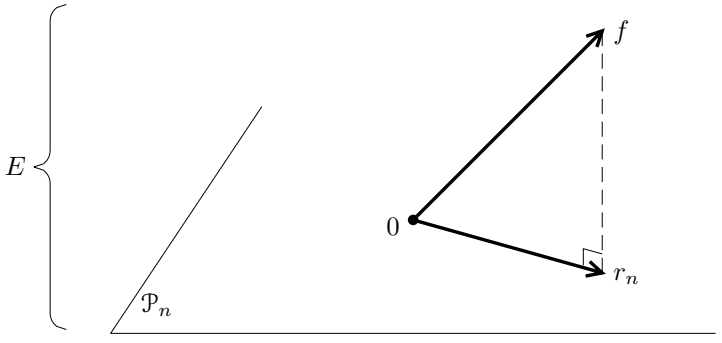
Le polynôme $p_n q$ admet dans $]a, b[$ les zéros x_i avec multiplicité paire $m_i + \varepsilon_i$, donc $p_n q$ est de signe constant dans $]a, b[\setminus \{x_1, \ldots, x_k\}$. Par conséquent

$$\langle p_n, q\rangle = \int_a^b p_n(x)q(x)w(x)dx \neq 0.$$

Comme p_n est orthogonal à $\mathcal{P}_{n-1}$, on a nécessairement $\deg q = n$, donc $k = n$ et $m_1 = \ldots = m_k = 1$. ∎

Plusieurs méthodes d'approximation de fonctions continues par des polynômes ont déjà été vues. En voici encore une autre.

Théorème 4 – *Soit $f \in E$. Alors il existe une unique polynôme $r_n \in \mathcal{P}_n$ tel que $\|f - r_n\|_2 = d_2(f, \mathcal{P}_n)$; r_n est appelé polynôme de meilleure approximation quadratique de f à l'ordre n.*



Puisqu'on travaille dans un espace euclidien, le point de $\mathcal{P}_n$ le plus proche de f n'est autre que la projection orthogonale de f sur $\mathcal{P}_n$. Si on écrit $r_n = \sum \alpha_k p_k$, il vient $\langle f, p_k\rangle = \langle r_n, p_k\rangle = \alpha_k \|p_k\|_2^2$, d'où la formule

$$r_n(x) = \sum_{k=0}^{n} \frac{\langle f, p_k\rangle}{\|p_k\|_2^2}\, p_k(x). \qquad \blacksquare$$

On va maintenant étudier la convergence de r_n quand n tend vers $+\infty$. L'inégalité évidente

$$\int_a^b |f(x)|^2 w(x)dx \le (\sup|f(x)|)^2 \int_a^b w(x)dx$$

entraîne

$$\|f\|_2 \le C_w \|f\|, \quad \text{où} \quad C_w = \Big(\int_a^b w(x)dx\Big)^{1/2}.$$

Ceci permet de contrôler $\|\ \|_2$ à l'aide de la norme uniforme.

Théorème 5 – *Si* $]a,b[$ *est borné, alors* $\lim_{n\to+\infty} \|f - r_n\|_2 = 0$ *pour tout* $f \in E$.

Remarque – Le théorème 5 peut être faux si $]a,b[$ est non borné.

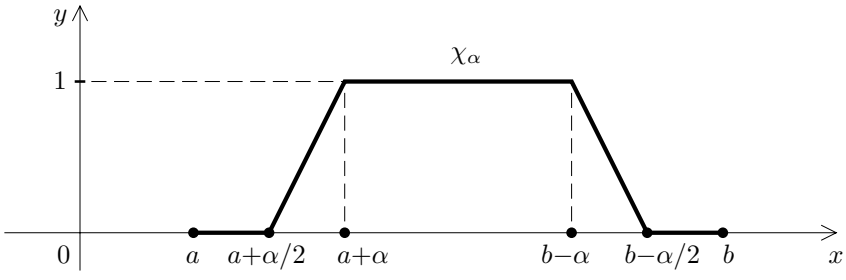
Démonstration

• Supposons d'abord que $f \in \mathcal{C}([a,b])$. Dans ce cas, soit q_n le polynôme de meilleure approximation uniforme de f. On a

$$\|f - r_n\|_2 \le \|f - q_n\|_2 \le C_w \|f - q_n\|,$$

et on sait que $\lim_{n\to+\infty} \|f - q_n\| = 0$.

• Supposons maintenant $f \in E$ quelconque. Soit χ_α la fonction plateau définie par le schéma ci-dessous :



Comme $f \in \mathcal{C}(]a,b[)$, on a $f\chi_\alpha \in \mathcal{C}([a,b])$ si l'on convient que $f\chi_\alpha(a) = f\chi_\alpha(b) = 0$. De plus

$$\|f - f\chi_\alpha\|_2^2 \leq \int_a^{a+\alpha} |f(x)|^2 w(x)dx + \int_{b-\alpha}^{b} |f(x)|^2 w(x)dx,$$

de sorte que $\lim_{\alpha \to 0^+} \|f - f\chi_\alpha\|_2 = 0$. Soit $r_{\alpha,n}$ le polynôme de meilleure approximation quadratique de $f\chi_\alpha$. On a

$$\|f - r_n\|_2 \leq \|f - r_{\alpha,n}\|_2 \leq \|f - f\chi_\alpha\|_2 + \|f\chi_\alpha - r_{\alpha,n}\|_2.$$

Soit $\varepsilon > 0$ fixé. On peut d'abord choisir $\alpha > 0$ tel que $\|f - f\chi_\alpha\|_2 < \frac{\varepsilon}{2}$; α étant ainsi fixé, on peut choisir n_0 tel que $n > n_0$ entraîne $\|f\chi_\alpha - r_{\alpha,n}\|_2 < \frac{\varepsilon}{2}$ et donc $\|f - r_n\|_2 < \varepsilon$. ■

Mise en œuvre numérique – Si les polynômes p_n sont connus, le calcul des r_n est possible dès lors qu'on sait évaluer les intégrales $\langle f, p_k \rangle$: les méthodes d'intégration numérique feront précisément l'objet du prochain chapitre. Si les polynômes p_n ne sont pas connus, on peut les calculer numériquement par la formule de récurrence du théorème 2. Le coût global de ces calculs est en général beaucoup plus élevé que celui des méthodes d'interpolation.

6. Problèmes

6.1. On note $\mathcal{C}([a,b],\mathbb{R})$ l'espace des fonctions continues sur l'intervalle $[a,b]$ à valeurs dans $\mathbb{R}$, muni de la norme $\| \ \|_\infty$ de la convergence uniforme. On considère l'application

$$\begin{aligned} \phi : \mathcal{C}([a,b],\mathbb{R}) &\longrightarrow \mathbb{R}^{n+1} \\ f &\longmapsto (m_0(f), m_1(f), \ldots, m_n(f)) \end{aligned}$$

telle que $m_i(f) = \frac{1}{2}\ (f(x_i) + f(x_i'))$, où

$$x_0 < x_0' < x_1 < x_1' < \ldots < x_n < x_n'$$

sont des points fixés de $[a,b]$.

(a) Soit $f \in \mathcal{C}([a,b],\mathbb{R})$ telle que $\phi(f) = 0$. Montrer que pour tout i il existe $\xi_i \in [x_i, x_i']$ tel que $f(\xi_i) = 0$.

(b) Montrer que la restriction $\phi : \mathcal{P}_n \to \mathbb{R}^{n+1}$ de ϕ à l'espace $\mathcal{P}_n$ des polynômes de degré $\leq n$ est injective. En déduire que pour tout $f \in \mathcal{C}([a,b]),\mathbb{R})$ il existe un unique polynôme $P_n \in \mathcal{P}$ tel que $\phi(p_n) = \phi(f)$.

(c) On suppose ici que f est de classe C^{n+1}. En utilisant (a), majorer $\|p_n - f\|_\infty$ en fonction de $f^{(n+1)}$ et $b-a$.

(d) Calculer explicitement p_2 en fonction de $m_0(f)$, $m_1(f)$, $m_2(f)$ pour la subdivision $x_0 < x_0' < x_1 < x_1' < x_2 < x_2'$ de $[a,b] = [-1,1]$ de pas constant $\frac{2}{5}$.

6.2. On note t_n le polynôme de Tchebychev de degré n et c un réel tel que $|c| < 1$.

(a) Montrer qu'il existe une fonction continue ψ à valeurs réelles, définie sur $[0,\pi]$ avec $\psi(0) = \psi(\pi) = 0$ et vérifiant

$$e^{i\psi(\theta)} = \frac{1 - ce^{-i\theta}}{1 - ce^{i\theta}}.$$

(b) Pour $n \in \mathbb{N}$ on note $g(\theta) = (n+1)\theta + \psi(\theta)$.

(α) Soit $\theta_1 = \pi$. Calculer $g(\theta_1) - n\pi$ et $g(0) - n\pi$.
En déduire qu'il existe θ_2 vérifiant $0 < \theta_2 < \theta_1$ et $g(\theta_2) = n\pi$.

(β) Montrer qu'il existe une suite strictement décroissante θ_k de $[0,\pi]$ telle que $g(\theta_k) = (n-k+2)\pi$ pour $k = 1, \dots n+2$.

(γ) On note $\varphi_n(x) = \mathrm{Re}\left(e^{i(n+1)\theta}\ \frac{1-ce^{-i\theta}}{1-ce^{i\theta}}\right)$ avec $\theta = \mathrm{Arc}\cos x$, $x \in [-1,1]$.
Calculer $\|\varphi_n\|$. Montrer que φ_n équioscille sur $n+2$ points de $[-1,1]$.

(c) (α) Montrer que la série $-\frac{1}{2} + \sum_{k=0}^{+\infty} c^k t_k(x)$ converge uniformément sur $[-1,1]$ vers une fonction $f_c(x)$ que l'on explicitera.

(β) On note $p_n(x) = -\dfrac{1}{2} + \displaystyle\sum_{k=0}^{n-1} c^k t_k(x) + \frac{c^n}{1-c^2}\ t_n(x)$.
Montrer que p_n est le polynôme de meilleure approximation uniforme de degré n de f_c. Calculer $\|f_c - p_n\|$.

(d) (α) Montrer que l'on peut choisir c et λ tels que pour tout $x \in [0,1]$ on ait $f_c(2x-1) = \frac{\lambda}{1+x}$.

(β) Montrer qu'il existe une suite de polynômes q_n de $\mathcal{P}_n$ tels que la suite

$$\alpha_n = \mathrm{Sup}_{x\in[0,1]} \left| \frac{1}{1+x} - q_n(x) \right|$$

vérifie pour tout $k \in \mathbb{N}$, $\lim\limits_{n\to+\infty} n^k \alpha_n = 0$.

6.3. Soit $f : [a,b] \to \mathbb{R}$ une fonction continue, indéfiniment dérivable sur $]a,b[$. Soient $x_0, x_1, \ldots, x_n \in [a,b]$. Pour chaque $i \in \{0,1,\ldots,n\}$, soit α_i un entier positif.

On cherche un polynôme $P(x)$ de degré $< \beta = \sum(\alpha_i + 1)$ tel que :

$$P^{(j)}(x_i) = f^{(j)}(x_i) \quad \text{pour} \quad i = 0,1,\ldots,n \quad \text{et} \quad j = 0,1,\ldots,\alpha_i,$$

où (j) désigne l'ordre de dérivation.

(a) Démontrer l'unicité de P, puis son existence grâce à un raisonnement d'algèbre linéaire.

(b) On suppose P solution du problème. Soient $R_i(x)$ et $p_i(x)$ des polynômes vérifiant les relations

$$R_i(x) = p_i(x) + (x - x_i)^{\alpha_i+1} R_{i+1}(x), \quad \deg p_i \leq \alpha_i,$$
$$R_0(x) = P(x), \quad R_{n+1}(x) = 0.$$

(α) Montrer que

$$p_i^{(j)}(x_i) = R_i^{(j)}(x_i) \quad \text{pour} \quad j = 0,1,\ldots,\alpha_i.$$

(β) Montrer que l'on peut écrire $p_i(x)$ sous la forme :

$$p_i(x) = \sum_{k=0}^{\alpha_i} a_{ik}(x - x_i)^k.$$

Calculer les coefficients a_{ik} en fonction de $R_i^{(k)}(x_i)$.

(γ) Montrer que $P(x)$ peut s'écrire :

$$P(x) = p_0(x) + \sum_{i=1}^{n} \Big(p_i(x) \prod_{r=0}^{i-1} (x - x_r)^{\alpha_r+1} \Big).$$

(δ) Indiquer une méthode de récurrence pour calculer $p_0(x)$, puis $R_1(x)$ et $a_{1k}, \ldots$, puis $R_j(x)$ et a_{jk} en fonction de f et de ses dérivées $f^{(j)}(x_i)$. Montrer que l'on peut ainsi calculer $P(x)$ en fonction des données du problème.

(ε) Que se passe-t-il dans le cas particulier où $n = 0$?

(c) On suppose que les α_i sont rangés par ordre croissant. Montrer qu'il existe $t \in [a,b]$ tel que :

$$f(x) = P(x) + (x - x_0)^{\alpha_0+1}(x - x_1)^{\alpha_1+1} \ldots (x - x_n)^{\alpha_n+1} \, \frac{f^{(\beta)}(t)}{\beta!}$$

Indication : on pourra considérer la fonction

$$g(x) = f(x) - P(x) - (x - x_0)^{\alpha_0+1} \ldots (x - x_n)^{\alpha_n+1} K,$$

et examiner combien de fois s'annulent $g(x)$, $g'(x) \ldots$, $g^{(\alpha_0)}(x), \ldots$, $g^{(\alpha_n)}(x), \ldots$, $g^{(\beta)}(x)$.

CHAPITRE III

INTÉGRATION NUMÉRIQUE

L'objet de ce chapitre est de décrire quelques méthodes numériques classiques (Newton-Cotes, Gauss, Romberg) permettant d'évaluer des intégrales de fonctions dont les valeurs sont connues en un nombre fini de points. On s'attachera à expliciter le plus complètement possible les formules d'erreurs dans chacun des cas.

1. MÉTHODES DE QUADRATURE ÉLÉMENTAIRES ET COMPOSÉES

1.1. PRINCIPE DES MÉTHODES NUMÉRIQUES

Soit $f : [\alpha, \beta] \to \mathbb{R}$ une fonction continue. On se propose de chercher des formules approchées pour l'intégrale $\int_\alpha^\beta f(x)dx$. Pour cela, on choisit d'abord une subdivision

$$\alpha = \alpha_0 < \alpha_1 < \ldots < \alpha_k = \beta$$

de l'intervalle $[\alpha, \beta]$. La formule de Chasles donne

$$\int_\alpha^\beta f(x)dx = \sum_{i=0}^{k-1} \int_{\alpha_i}^{\alpha_{i+1}} f(x)dx.$$

On est donc ramené au problème d'évaluer l'intégrale de f sur un petit intervalle $[\alpha_i, \alpha_{i+1}]$. Ce calcul est effectué au moyen de formules approchées (qui peuvent être *a priori* différentes sur chacun des intervalles $[\alpha_i, \alpha_{i+1}]$), appelées *méthodes de quadrature élémentaires*, du type suivant :

Méthodes de quadrature élémentaires

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)dx \simeq (\alpha_{i+1} - \alpha_i) \sum_{j=0}^{l_i} \omega_{i,j} f(\xi_{i,j}),$$

où $\xi_{i,j} \in [\alpha_i, \alpha_{i+1}]$, $0 \leq j \leq l_i$ et $\sum_{j=0}^{l_i} \omega_{i,j} = 1$.

La sommation peut être interprétée comme une valeur moyenne de f sur $[\alpha, \alpha_{i+1}]$. Le problème est de choisir convenablement les points $\xi_{i,j}$ et les coefficients $\omega_{i,j}$ de façon à minimiser l'erreur. Ceci se fera en général en évaluant l'intégrale $\int_{\alpha_i}^{\alpha_{i+1}} f(x)dx$ au moyen d'une interpolation de f aux points $\xi_{i,j}$.

La méthode de quadrature composée associée sera

$$\int_\alpha^\beta f(x)dx \simeq \sum_{i=0}^{k-1}(\alpha_{i+1}-\alpha_i)\sum_{j=0}^{l_i}\omega_{i,j}f(\xi_{i,j})$$

Définition – *On dit qu'une méthode de quadrature (élémentaire ou composée) est d'ordre N si la formule approchée est exacte pour tout $f \in \mathcal{P}_N$ et inexacte pour au moins un $f \in \mathcal{P}_{N+1}$.*

On observera que les formules sont toujours exactes pour $f(x) = 1$ à cause de l'hypothèse $\sum_j \omega_{i,j} = 1$. Par linéarité, elles sont donc exactes au moins pour $f \in \mathcal{P}_0$.

1.2. Exemples

(a) Cas le plus simple : $l_i = 0$, quel que soit i.

On choisit alors un seul point $\xi_i \in [\alpha_i, \alpha_{i+1}]$ et on remplace f sur $[\alpha_i, \alpha_{i+1}]$ par le polynôme de degré 0 : $p_0(x) = f(\xi_i)$. On a alors

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)dx \simeq (\alpha_{i+1}-\alpha_i)f(\xi_i),$$

$$\int_\alpha^\beta f(x)dx \simeq \sum_{i=0}^{k-1}(\alpha_{i+1}-\alpha_i)f(\xi_i),$$

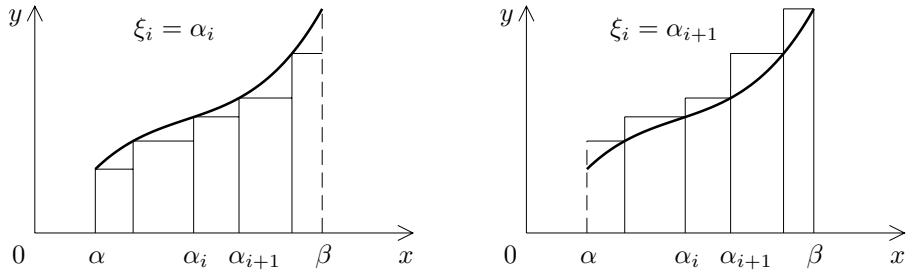
c'est-à-dire qu'on approxime l'intégrale par une somme de Riemann relative à la subdivision (α_i). Voici les choix les plus courants :

• $\xi_i = \alpha_i$: *méthode des rectangles à gauche*

$$\int_\alpha^\beta f(x)dx \simeq \sum_{i=0}^{k-1}(\alpha_{i+1}-\alpha_i)f(\alpha_i).$$

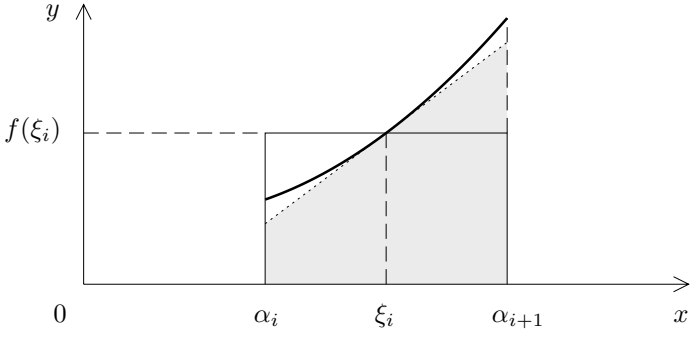
• $\xi_i = \alpha_{i+1}$: *méthode des rectangles à droite*

$$\int_\alpha^\beta f(x)dx \simeq \sum_{i=0}^{k-1}(\alpha_{i+1}-\alpha_i)f(\alpha_{i+1}).$$



Ces méthodes sont d'ordre 0.

• $\xi_i = \frac{\alpha_i + \alpha_{i+1}}{2}$: *méthode du point milieu*



L'aire du rectangle coïncide avec l'aire du trapèze indiqué en grisé. La formule approchée est donc exacte si f est une fonction affine, par suite la méthode est d'ordre 1.

(b) Cas d'une interpolation linéaire : on choisit

$$l_i = 1, \quad \forall i, \quad \xi_{i,0} = \alpha_i, \quad \xi_{i,1} = \alpha_{i+1}$$

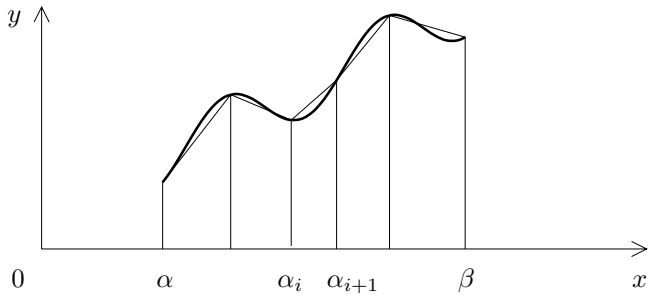
et on remplace f sur $[\alpha_i, \alpha_{i+1}]$ par la fonction linéaire p_1 qui interpole f aux points α_i, α_{i+1} :

$$p_1(x) = \frac{(x-\alpha_i)f(\alpha_{i+1}) - (x-\alpha_{i+1})f(\alpha_i)}{\alpha_{i+1} - \alpha_i}.$$

On obtient les formules suivantes, correspondant à la *méthode dite des trapèzes :*

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)dx \simeq \int_{\alpha_i}^{\alpha_{i+1}} p_1(x)dx = (\alpha_{i+1} - \alpha_i)\Big(\frac{1}{2}\, f(\alpha_i) + \frac{1}{2}\, f(\alpha_{i+1})\Big)$$

$$\int_{\alpha}^{\beta} f(x)dx \simeq \sum_{i=0}^{k-1} (\alpha_{i+1} - \alpha_i)\Big(\frac{1}{2}\, f(\alpha_i) + \frac{1}{2}\, f(\alpha_{i+1})\Big)$$



L'ordre de cette méthode est 1 comme dans le cas précédent.

(c) Méthodes de Newton-Cotes

Dans la méthode de Newton-Cotes de rang l, qu'on désignera dans la suite par NC_l, on prend $l_i = l$ pour tout i, et les points $\xi_{i,j}$, $0 \leq j \leq l$, sont les points équidistants

$$\xi_{i,j} = \alpha_i + j\ \frac{\alpha_{i+1} - \alpha_i}{l}$$

divisant $[\alpha_i, \alpha_{i+1}]$ en l sous-intervalles égaux. Pour déterminer la formule de quadrature élémentaire, on se ramène par changement de variable à l'intervalle $[\alpha_i, \alpha_{i+1}] = [-1, 1]$, subdivisé par les points $\tau_j = -1 + j\,\frac{2}{l}$. Le polynôme d'interpolation d'une fonction $f \in \mathcal{C}([-1,1])$ est donné par

$$p_l(x) = \sum_{j=0}^{l} f(\tau_j) L_j(x)$$

avec $L_j(x) = \prod_{k \neq j} \dfrac{x - \tau_k}{\tau_j - \tau_k}$. On a donc

$$\int_{-1}^{1} f(x)dx \simeq \int_{-1}^{1} p_l(x)dx = 2\sum_{j=0}^{l} \omega_j f(\tau_j)$$

avec $\omega_j = \frac{1}{2} \int_{-1}^{1} L_j(x)dx$. Par suite de la symétrie des points τ_j autour de 0, on a

$$\tau_{l-j} = -\tau_j, \quad L_{l-j}(x) = L_j(-x), \quad \omega_{l-j} = \omega_j.$$

Pour $l = 2$ par exemple, il vient

$$\tau_0 = -1, \quad \tau_1 = 0, \quad \tau_2 = 1, \quad L_1(x) = 1 - x^2, \quad \omega_1 = \frac{1}{2}\int_{-1}^{1}(1 - x^2)dx = \frac{2}{3},$$

d'où $\omega_0 = \omega_2 = \frac{1}{2}\ (1 - \omega_1) = \frac{1}{6}$. Après changement de variable, les coefficients ω_j restent inchangés (le lecteur le vérifiera à titre d'exercice), donc on obtient les

formules

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)dx \simeq (\alpha_{i+1}-\alpha_i)\sum_{j=0}^{l}\omega_j f(\xi_{i,j}),$$

$$\int_{\alpha}^{\beta} f(x)dx \simeq \sum_{i=0}^{k-1}(\alpha_{i+1}-\alpha_i)\sum_{j=0}^{l}\omega_j f(\xi_{i,j}).$$

Si $f \in \mathcal{P}_l$, alors $p_l = f$, donc la méthode de Newton-Cotes de rang l est d'ordre $\geq l$. De plus, lorsque $f \in \mathcal{C}[-1,1])$ est un polynôme impair, on a

$$\int_{-1}^{1} f(x)dx = 0 = 2\sum_{j=0}^{l}\omega_j f(\tau_j).$$

Si l est pair, les formules sont donc encore exactes pour $f(x) = x^{l+1}$, et plus généralement pour $f \in \mathcal{P}_{l+1}$ par linéarité. On démontre en fait le résultat suivant que nous admettrons :

Proposition – *Si l est pair, l'ordre de NC_l est $l+1$,*
si l est impair, l'ordre de NC_l est l.

Ceci fait que, hormis le cas $l = 1$, les méthodes de Newton-Cotes ne sont utilisées que pour l pair :

• $l = 1$: méthode des trapèzes (ordre 1)

$$\omega_0 = \omega_1 = \frac{1}{2}$$

• $l = 2$: méthode de Simpson (ordre 3)

$$\omega_0 = \omega_2 = \frac{1}{6}, \quad \omega_1 = \frac{2}{3}.$$

• $l = 4$: méthode de Boole-Villarceau (ordre 5)

$$\omega_0 = \omega_4 = \frac{7}{90}, \quad \omega_1 = \omega_3 = \frac{16}{45}, \quad \omega_2 = \frac{2}{15}$$

• $l = 6$: méthode de Weddle-Hardy (ordre 7)

$$\omega_0 = \omega_6 = \frac{41}{840}, \quad \omega_1 = \omega_5 = \frac{9}{35}, \quad \omega_2 = \omega_4 = \frac{9}{280}, \quad \omega_3 = \frac{34}{105}.$$

Pour $l \geq 8$, il apparaît des coefficients $\omega_j < 0$, ce qui a pour effet de rendre les formules beaucoup plus sensibles aux erreurs d'arrondis (cf. § 1.3). Les méthodes NC_l ne sont donc utilisées en pratique que dans les 4 cas ci-dessus.

1.3. INFLUENCE DES ERREURS D'ARRONDI

Supposons que les valeurs de f soient calculées avec des erreurs d'arrondi de valeur absolue $\leq \varepsilon$. L'erreur qui va en résulter par application d'une méthode de quadrature composée sera majorée par

$$\varepsilon \sum_{i=0}^{k-1} (\alpha_{i+1} - \alpha_i) \sum_{j=0}^{l_i} |\omega_{i,j}|.$$

Si les coefficients $\omega_{i,j}$ sont ≥ 0, on a

$$\sum_{j=0}^{l_i} |\omega_{i,j}| = \sum_{j=0}^{l_i} \omega_{i,j} = 1.$$

L'erreur est donc majorée par $\varepsilon(\beta - \alpha)$; ce résultat est manifestement optimal puisque le calcul exact de $\int_\alpha^\beta f(x)dx$ peut conduire à une erreur $\varepsilon(\beta - \alpha)$ si l'erreur sur f est constante de valeur absolue ε.

Si par contre les coefficients $\omega_{i,j}$ ne sont pas tous ≥ 0, alors $\sum_j |\omega_{i,j}| > \sum_j \omega_{i,j} = 1$, donc l'erreur due aux arrondis des $f(\xi_{i,j})$ peut dépasser $\varepsilon(\beta - \alpha)$.

1.4. CONVERGENCE QUAND LE NOMBRE k DE SUBDIVISIONS TEND VERS $+\infty$

Le résultat théorique suivant de convergence justifie en partie l'intérêt des méthodes composées.

Théorème – *On suppose que les méthodes de quadrature élémentaire font intervenir un nombre de points $l_i = l$ fixe et que les coefficients $\omega_{i,j} = \omega_j$ ne dépendent pas de i, k. Alors l'approximation donnée par la méthode composée, soit*

$$T_k(f) = \sum_{i=0}^{k-1} (\alpha_{i+1} - \alpha_i) \sum_{j=0}^{l} \omega_j f(\xi_{i,j})$$

converge vers $\displaystyle\int_\alpha^\beta f(x)dx$ *quand $k \to +\infty$ et quand le maximum du pas, à savoir $h_{\max} = \max(\alpha_{i+1} - \alpha_i)$, tend vers 0.*

Démonstration. On peut écrire $T_k(f) = \sum_{j=0}^{l} \omega_j S_{j,k}(f)$ où

$$S_{j,k}(f) = \sum_{i=0}^{k-1} (\alpha_{i+1} - \alpha_i) f(\xi_{i,j})$$

est une somme de Riemann de f relative à la subdivision (α_i). Pour tout $j = 0, 1, \ldots, l$ fixé, $S_{j,k}(f)$ converge vers $\int_\alpha^\beta f(x)dx$ quand $h_{\max}$ tend vers 0. Par conséquent $T_k(f)$ converge aussi vers $\int_\alpha^\beta f(x)dx$ quand $h_{\max}$ tend vers 0. ■

Exercice – *Montrer que dans le cas général*

$$\sum_{i=0}^{k-1}(\alpha_{i+1}-\alpha_i)\sum_{j=0}^{l_i}\omega_{i,j}f(\xi_{i,j})$$

converge encore vers $\int_\alpha^\beta f(x)dx$ *quand* $h_{\max}$ *tend vers 0, pourvu que* $\omega_{i,j}\geq 0$. [Indication : *revenir à la définition de l'intégrale en encadrant* f *par des fonctions en escalier.*]

Remarque – Dans le cas de la méthode NC_l élémentaire

$$\int_{-1}^{1} f(x)dx \simeq 2\sum_{j=0}^{l}\omega_j f(\tau_j)$$

on peut donner des exemples montrant qu'il n'y a pas nécessairement convergence quand $l\to+\infty$ (ceci est lié au phénomène de Runge II 2.3). C'est une des principales raisons pour lesquelles on est amené à considérer des méthodes composées avec k assez grand, plutôt que d'augmenter l'entier l.

2. Évaluation de l'erreur

Nous allons montrer que lorsque la fonction f à intégrer est suffisamment régulière, l'erreur d'intégration numérique peut s'exprimer de manière assez simple en fonction d'une certaine dérivée de f. Auparavant, nous aurons besoin de quelques rappels d'Analyse.

2.1. Rappels préliminaires d'analyse

Énonçons tout d'abord une version de la formule de Taylor fournissant une expression exacte du reste. Ceci est possible à l'aide d'intégrations par parties successives, permettant d'exprimer le reste comme une intégrale où figurent les dérivées de la fonction considérée.

Formule de Taylor avec reste intégral – *Soit* f *une fonction de classe* C^{N+1} *sur* $[\alpha,\beta]$. *Alors pour tout* $x\in[\alpha,\beta]$

$$f(x)=\sum_{k=0}^{N}\frac{1}{k!}f^{(k)}(\alpha)(x-\alpha)^k+\int_\alpha^x\frac{1}{N!}(x-t)^N f^{(N+1)}(t)dt.$$

Démonstration. Par récurrence sur N. Pour $N=0$, la formule se réduit simplement à

$$f(x)=f(\alpha)+\int_\alpha^x f'(t)dt.$$

Si la formule est vraie à l'ordre $N-1$, le reste intégral s'écrit, après intégration par parties :

$$\begin{aligned}\int_\alpha^x \frac{1}{(N-1)!}(x-t)^{N-1}f^{(N)}(t)dt &= \left[-\frac{1}{N!}(x-t)^N f^{(N)}(t)\right]_\alpha^x - \int_\alpha^x -\frac{1}{N!}(x-t)^N f^{(N+1)}(t)dt \\ &= \frac{1}{N!}(x-\alpha)^N f^{(N)}(\alpha) + \int_\alpha^x \frac{1}{N!}(x-t)^N f^{(N+1)}(t)dt.\end{aligned}$$

La formule est donc encore vraie à l'ordre N. ■

Notons $x_+ = \max(x,0) = x$ si $x \geq 0$, $x_+ = 0$ si $x \leq 0$. Avec la convention $x_+^0 = 1$ si $x \geq 0$, $x_+^0 = 0$ si $x < 0$, la formule se récrit

$$f(x) = p_N(x) + \int_\alpha^\beta \frac{1}{N!}(x-t)_+^N f^{(N+1)}(t)dt$$

où p_N est le polynôme de Taylor de f d'ordre N au point α.

Formule de la moyenne – *Soit $w \geq 0$ une fonction intégrable sur $]\alpha,\beta[$ telle que $\int_\alpha^\beta w(x)dx$ converge. Alors pour toute $f \in \mathcal{C}([\alpha,\beta])$, il existe un point $\xi \in]\alpha,\beta[$ tel que*

$$\int_\alpha^\beta f(x)w(x)dx = f(\xi)\int_\alpha^\beta w(x)dx.$$

Démonstration. Soient m, M respectivement le minimum et le maximum de f sur $[\alpha,\beta]$. Comme $w \geq 0$, il vient

$$m\int_\alpha^\beta w(x)dx \leq \int_\alpha^\beta f(x)w(x)dx \leq M\int_\alpha^\beta w(x)dx$$

Si $\int_\alpha^\beta w(x)dx = 0$, le résultat est vrai pour ξ quelconque. Supposons donc $\int_\alpha^\beta w(x)dx > 0$ et soit alors q le quotient

$$q = \int_\alpha^\beta f(x)w(x)dx \Big/ \int_\alpha^\beta w(x)dx \in [m,M].$$

Le théorème des valeurs intermédiaires montre que $f(]\alpha,\beta[)$ est un intervalle ayant pour bornes m, M. Si $q \in]m,M[$, il existe donc $\xi \in]\alpha,\beta[$ tel que $q = f(\xi)$. Restent les cas $q = m$ et $q = M$. Si $q = m$ et si $f(\xi) > m$ pour tout $\xi \in]\alpha,\beta[$, alors

$$\int_\alpha^\beta (f(x)-m)w(x)dx > 0$$

puisque $\int_\alpha^\beta w(x)dx > 0$, ce qui est contradictoire. Il existe donc dans ce cas $\xi \in]\alpha,\beta[$ tel que $f(\xi) = m$. Le cas $q = M$ est analogue. ■

2.2. Noyau de Peano

En vue de l'étude des méthodes de Gauss au § 3, on se place ici dans une situation un peu plus générale.

Situation étudiée – On se donne un poids w sur $]\alpha,\beta[$, c'est-à-dire une fonction continue >0 telle que $\int_\alpha^\beta w(x)dx$ converge. On cherche à évaluer l'intégrale $\int_\alpha^\beta f(x)w(x)dx$ par une formule approchée

$$\int_\alpha^\beta f(x)w(x)dx \simeq \sum_{j=0}^{l} \lambda_j f(x_j), \quad x_j \in [\alpha,\beta].$$

On notera que les formules du § 1 rentrent dans ce cadre (avec $w \equiv 1$); en général, on a $\sum \lambda_j \neq 1$. L'erreur due à la méthode est donnée par :

$$E(f) = \int_\alpha^\beta f(x)w(x)dx - \sum_{j=0}^{l} \lambda_j f(x_j).$$

Théorème et définition – *On suppose que la méthode est d'ordre $N \geq 0$. Si f est de classe C^{N+1} sur $[\alpha,\beta]$, alors*

$$E(f) = \frac{1}{N!}\int_\alpha^\beta K_N(t) f^{(N+1)}(t)dt,$$

où K_N est une fonction sur $[\alpha,\beta]$, appelée noyau de Peano associé à la méthode, définie par

$$K_N(t) = E\Big(x \mapsto (x-t)_+^N\Big), \quad t \in [\alpha,\beta].$$

Démonstration. On observe d'abord que $f \mapsto E(f)$ est une forme linéaire sur $\mathcal{C}([\alpha,\beta])$. Si $g : (x,t) \mapsto g(x,t)$ est une fonction intégrable sur $[\alpha,\beta]\times I$, le théorème de Fubini implique par ailleurs

$$E\Big(x \mapsto \int_{t\in I} g(x,t)dt\Big) = \int_{t\in I} E\Big(x \mapsto g(x,t)\Big)dt.$$

La formule de Taylor avec reste intégral donne

$$f(x) = p_N(x) + \int_\alpha^\beta \frac{1}{N!}(x-t)_+^N f^{(N+1)}(t)dt.$$

Comme $p_N \in \mathcal{P}_N$, on a $E(p_N) = 0$ par hypothèse, d'où

$$\begin{aligned} E(f) &= E\Big(x \mapsto \int_\alpha^\beta \frac{1}{N!}(x-t)_+^N f^{(N+1)}(t)dt\Big) \\ &= \int_\alpha^\beta E\Big(x \mapsto \frac{1}{N!}(x-t)_+^N f^{(N+1)}(t)\Big)dt \\ &= \int_\alpha^\beta \frac{1}{N!} f^{(N+1)}(t)\cdot E\Big(x \mapsto (x-t)_+^N\Big)dt \\ &= \frac{1}{N!}\int_\alpha^\beta K_N(t) f^{(N+1)}(t)dt. \end{aligned}$$

■

Notons que si $N \geq 1$, la fonction $(x,t) \mapsto (x-t)_+^N$ est continue sur $[\alpha,\beta] \times [\alpha,\beta]$, donc K_N est continue sur $[\alpha,\beta]$. Ceci n'est pas vrai en général si $N = 0$.

Corollaire 1 – *On a la majoration*

$$E(f) \leq \frac{1}{N!} \|f^{(N+1)}\|_\infty \cdot \int_\alpha^\beta |K_N(t)| dt.$$

Corollaire 2 – *On suppose que K_N est de signe constant. Alors pour toute $f \in C^{N+1}([\alpha,\beta])$ il existe $\xi \in]\alpha,\beta[$ tel que*

$$E(f) = \frac{1}{N!} f^{(N+1)}(\xi) \int_\alpha^\beta K_N(t) dt.$$

De plus $\int_\alpha^\beta K_N(t)dt = \frac{1}{N+1} E(x \mapsto x^{N+1})$, donc

$$E(f) = \frac{1}{(N+1)!} f^{(N+1)}(\xi)\, E(x \mapsto x^{N+1}).$$

Démonstration. La première égalité résulte du théorème et de la formule de la moyenne appliquée à la fonction $f^{(N+1)}$ et au poids $w = K_N$ (ou $w = -K_N$ si $K_N \leq 0$). La deuxième égalité s'obtient en prenant

$$f(x) = x^{N+1}, \quad \text{qui donne} \quad f^{(N+1)}(x) = (N+1)!.$$

La troisième découle des 2 premières. ■

2.3. EXEMPLES

On verra au § 2.4 comment on peut déduire le noyau de Peano d'une méthode composée de celui de la méthode élémentaire utilisée. On se contentera donc ici de regarder le cas des méthodes élémentaires sur l'intervalle de référence $[-1,1]$.

• **Méthode du point milieu**

$$E(f) = \int_{-1}^1 f(x)dx - 2f(0).$$

Cette méthode est d'ordre 1, le noyau de Peano est donné par :

$$\begin{aligned} K_1(t) &= E\Big(x \mapsto (x-t)_+\Big) \\ &= \int_{-1}^1 (x-t)_+ dx - 2(-t)_+ \\ &= \int_t^1 (x-t)dx - 2t_- \\ &= \Big[\frac{1}{2}(x-t)^2\Big]_t^1 - 2t_- = \frac{1}{2}(1-t)^2 - 2t_-. \end{aligned}$$

On a donc

$$K_1(t) = \begin{cases} \frac{1}{2}(1-t)^2 & \text{si } t \geq 0 \\ \frac{1}{2}(1-t)^2 + 2t = \frac{1}{2}(1+t)^2 & \text{si } t \leq 0, \end{cases}$$

soit $K_1(t) = \frac{1}{2}(1-|t|)^2 \geq 0$ sur $[-1,1]$. Comme $\int_{-1}^{1} K_1(t) = \int_0^1 (1-t)^2 dt = \frac{1}{3}$, le corollaire 2 implique

$$E(f) = \frac{1}{3} f''(\xi), \quad \xi \in]-1,1[.$$

• **Méthode des trapèzes** (ordre 1)

$$\begin{aligned} E(f) &= \int_{-1}^{1} f(x)dx - (f(-1)+f(1)), \\ K_1(t) &= \int_{-1}^{1} (x-t)_+ dx - ((-1-t)_+ + (1-t)_+) \\ &= \int_t^1 (x-t)dx - (1-t), \\ K_1(t) &= -\frac{1}{2}(1-t^2) \leq 0 \quad \text{sur} \quad [-1,1]. \end{aligned}$$

Comme $\int_{-1}^{1} K_1(t)dt = -\frac{2}{3}$, on en déduit

$$E(f) = -\frac{2}{3} f''(\xi), \quad \xi \in]-1,1[.$$

• **Méthode de Simpson** (ordre 3)

$$\begin{aligned} E(f) &= \int_{-1}^{1} f(x)dx - 2\Big(\frac{1}{6} f(-1) + \frac{2}{3} f(0) + \frac{1}{6} f(1)\Big). \\ K_3(t) &= E\Big(x \mapsto (x-t)_+^3\Big) \\ K_3(t) &= \int_{-1}^{1} (x-t)_+^3 dx - 2\Big(0 + \frac{2}{3}(-t)_+^3 + \frac{1}{6}(1-t)_+^3\Big) \\ &= \int_t^1 (x-t)^3 dx - 2\Big(\frac{2}{3} t_-^3 + \frac{1}{6}(1-t)^3\Big) \end{aligned}$$

Si $t \geq 0$, on obtient donc

$$\begin{aligned} K_3(t) &= \frac{1}{4}(1-t)^4 - \frac{1}{3}(1-t)^3 \\ &= \frac{1}{12}(1-t)^3[3(1-t)-4] = -\frac{1}{12}(1-t)^3(1+3t). \end{aligned}$$

Si $t \leq 0$, on a

$$K_3(t) = -\frac{1}{12}(1-t)^3(1+3t) + \frac{4}{3}t^3 = -\frac{1}{12}(1+t)^3(1-3t).$$

On aurait pu également observer que $K_3(-t) = K_3(t)$ comme il résulte de l'exercice suivant.

Exercice – *Montrer que le noyau de Peano K_N d'une méthode élémentaire*

$$\int_{-1}^{1} f(x)dx \simeq 2\sum_{j=0}^{l} \omega_j f(\xi_j)$$

est pair *dès que $\omega_{l-j} = \omega_j$ et $\xi_{l-j} = -\xi_j$ (points et coefficients répartis symétriquement autour de* 0*).*

Indication : $(x+t)_+^N - (-x-t)_+^N = (x+t)^N$.

On a donc ici

$$K_3(t) = -\frac{1}{12}(1-|t|)^3(1+3|t|) \leq 0 \quad \text{sur} \quad [-1,1],$$
$$\int_{-1}^{1} K_3(t) = 2\int_0^1 \Big(\frac{1}{4}(1-t)^4 - \frac{1}{3}(1-t)^3\Big)dt = 2\Big(\frac{1}{20} - \frac{1}{12}\Big) = -\frac{1}{15},$$
$$E(f) = -\frac{1}{15 \cdot 3!} f^{(4)}(\xi).$$

Nous admettrons le résultat général suivant.

Théorème de Steffensen – *Dans les méthodes de Newton-Cotes, le noyau de Peano est de signe constant.*

Le corollaire 2 du §2.2 est donc toujours applicable dans ce cas.

2.4. Noyau de Peano d'une méthode composée

On suppose qu'on s'est fixé une méthode de quadrature élémentaire

$$\int_{-1}^{1} g(x)dx \simeq 2\sum_{j=0}^{l} \omega_j g(\tau_j), \quad \tau_j \in [-1,1].$$

L'erreur correspondante est

$$E_{\text{elem}}(g) = \int_{-1}^{1} g(x)dx - 2\sum_{j=0}^{l} \omega_j g(\tau_j).$$

On notera k_n le noyau de Peano associé.

On considère maintenant une subdivision de $[\alpha, \beta]$:

$$\alpha = \alpha_0 < \alpha_1 < \ldots < \alpha_k = \beta$$

de pas $h_i = \alpha_{i+1} - \alpha_i$. L'erreur de la méthode composée associée à la méthode élémentaire ci-dessus est

$$E_{\text{comp}}(f) = \int_\alpha^\beta f(x)dx - \sum_{i=0}^{k-1} h_i \sum_{j=0}^{l} \omega_j f(\xi_{i,j})$$

où $\xi_{i,j}$ se déduit de τ_j par le changement de variable

$$\begin{aligned} [-1,1] &\longrightarrow [\alpha_i, \alpha_{i+1}] \\ u &\longmapsto x = \frac{\alpha_i + \alpha_{i+1}}{2} + u\,\frac{h_i}{2}. \end{aligned}$$

Définissons $g_i \in \mathcal{C}([-1,1])$ par

$$g_i(u) = f\Big(\frac{\alpha_i + \alpha_{i+1}}{2} + u\,\frac{h_i}{2}\Big).$$

Comme $dx = \frac{h_i}{2}\,du$, il vient

$$E_{\text{comp}}(f) = \sum_{i=0}^{k-1} \left(\frac{h_i}{2} \int_{-1}^{1} g_i(u)du - h_i \sum_{j=0}^{l} \omega_j g_i(\tau_j) \right) = \sum_{i=0}^{k-1} \frac{h_i}{2}\; E_{\text{elem}}(g_i).$$

Le noyau de Peano de la méthode composée est donc

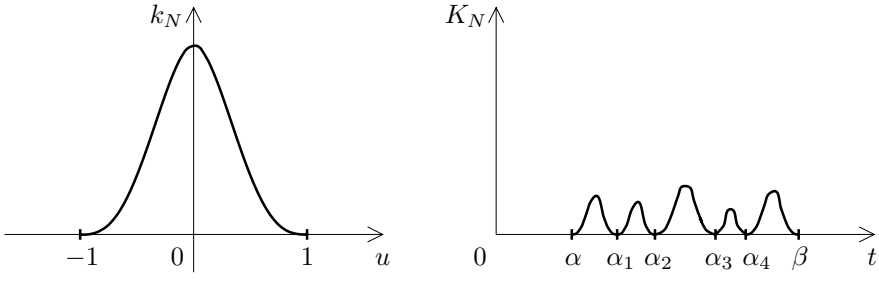
$$\begin{aligned} K_N(t) &= E_{\text{comp}}\Big(x \mapsto (x-t)_+^N\Big) \\ &= \sum_{i=0}^{k-1} \frac{h_i}{2}\; E_{\text{elem}}\left(u \mapsto \Big(\frac{\alpha_i + \alpha_{i+2}}{2} + u\,\frac{h_i}{2} - t\Big)_+^N\right) \\ &= \sum_{i=0}^{k-1} \frac{h_i}{2}\; E_{\text{elem}}\left(u \mapsto \Big(\frac{h_i}{2}\Big)^N \Big(u - \frac{2}{h_i}\Big(t - \frac{\alpha_i + \alpha_{i+i}}{2}\Big)\Big)_+^N\right) \\ K_N(t) &= \sum_{i=0}^{k-1} \Big(\frac{h_i}{2}\Big)^{N+1}\; E_{\text{elem}}\left(u \mapsto \Big(u - \frac{2}{h_i}\Big(t - \frac{\alpha_i + \alpha_{i+1}}{2}\Big)\Big)_+^N\right) \end{aligned}$$

Supposons $t \in [\alpha_j, \alpha_{j+1}]$, et soit $\theta_i = \dfrac{2}{h_i}\,\Big(t - \dfrac{\alpha_i + \alpha_{i+1}}{2}\Big)$. Alors $\theta_i \in [-1,1]$ si et seulement si $i = j$. Si $i \neq j$ on a

- ou bien $\theta_i > 1$, et $(u - \theta_i)_+^N \equiv 0$ pour $u \in [-1,1]$:
- ou bien $\theta_i < -1$, et $(u - \theta_i)_+^N \equiv (u - \theta_i)^N$ pour $u \in [-1,1]$.

Dans les 2 cas $u \mapsto (u-\theta_i)_+^N$ est un polynôme de degré $\leq N$ sur $[-1,1]$ donc $E_{\text{elem}}\big(u \mapsto (u-\theta_i)_+^N\big) = 0$. Dans la sommation, il n'y a donc que le terme $i=j$, d'où :

$$K_N(t) = \Big(\frac{h_j}{2}\Big)^{N+1} k_N\Big(\frac{2}{h_j}\,\Big(t - \frac{\alpha_j+\alpha_{j+1}}{2}\Big)\Big), \quad t \in [\alpha_j, \alpha_{j+1}].$$



Théorème – *On suppose que k_N est de signe constant et que le pas h_i est constant, égal à $h = \frac{\beta-\alpha}{k}$. On note $C_N = \int_{-1}^{1} k_N(t)dt$. Alors pour tout $f \in C^{N+1}([\alpha,\beta])$, il existe un point $\xi \in\,]\alpha,\beta[$ tel que*

$$E_{\text{comp}}(f) = \frac{C_N}{N!\,2^{N+2}}\; h^{N+1} f^{(N+1)}(\xi)(\beta-\alpha).$$

On voit donc que lorsque le pas h tend vers 0 l'ordre de grandeur de l'erreur dans une méthode composée d'ordre N est approximativement h^{N+1}. Ce résultat justifie l'intérêt des méthodes d'ordre élevé, qui donnent une précision plus grande pourvu que f soit très régulière.

Démonstration. K_N étant lui aussi de signe constant, le corollaire 2 du § 2.2 montre l'existence de $\xi \in\,]\alpha,\beta[$ tel que

$$E_{\text{comp}}(f) = \frac{1}{N!} f^{(N+1)}(\xi) \int_\alpha^\beta K_N(t)dt.$$

D'après l'expression de K_N, on obtient

$$\begin{aligned}\int_\alpha^\beta K_N(t)dt &= k\int_{\alpha_0}^{\alpha_1} K_N(t)dt \\ &= k\Big(\frac{h}{2}\Big)^{N+1} \int_{\alpha_0}^{\alpha_1} k_n\Big(\frac{2}{h}\Big(t - \frac{\alpha_0+\alpha_1}{2}\Big)\Big)\,dt\end{aligned}$$

Le changement de variable $t = \frac{\alpha_0+\alpha_1}{2} + \frac{h}{2}\,u$, $dt = \frac{h}{2}\,du$ fournit

$$\begin{aligned}\int_\alpha^\beta K_N(t)dt &= k\,\Big(\frac{h}{2}\Big)^{N+2} \int_{-1}^{1} k_n(u)du \\ &= kh\,\frac{h^{N+1}}{2^{N+2}}\,C_N = \frac{C_N}{2^{N+2}}\; h^{N+1}\,(\beta-\alpha),\end{aligned}$$

d'où le théorème. ■

Les exemples du § 2.3 donnent en particulier :

- *Point milieu :* $N = 1$, $C_1 = \dfrac{1}{3}$, $E_{\text{comp}}(f) = \dfrac{1}{24}\, h^2 f''(\xi)(\beta - \alpha)$,
- *Trapèzes :* $N = 1$, $C_1 = -\dfrac{2}{3}$, $E_{\text{comp}}(f) = -\dfrac{1}{12}\, h^2 f''(\xi)(\beta - \alpha)$,
- *Simpson :* $N = 3$, $C_3 = -\dfrac{1}{15}$, $E_{\text{comp}}(f) = -\dfrac{1}{2880}\, h^4 f^{(4)}(\xi)(\beta - \alpha)$.

3. Méthodes de Gauss

Les méthodes de Gauss concernent le calcul numérique d'intégrales faisant intervenir un poids. Elles constituent une application directe de la théorie des polynômes orthogonaux.

3.1. Description et formule d'erreur

Soit w une fonction poids fixée sur $]\alpha, \beta[$. On étudie les méthodes d'intégration approchée du type

$$\int_\alpha^\beta f(x)w(x)dx \simeq \sum_{j=0}^{l} \lambda_j f(x_j), \quad x_j \in [\alpha, \beta].$$

Théorème 1 – *Il existe un choix et un seul des points x_j et des coefficients λ_j de sorte que la méthode soit d'ordre $N = 2l + 1$. Les points x_j appartiennent à $]\alpha, \beta[$ et sont les racines du $(l+1)$-ième polynôme orthogonal pour le poids w.*

Unicité. Supposons qu'on ait des points x_j et des coefficients λ_j pour lesquels la méthode est d'ordre $\geq 2l + 1$. Posons

$$\pi_{l+1}(x) = \prod_{j=0}^{l} (x - x_j).$$

Pour tout $p \in \mathcal{P}_l$, $\deg(p\pi_{l+1}) \leq 2l + 1$, donc

$$\int_\alpha^\beta p(x)\pi_{l+1}(x)w(x)dx = \sum_{j=0}^{l} \lambda_j p(x_j)\pi_{l+1}(x_j) = 0.$$

Ceci entraîne que π_{l+1} est orthogonal à $\mathcal{P}_l$. Comme π_{l+1} est unitaire, c'est donc le $(l+1)$-ième polynôme orthogonal associé au poids w. Les points x_j ne sont autres que les racines de ce polynôme.

Soit $L_i \in \mathcal{P}_l$ tel que $\begin{cases} L_i(x_j) = 1 & \text{si} \quad i = j, \\ L_i(x_j) = 0 & \text{si} \quad i \neq j. \end{cases}$

Les coefficients λ_i sont donnés nécessairement par

$$\lambda_i = \sum_{j=0}^{l} \lambda_j L_i(x_j) = \int_\alpha^\beta L_i(x) w(x) dx.$$

Ces coefficients sont donc eux aussi uniques. ■

Existence. On sait que le polynôme orthogonal $\pi_{l+1} \in \mathcal{P}_{l+1}$ possède $l+1$ racines distinctes dans $]\alpha, \beta[$. Soient $x_0, \ldots, x_l$ ces racines et soit

$$\lambda_j = \int_\alpha^\beta L_j(x) w(x) dx.$$

Si $f \in \mathcal{C}([\alpha, \beta])$, le polynôme d'interpolation de Lagrange est

$$p_l(x) = \sum_{j=0}^{l} f(x_j) L_j(x);$$

par définition des coefficients λ_j il vient donc

$$\int_\alpha^\beta p_l(x) w(x) = \sum_{j=0}^{l} \lambda_j f(x_j).$$

Si $f \in \mathcal{P}_l$ alors $p_l = f$, donc la méthode est d'ordre $\geq l$. Montrons que l'ordre est en fait $\geq 2l+1$. En effet, lorsque $f \in \mathcal{P}_{2l+1}$, la division euclidienne de f par π_{l+1} donne

$$f(x) = q(x)\pi_{l+1}(x) + r(x), \quad \text{avec} \quad \deg q \leq l, \quad \deg r \leq l.$$

Comme $\pi_{l+1} \perp \mathcal{P}_l$, il vient $\int_\alpha^\beta q(x)\pi_{l+1}(x)w(x)dx = 0$, d'où

$$\int_\alpha^\beta f(x) w(x) dx = \int_\alpha^\beta r(x) w(x) dx = \sum_{j=0}^{l} \lambda_j r(x_j)$$

Comme $f(x_j) = r(x_j)$, on a donc bien $E(f) = 0$. Il reste seulement à voir que l'ordre n'est pas $> 2l+1$, ce qui résulte du théorème ci-dessous.

Théorème 2 – *Le noyau de Peano K_{2l+1} est ≥ 0, et pour tout $f \in C^{2l+2}([\alpha, \beta])$, il existe $\xi \in]\alpha, \beta[$ tel que*

$$E(f) = \frac{f^{(2l+2)}(\xi)}{(2l+2)!} \int_\alpha^\beta \pi_{l+1}(x)^2 w(x) dx.$$

On notera en particulier que ceci entraîne

$$E(x \mapsto x^{2l+2}) = \int_\alpha^\beta \pi_{l+1}(x)^2 w(x) dx > 0,$$

donc la méthode n'est pas d'ordre $2l+2$.

Démonstration.* D'après le § 2.2, on a

$$E(f) = \frac{1}{(2l+1)!}\int_\alpha^\beta K_{2l+1}(t) f^{(2l+2)}(t)dt.$$

Inversement, si $\varphi \in \mathcal{C}([\alpha,\beta])$, on obtient

$$\int_\alpha^\beta K_{2l+1}(t)\varphi(t)dt = (2l+1)!\, E(\Phi)$$

où Φ est une primitive d'ordre $2l+2$ de φ. Supposons par l'absurde qu'il existe $t_0 \in [\alpha,\beta]$ tel que $K_{2l+1}(t_0) < 0$. Notons $K_{2l+1}^- = \max(-K_{2l+1}, 0) \in \mathcal{C}([\alpha,\beta])$ la partie négative de la fonction K_{2l+1}, et soit φ un polynôme qui approche $K_{2l+1}^- + \varepsilon$ uniformément à ε près sur $[\alpha,\beta]$. On a donc en particulier

$$0 \le K_{2l+1}^- < \varphi < K_{2l+1}^- + 2\varepsilon,$$
$$\left|\int_\alpha^\beta K_{2l+1}(t)\varphi(t)dt - \int_\alpha^\beta K_{2l+1}(t)K_{2l+1}^-(t)dt\right| \le 2\varepsilon\int_\alpha^\beta |K_{2l+1}(t)|dt.$$

Comme $\int_\alpha^\beta K_{2l+1}(t)K_{2l+1}^-(t)dt = -\int_\alpha^\beta (K_{2l+1}^-(t))^2 dt < 0$, on en déduit pour ε assez petit :

$$\int_\alpha^\beta K_{2l+1}(t)\varphi(t)dt < 0.$$

Soit Φ une primitive d'ordre $2l+2$ de φ; Φ est un polynôme. Écrivons la division euclidienne de Φ par π_{l+1}^2 :

$$\Phi(x) = \pi_{l+1}^2(x)q(x) + r(x)$$

avec $\deg r \le \deg(\pi_{l+1}^2) - 1 = 2l+1$. Il vient $E(r) = 0$ d'où

$$E(\Phi) = E(\pi_{l+1}^2 q) = \int_\alpha^\beta \pi_{l+1}^2(x)q(x)w(x)dx - 0.$$

La formule de la moyenne implique qu'il existe $\theta \in]\alpha,\beta[$ tel que

$$E(\Phi) = q(\theta)\int_\alpha^\beta \pi_{l+1}^2(x)w(x)dx,$$

et par ailleurs

$$E(\Phi) = \frac{1}{(2l+1)!}\int_\alpha^\beta K_{2l+1}(t)\varphi(t)dt < 0.$$

On va obtenir une contradiction en montrant que $q(\theta) > 0$. Considérons le polynôme

$$g(x) = \Phi(x) - r(x) - \pi_{l+1}^2(x)q(\theta) = \pi_{l+1}^2(x)(q(x) - q(\theta)).$$

g admet $x_0, \ldots, x_l$ comme zéros de multiplicité 2, et θ de multiplicité ≥ 1, c'est-à-dire au moins $2l+3$ zéros. Il existe donc un point η intermédiaire entre les points x_j, θ, tel que $g^{(2l+2)}(\eta) = 0$. Par suite

$$0 = g^{(2l+2)}(\eta) = \Phi^{(2l+2)}(\eta) - (2l+2)!\, q(\theta) = \varphi(\eta) - (2l+2)!\, q(\theta)$$

et comme $\varphi(\eta) > 0$ on en déduit bien $q(\theta) > 0$, contradiction. Par suite $K_{2l+1} \geq 0$ et le corollaire 2 du § 2.2 donne

$$E(f) = \frac{1}{(2l+2)!}\, f^{(2l+2)}(\xi)\, E(x \mapsto x^{2l+2}).$$

Comme π_{l+1} est unitaire, on a $x^{2l+2} = \pi_{l+1}(x)^2 + r(x)$ où $r \in \mathcal{P}_{2l+1}$, donc $E(x \mapsto x^{2l+2}) = E(\pi_{l+1}^2) = \int_\alpha^\beta \pi_{l+1}(x)^2 w(x) dx$, ce qui démontre le théorème. ■

3.2. *Cas particuliers d'usage fréquent*

L'intérêt des méthodes de Gauss est de réaliser l'ordre N maximal pour un nombre fixé $l+1$ de points d'interpolation. Néanmoins, la complexité du calcul des polynômes orthogonaux fait que les méthodes de Gauss ne sont guère utilisées que dans les deux cas suivants.

• $w(x) = 1$ sur $[-1, 1]$: méthode de Gauss-Legendre.

Les polynômes orthogonaux successifs et les points x_j correspondant sont donnés par le tableau :

l	$\pi_{l+1}(x)$	$x_0, \ldots, x_l$	$\lambda_0, \ldots, \lambda_l$	ordre N
-1	1			
0	x	0	2	1
1	$x^2 - \frac{1}{3}$	$-\frac{1}{\sqrt{3}},\ \frac{1}{\sqrt{3}}$	$1,\ 1$	3
2	$x^3 - \frac{3}{5}x$	$-\sqrt{\frac{3}{5}},\ 0,\ \sqrt{\frac{3}{5}}$	$\frac{5}{9},\ \frac{8}{9},\ \frac{5}{9}$	5
3	$x^4 - \frac{6}{7}x^2 + \frac{3}{35}$	$\pm\sqrt{\frac{3}{7} \pm \frac{2}{7}\sqrt{\frac{6}{5}}}$	$\frac{1}{2} - \frac{1}{6}\sqrt{\frac{5}{6}},\ \frac{1}{2} + \frac{1}{6}\sqrt{\frac{5}{6}}$	7
4	$x^5 - \frac{10}{9}x^3 + \frac{5}{21}x$	$0,\ \pm\sqrt{\frac{5}{9} \pm \frac{2}{9}\sqrt{\frac{10}{7}}}$	compliqués !	9

• $w(x) = \dfrac{1}{\sqrt{1-x^2}}$ sur $]-1, 1[$: méthode de Gauss-Tchebychev.

Les points x_j sont alors les points d'interpolation de Tchebychev dans l'intervalle $]-1, 1[$:

$$x_j = \cos\frac{2j+1}{2l+2}\,\pi, \quad 0 \leq j \leq l,$$

et on peut démontrer (voir par exemple le livre de Crouzeix-Mignot, exercice 2.4) que $\lambda_j = \frac{\pi}{l+1}$. On obtient donc une méthode approchée d'ordre $2l+1$ s'écrivant :

$$\int_{-1}^{1} f(x)\frac{dx}{\sqrt{1-x^2}} \simeq \frac{\pi}{l+1}\sum_{j=0}^{l} f\Big(\cos\,\frac{2j+1}{2l+2}\,\pi\Big).$$

4. Formule d'Euler-Maclaurin et développements asymptotiques

Nous allons quitter ici quelque peu le fil directeur des paragraphes précédents. Notre objectif est d'obtenir une formule théorique pour le calcul du développement limité des approximations numériques en fonction du pas de la subdivision. Ceci conduit, pour des fonctions suffisamment régulières, à des procédés numériques en général très performants.

4.1. Polynômes et nombres de Bernoulli

Soit f une fonction de classe C^p sur $[0,1]$ avec $p \geq 1$. Une intégration par parties donne

$$\int_0^1 f(x)dx = \Big[\Big(x-\frac{1}{2}\Big)f(x)\Big]_0^1 - \int_0^1 \Big(x-\frac{1}{2}\Big)f'(x)dx,$$

ce qui peut se récrire

$$\frac{1}{2}\,f(0)+\frac{1}{2}\,f(1) = \int_0^1 f(x)dx + \int_0^1 B_1(x)f'(x)dx$$

avec $B_1(x) = x - \frac{1}{2}$, ce choix ayant l'intérêt que $\int_0^1 B_1(x)dx = 0$. L'idée consiste à répéter les intégrations par parties en introduisant des primitives successives de B_1 dont l'intégrale sur $[0,1]$ est nulle. De façon précise, on choisit B_p en sorte que

$$B_p'(x) = pB_{p-1}(x), \quad \int_0^1 B_p(x)dx = 0,$$

la deuxième condition permettant de fixer la constante d'intégration de manière unique. On trouve ainsi

$$\int_0^1 B_{p-1}(x)f^{(p-1)}(x)dx = \Big[\frac{1}{p}\,B_p(x)f^{(p-1)}(x)\Big]_0^1 - \int_0^1 \frac{1}{p}\,B_p(x)f^{(p)}(x)dx,$$

$$\int_0^1 \frac{B_{p-1}(x)}{(p-1)!}\;f^{(p-1)}(x)dx = \frac{b_p}{p!}\left(f^{(p-1)}(1)-f^{(p-1)}(0)\right) - \int_0^1 \frac{B_p(x)}{p!}\;f^{(p)}(x)dx,$$

où $b_p = B_p(0) = B_p(1)$ par définition (noter que $B_p(1) - B_p(0) = \int_0^1 pB_{p-1}(x)dx$ est nulle pour $p \geq 2$). De ceci on déduit facilement par récurrence la formule

$$\frac{1}{2}\,f(0)+\frac{1}{2}\,f(1) = \int_0^1 f(x)dx + \sum_{m=2}^{b}(-1)^m\frac{b_m}{m!}\left(f^{(m-1)}(1)-f^{(m-1)}(0)\right)$$
$$+(-1)^{p+1}\int_0^1 \frac{B_p(x)}{p!}\,f^{(p)}(x)dx. \qquad (*)$$

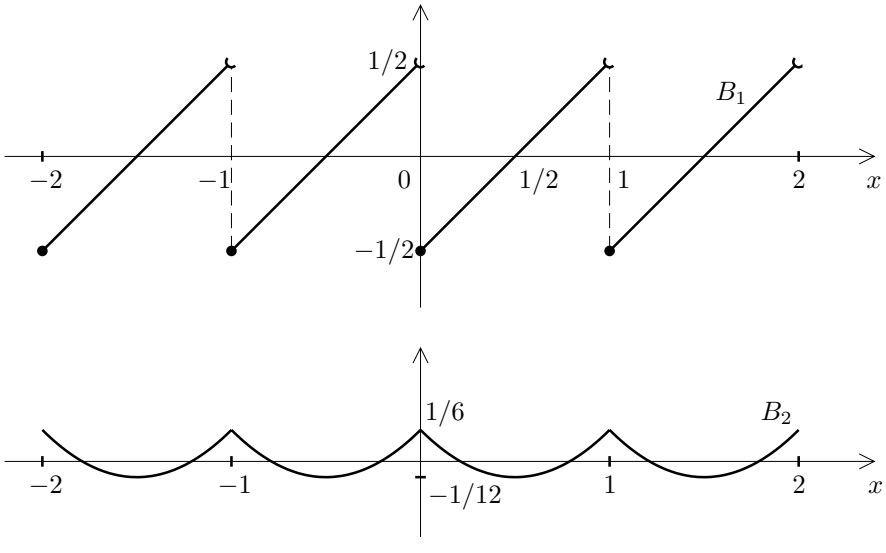
Calculons par exemple B_2. On a par définition

$$B_2'(x) = 2B_1(x) = 2x - 1, \quad \text{d'où}$$
$$B_2(x) = x^2 - x + C, \quad x \in [0,1],$$

et la condition $\int_0^1 B_2(x)dx = 0$ implique $C = \frac{1}{6}$. On voit facilement par récurrence que B_p est un polynôme unitaire de degré p à coefficients rationnels, tel que $B_p(0) = B_p(1)$ pour $p \geq 2$. On convient d'étendre B_p à $\mathbb{R}$ en posant

$$B_p(x) = B_p(x - E(x)) \quad \text{si} \quad x \notin [0,1[.$$

On obtient ainsi une fonction périodique de période 1 qui est un polynôme en restriction à $[0,1[$ (mais qui, bien entendu, n'*est pas* un polynôme sur $\mathbb{R}$ tout entier).



Théorème et définition – *Les polynômes B_p sont appelés polynômes de Bernoulli. Les nombres de Bernoulli sont les réels b_p définis par*

$$b_0 = 1, \quad b_1 = -\frac{1}{2}, \quad b_p = B_p(0) \quad \text{si} \quad p \geq 2.$$

On a les formules :

(1) $B_p(x) = \sum_{m=0}^{p} C_p^m b_m x^{p-m}$, $p \geq 1$, $x \in [0,1[$.

(2) $b_p = \sum_{m=0}^{p} C_p^m b_m$, $p \geq 2$.

(3) $B_p(1-x) = (-1)^p B_p(x)$, $p \geq 1$.

(4) $b_m = 0$ *si m est impair* ≥ 3.

Démonstration

(1) La formule est vraie pour $p = 1$ d'après la définition de b_0, b_1. Supposons la formule vraie à l'ordre $p-1$:

$$B_{p-1}(x) = \sum_{m=0}^{p-1} C_{p-1}^m b_m x^{p-1-m}.$$

On a alors

$$B'_p(x) = pB_{p-1}(x) = \sum_{m=0}^{p-1} pC_{p-1}^m b_m x^{p-1-m},$$

$$\begin{aligned} B_p(x) &= B_p(0) + \sum_{m=0}^{p-1} \frac{p}{p-m} C_{p-1}^m b_m x^{p-m} \\ &= b_p + \sum_{m=0}^{p-1} C_p^m b_m x^{p-m}, \end{aligned}$$

donc la formule est encore vraie à l'ordre p.

(2) D'après ce qui précède, B_p est continue sur $\mathbb{R}$ pour tout $p \geq 2$ et vérifie $B_p(1) = B_p(0) = b_p$, par conséquent (2) est un cas particulier de (1).

(3) Par récurrence sur p, on voit que $(-1)^p B_p(1-x)$ a pour dérivée

$$-(-1)^p B'_p(1-x) = p(-1)^{p-1} B_{p-1}(1-x) = pB_{p-1}(x) = B'_p(x).$$

Comme $(-1)^p B_p(1-x)$ est d'intégrale nulle sur $[0,1]$, on en déduit que $(-1)^p B_p(1-x)$ et $B_p(x)$ coïncident.

(4) Pour $p \geq 2$ et $x = 0$, (3) donne $b_p = (-1)^p b_p$, donc $b_p = 0$ si p est impair. ■

La relation (2) appliquée à $p = 2k+1$ donne

$$0 = C_{2k+1}^{2k} b_{2k} + C_{2k+1}^{2k-2} b_{2k-2} + \ldots + C_{2k+1}^2 b_2 + C_{2k+1}^1 b_1 + 1.$$

Ceci permet de calculer par récurrence les nombres de Bernoulli successifs :

$$b_0 = 1,\ b_1 = -\frac{1}{2},\ b_2 = \frac{1}{6},\ b_4 = -\frac{1}{30},\ b_6 = \frac{1}{42},\ b_8 = -\frac{1}{30},\ b_{10} = \frac{5}{66}, \ldots$$

Supposons maintenant donnée une fonction f de classe C^p sur $[\alpha, \beta]$ où α, β sont des entiers. Grâce à la périodicité des fonctions B_p, la formule $(*)$ ci-dessus est vraie sur chaque intervalle $[\alpha, \alpha+1], \ldots, [\beta-1, \beta]$. Par sommation, on en déduit

$$\begin{aligned} \frac{1}{2} f(\alpha) + f(\alpha+1) + \ldots + f(\beta-1) + \frac{1}{2} f(\beta) &= \int_\alpha^\beta f(x)dx \\ &+ \sum_{m=2}^{p} (-1)^m \frac{b_m}{m!} \left(f^{(m-1)}(\beta) - f^{(m-1)}(\alpha) \right) + (-1)^{p+1} \int_\alpha^\beta \frac{B_p(x)}{p!} f^{(p)}(x)dx. \end{aligned}$$

En appliquant ceci pour $p = 2k$ et en tenant compte du fait que $b_m = 0$ si m est impair ≥ 3, on obtient la

Formule d'Euler-Maclaurin – Soit f une fonction de classe C^k sur $[\alpha, \beta]$ où $\alpha, \beta \in \mathbb{Z}$ et soit $T(f) = \frac{1}{2} f(\alpha) + f(\alpha+1) + \ldots + f(\beta-1) + \frac{1}{2} f(\beta)$ la somme des trapèzes associées à f. Alors

$$T(f) = \int_\alpha^\beta f(x)dx + \sum_{m=1}^{k} \frac{b_{2m}}{(2m)!} \left(f^{(2m-1)}(\beta) - f^{(2m-1)}(\alpha) \right) - \int_\alpha^\beta \frac{B_{2k}(x)}{(2k)!}\ f^{(2k)}(x)dx.$$

Pour pouvoir exploiter cette formule à des fins numériques, il importe de savoir majorer la fonction $B_{2k}(x)$ qui intervient dans le reste intégral.

4.2. Lien avec les séries de Fourier et estimation de B_p

Comme B_p est périodique de période 1, il est tentant de rechercher un développement de B_p en série de Fourier. D'après la formule $(*)$ du § 4.1 appliquée à $f(x) = e^{-2\pi inx}$, il vient

$$1 = \int_0^1 e^{-2\pi inx}dx + 0 - (2\pi in)^p \int_0^1 \frac{B_p(x)}{p!}\ e^{-2\pi inx}dx,$$

et la première intégrale est nulle pour $n \neq 0$. On en déduit que le coefficient de Fourier d'indice n de B_p est

$$\begin{cases} \widehat{B}_p(n) = -\dfrac{p!}{(2\pi in)^p} & \text{si } n \neq 0, \\ \widehat{B}_p(0) = \displaystyle\int_0^1 B_p(x) = 0 & \text{si } n = 0. \end{cases}$$

Pour $p \geq 2$, la série de Fourier est absolument convergente et B_p est continue, donc

$$B_p(x) = -p! \sum_{n \in \mathbb{Z}^*} \frac{e^{2\pi inx}}{(2\pi in)^p}, \quad (\forall x \in \mathbb{R}).$$

Pour $p = 1$, la fonction B_1 est de classe C^1 par morceaux, donc la série converge vers $B_1(x)$ en tout point $x \notin \mathbb{Z}$ et vers $\frac{1}{2}\left(B_1(x+0) + B_1(x-0)\right) = 0$ si $x \in \mathbb{Z}$. La formule ci-dessus peut se récrire

$$B_{2k}(x) = \frac{(-1)^{k+1} 2(2k)!}{(2\pi)^{2k}} \sum_{n=1}^{+\infty} \frac{\cos 2\pi nx}{n^{2k}},$$

$$B_{2k+1}(x) = \frac{(-1)^{k+1} 2(2k+1)!}{(2\pi)^{2k+1}} \sum_{n=1}^{+\infty} \frac{\sin 2\pi nx}{n^{2k+1}}.$$

En particulier, si l'on introduit la fonction ζ de Riemann

$$\zeta(s) = \sum_{n=1}^{+\infty} \frac{1}{n^s},$$

on obtient

$$b_{2k} = \frac{(-1)^{k+1} 2(2k)!}{(2\pi)^{2k}}\ \zeta(2k).$$

Comme $\zeta(s) \leq 1 + \int_1^{+\infty} dx/x^s = 1 + 1/(s-1)$, on voit que $\lim_{s\to+\infty} \zeta(s) = 1$ et en particulier on a

$$b_{2k} \sim \frac{(-1)^{k+1} 2(2k)!}{(2\pi)^{2k}} \quad \text{quand} \quad k \to +\infty.$$

Les coefficients $|b_{2k}|$ tendent donc vers $+\infty$ assez vite. Par ailleurs, il est clair que $B_{2k}(x)$ atteint sa valeur absolue maximum pour $x = 0$; on obtient donc

$$|B_{2k}(x)| \leq |b_{2k}|, \quad \forall x \in \mathbb{R}. \tag{$**$}$$

Comme $B_{2k+1}(0) = 0$ pour $k \geq 1$, on a d'autre part

$$B_{2k+1}(x) = (2k+1) \int_0^x B_{2k}(t)dt$$

donc $|B_{2k+1}(x)| \leq (2k+1)|x|\,|b_{2k}|$. On en déduit l'inégalité

$$|B_{2k+1}(x)| \leq \Big(k + \frac{1}{2}\Big)|b_{2k}|$$

d'abord pour $x \in \big[0, \frac{1}{2}\big]$, puis pour $x \in \big[\frac{1}{2}, 1\big]$ grâce à la formule (3) du § 4.1, puis pour tout $x \in \mathbb{R}$ par périodicité.

4.3. Application à la recherche de développements asymptotiques

Soit f une fonction de classe C^∞ sur $[\alpha, +\infty[$ où $\alpha \in \mathbb{Z}$. Pour tout entier $n \geq \alpha$, on cherche à obtenir un développement limité à tout ordre de la somme

$$S_n(f) = f(\alpha) + f(\alpha+1) + \ldots + f(n)$$

lorsque n tend vers $+\infty$. Un tel développement est appelé *développement asymptotique* de $S_n(f)$; il permet généralement d'obtenir de très bonnes valeurs approchées de $S_n(f)$ lorsque n est grand.

Théorème – *On suppose qu'il existe un entier $m_0 \in \mathbb{N}$ et un réel x_0 tels que pour $m \geq m_0$ les dérivées $f^{(m)}(x)$ soient de signe constant sur $[x_0, +\infty[$, avec $\lim_{x\to+\infty} f^{(m)}(x) = 0$. Alors il existe une constante C indépendante de n et k, telle que pour tout $n \geq x_0$ et tout $k > \frac{m_0}{2}$ on ait :*

$$S_n(f) = C + \frac{1}{2}\, f(n) + \int_\alpha^n f(x)dx + \sum_{m=1}^{k-1} \frac{b_{2m}}{(2m)!}\ f^{(2m-1)}(n) + R_{n,k}$$

avec

$$R_{n,k} = \theta\ \frac{b_{2k}}{(2k)!}\ f^{(2k-1)}(n) = \theta \times (1^{\text{er}}\ \textit{terme omis}), \qquad \theta \in [0,1].$$

Démonstration. On a par définition $S_n(f) = \frac{1}{2}\,f(\alpha) + \frac{1}{2}\,f(n) + T(f)$ où $T(f)$ est la somme des trapèzes de f sur $[\alpha, n]$. La formule d'Euler Maclaurin entraîne

$$\begin{aligned} S_n(f) = \frac{1}{2}\,f(\alpha) + \frac{1}{2}\,f(n) + \int_\alpha^n f(x)dx + \sum_{m=1}^{k} \frac{b_{2m}}{(2m)!}\; f^{(2m-1)}(n) \\ - \sum_{m=1}^{k} \frac{b_{2m}}{(2m)!}\; f^{(2m-1)}(\alpha) - \int_\alpha^{+\infty} \frac{B_{2k}(x)}{(2k)!}\; f^{(2k)}(x)dx \\ + \int_n^{+\infty} \frac{B_{2k}(x)}{(2k)!}\; f^{(2k)}(x)dx. \end{aligned}$$

On obtient donc le développement du théorème avec une constante $C = C_k$ dépendant *a priori* de k et un reste $R_{n,k}$ donnés par

$$\begin{aligned} C_k &= \frac{1}{2}\,f(\alpha) - \sum_{m=1}^{k} \frac{b_{2m}}{(2m)!}\, f^{(2m-1)}(\alpha) - \int_\alpha^{+\infty} \frac{B_{2k}(x)}{(2k)!}\; f^{(2k)}(x)dx, \\ R_{n,k} &= \frac{b_{2k}}{(2k)!}\; f^{(2k-1)}(n) + \int_n^{+\infty} \frac{B_{2k}(x)}{(2k)!}\, f^{(2k)}(x)dx, \end{aligned}$$

à condition de montrer que les intégrales convergent. Comme $k > \frac{m_0}{2}$, $f^{(2k)}$ est de signe constant sur $[x_0, +\infty[$. D'après l'inégalité $(**)$ du § 4.2, il vient

$$\Big|\int_n^{+\infty} \frac{B_{2k}(x)}{(2k)!}\, f^{(2k)}(x)dx\Big| \le \frac{|b_{2k}|}{(2k)!}\,\Big|\int_n^{+\infty} f^{(2k)}(x)\Big|,$$

$$\int_n^{+\infty} f^{(2k)}(x)dx = \lim_{N\to+\infty}\int_n^N = \lim_{N\to+\infty}\Big(f^{(2k-1)}(N) - f^{(2k-1)}(n)\Big) = -f^{(2k-1)}(n).$$

On a donc bien convergence et nos estimations montrent par ailleurs que l'intégrale figurant dans $R_{n,k}$ est de valeur absolue plus petite que le premier terme, donc

$$R_{n,k} = \theta\,\frac{b_{2k}}{(2k)!}\; f^{(2k-1)}(n), \quad \theta \in [0,2].$$

Il reste à voir qu'on a en fait $\theta \in [0,1]$ et que C_k ne dépend pas de k. Appliquons la formule à l'ordre $k+1$ et identifions avec la formule donnant $S_n(f)$ à l'ordre k. Il vient

$$C_k + R_{n,k} = C_{k+1} + \frac{b_{2k}}{(2k)!}\; f^{(2k+1)}(n) + R_{n,k+1}.$$

En faisant tendre n vers $+\infty$, on trouve $C_k = C_{k+1}$, donc C_k est bien indépendante de k, et

$$R_{n,k} = \frac{b_{2k}}{(2k)!}\, f^{(2k+1)}(n) + R_{n,k+1}.$$

D'après ce qui précède, $R_{n,k}$ est de même signe que le terme $b_{2k}/(2k)!f^{(2k-1)}(n)$ tandis que $R_{n,k+1}$ est du signe opposé : le § 4.2 montre que signe $(b_{2k}) = (-1)^{k+1}$, tandis que signe $f^{(2k+1)} = -$signe $f^{(2k)} =$ signe $f^{(2k-1)}$. On a donc

$$R_{n,k}\Big/\Big(\frac{b_{2k}}{(2k)!}\, f^{(2k-1)}(n)\Big) \le 1,$$

ce qui implique $\theta \in [0,1]$. ■

Exemple – *Formule de Stirling avec reste.*

On applique la formule à $f(x) = \ln x$ sur $[1, +\infty[$:

$$S_n(f) = \ln 1 + \ldots + \ln(n) = \ln(n!),$$
$$\int_1^n \ln x dx = n(\ln(n) - 1) + 1,$$
$$f^{(m)}(x) = \frac{(-1)^{m-1}(m-1)!}{x^m}, \quad \text{d'où}$$
$$\ln(n!) = C' + \frac{1}{2}\ln(n) + n(\ln(n) - 1) + \sum_{m=1}^{k-1} \frac{b_{2m}}{2m(2m-1)} \frac{1}{n^{2m-1}} + R_{n,k}$$
$$n! = e^{C'}\sqrt{n}\Big(\frac{n}{e}\Big)^n \exp\Big(\sum_{m=1}^{k-1} \frac{b_{2m}}{2m(2m-1)} \frac{1}{n^{2m-1}} + R_{n,k}\Big)$$

On peut vérifier que $e^{C'} = \sqrt{2\pi}$ (exercice ci-dessous), d'où en particulier

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(\frac{1}{12n} - \frac{1}{360n^3} + \frac{1}{1260n^5} - \frac{\theta}{1680n^7}\right).$$

Exercice – *On pose* $I_n = \displaystyle\int_0^{\pi/2} \sin^n x\, dx$, $n \in \mathbb{N}$.

(a) *Montrer que* $I_n = \frac{n-1}{n} I_{n-2}$ *si* $n \geq 2$.
Calculer I_0, I_1 *puis* I_{2n}, I_{2n+1} *et* $I_{2n} \cdot I_{2n+1}$.

(b) *Montrer que* I_n *est décroissante et que* $I_{2n+1} \sim I_{2n}$.

(c) *En déduire* $\dfrac{(2n)!}{n!^2} \sim \dfrac{2^{2n}}{\sqrt{\pi n}}$ *et la valeur de* $e^{C'}$.

5. Méthode d'intégration de Romberg

On va montrer ici comment à partir de la formule d'Euler-Maclaurin on peut construire une méthode d'intégration basée sur l'accélération de la convergence de la méthode des trapèzes. On obtient ainsi un algorithme de calcul souple et performant, aisé à programmer et souvent préféré à tout autre dans la pratique.

5.1. Procédé d'extrapolation de Richardson

On suppose donnée une fonction A qui admet un développement limité à tout ordre au voisinage de 0 :

$$A(t) = a_0 + a_1 t + \ldots + a_k t^k + R_{k+1}(t)$$

avec $|R_{k+1}| \leq C_{k+1}|t|^{k+1}$.

La situation est la suivante : on suppose qu'on a un algorithme permettant de calculer $A(t_m)$ pour certains réels $t_m \to 0_+$, et on cherche à *extrapoler* ces valeurs pour obtenir $A(0) = a_0$. On construit pour cela un procédé d'*accélération de la convergence* consistant à éliminer successivement les termes $a_1 t, a_2 t^2, \ldots$ du développement limité de $A(t)$.

Principe de la méthode – Soit $r > 1$ un réel fixé. On a

$$A(rt) = a_0 + \ldots + a_n r^n t^n + \ldots + a_k r^k t^k + O(t^{k+1}).$$

Pour éliminer le terme en t^n, il suffit de former le quotient

$$\frac{r^n A(t) - A(rt)}{r^n - 1} = a_0 + b_1 t + \ldots + b_{n-1} t^{n-1} + 0 + b_{n+1} t^{n-1} + \ldots$$

Si on calcule successivement les quantités

$$\begin{aligned} A_0(t) &= A(t) \\ A_1(t) &= \frac{r A_0(t) - A_0(rt)}{r-1}, \ldots, \\ A_n(t) &= \frac{r^n A_{n-1}(t) - A_{n-1}(rt)}{r^n - 1}, \end{aligned}$$

alors on élimine successivement $t, t^2, \ldots, t^n$. De manière générale on aura

$$A_n(t) = a_0 + b_{n,n+1} t^{n+1} + \ldots + b_{n,k} t^k + O(t^{k+1})$$

donc $A_n(t) = A_0 + O(t^{n+1})$ est une meilleure approximation de a_0 que la fonction $A(t)$ initiale. Supposons en particulier qu'on sache calculer les quantités

$$A_{m,0} = A(r^{-m} t_0)$$

où $t_0 > 0$ est fixé (de sorte que $\lim_{m \to +\infty} A_{m,0} = a_0$). On a seulement *a priori* $A(t) = a_0 + O(t)$, donc

$$A_{m,0} = a_0 + O(r^{-m}).$$

Si on pose $A_{m,n} = A_n(r^{-m} t_0)$, il vient

$$A_{m,n} = a_0 + O(r^{-m(n+1)}) \quad \text{quand} \quad m \to +\infty,$$

de sorte que la convergence est sensiblement $(n+1)$-fois rapide que celle de $A_{m,0}$. Les nombres $A_{m,n}$ se calculent par la formule de récurrence

$$A_{m,n} = \frac{r^n A_{m,n-1} - A_{m-1,n-1}}{r^n - 1}.$$

Dans la pratique, on commence par ranger les valeurs $A_{m,0}$ dans un tableau TAB, puis on effectue le calcul des colonnes $A_{m,1}$, $A_{m,2}$, ... comme suit :

$$\begin{array}{lllll}
\text{TAB}[0] & A_{0,0} \nearrow & A_{1,1} \nearrow & A_{2,2} \nearrow & A_{3,3} \\
\text{TAB}[1] & A_{1,0} \nearrow & A_{2,1} \nearrow & A_{3,2} & \dots \\
\text{TAB}[2] & A_{2,0} \nearrow & A_{3,1} & \dots & \dots \\
\text{TAB}[3] & A_{3,0} & \dots & \dots & \dots \\
\dots & \dots & \dots & \dots & \dots
\end{array}$$

Chaque colonne est une suite convergeant vers a_0, mais la colonne d'indice n converge $n+1$ fois plus vite à l'infini que celle d'indice 0.

5.2. Méthode de Romberg

Soit $f \in C^\infty([\alpha,\beta])$. On considère la subdivision de $[\alpha,\beta]$ en l sous-intervalle égaux donnée par les points $x_j = \alpha + jh$, $0 \le j \le l$ où $h = \frac{\beta-\alpha}{l}$, et on note

$$T_f(h) = h\left(\frac{1}{2}f(\alpha) + f(\alpha+h) + \ldots + f(\beta-h) + \frac{1}{2}f(\beta)\right)$$

la somme des trapèzes associées. Appliquons la formule d'Euler-Maclaurin à la fonction

$$\begin{aligned} g(u) &= f(\alpha+uh), \quad u \in [0,l], \\ g^{(m)}(u) &= h^m f^{(m)}(\alpha+uh). \end{aligned}$$

Il vient

$$\begin{aligned} T_g(1) = \int_0^l f(\alpha+uh)du + \sum_{m=1}^{k} \frac{b_{2m}}{2m!}\, h^{2m-1}\Big(f^{(2m-1)}(\beta) - f^{(2m-1)}(\alpha)\Big) \\ - h^{2k}\int_0^l \frac{B_{2k}(u)}{2k!}\, f^{(2k)}(\alpha+uh)du \end{aligned}$$

d'où

$$\begin{aligned} T_f(h) = hT_g(1) = \int_\alpha^\beta f(x)dx + \sum_{m=1}^{k} \frac{b_{2m}}{(2m)!}\, h^{2m}\Big(f^{(2m-1)}(\beta) - f^{(2m-1)}(\alpha)\Big) \\ - h^{2k}\int_\alpha^\beta \frac{B_{2k}((x-\alpha)/h)}{2k!}\, f^{(2k)}(x)dx. \end{aligned}$$

On en déduit que $T_f(h)$ admet le développement limité

$$T_f(h) = \int_\alpha^\beta f(x)dx + \sum_{m=1}^{k-1} a_m h^{2m} + O(h^{2k})$$

avec $a_m = \frac{b_{2m}}{(2m)!}\Big(f^{(2m-1)}(\beta) - f^{(2m-1)}(\alpha)\Big)$.

On peut donc écrire $T_f(h) = A(h^2)$ où

$$A(t) = T_f(\sqrt{t}) = a_0 + a_1 t + \ldots + a_{k-1}t^{k-1} + O(t^k),$$

et il s'agit de calculer le coefficient

$$a_0 = \int_\alpha^\beta f(x)dx.$$

On utilise pour cela des dichotomies successives avec les pas $h = \frac{\beta-\alpha}{2^m}$. Ceci nous amène à calculer

$$A_{m,0} = T_f\Big(\frac{\beta-\alpha}{2^m}\Big) = A(4^{-m}(\beta-\alpha)^2\Big).$$

On applique donc le procédé d'extrapolation de Richardson avec $r = 4$, ce qui conduit à la formule de récurrence

$$A_{m,n} = \frac{4^n A_{m,n-1} - A_{m-1,n-1}}{4^n - 1}.$$

On a alors $A_{m,n} = \int_\alpha^\beta f(x)dx + O(4^{-m(n+1)})$ quand $m \to +\infty$. La valeur approchée retenue est celle correspondant aux indices m, n les plus élevés pour lesquels $A_{m,n}$ a été calculé.

Remarque 1 – On peut gagner du temps dans le calcul de $A_{m,0}$ en utilisant $A_{m-1,0}$ pour évaluer $A_{m,0}$. Si $h = \frac{\beta-\alpha}{2^m}$, on a en effet :

$$A_{m,0} = h\Big(\frac{1}{2}f(\alpha) + f(\alpha+h) + \ldots + f(\beta-h) + \frac{1}{2}f(\beta)\Big)$$
$$A_{m-1,0} = 2h\Big(\frac{1}{2}f(\alpha) + f(\alpha+2h) + \ldots + f(\beta-2h) + \frac{1}{2}f(\beta)\Big)$$

Il suffit de poser

$$A'_{m,0} = h\Big(f(\alpha+h) + f(\alpha+3h) + \ldots + f(\beta-h)\Big)$$

et alors on obtient

$$A_{m,0} = \frac{1}{2}\,A_{m-1,0} + A'_{m,0}.$$

Remarque 2 – Si $f \in C^\infty(\mathbb{R})$ est périodique de période $\beta-\alpha$, alors $f^{(m)}(\beta) = f^{(m)}(\alpha)$ pour tout m et on a donc un développement limité à tout ordre

$$T_f(h) = \int_\alpha^\beta f(x)dx + O(h^{2k})$$

réduit à son terme constant. Il est inutile dans ce cas d'appliquer le procédé d'extrapolation de Richardson : la dernière somme des trapèzes calculée $A_{m,0}$ donne déjà une très bonne approximation de l'intégrale.

Exercice – *Vérifier que $A_{m,1}$ (resp. $A_{m,2}$) est la méthode de Simpson composée sur 2^{m-1} sous-intervalles (resp. Boole-Villarceau sur 2^{m-2} sous-intervalles).*

Pour $n \geq 3$, on peut vérifier que $A_{m,n}$ ne correspond plus à une méthode de Newton-Cotes.

6. Problèmes

6.1. Soient x_1 et x_2 deux points de $[-1,1]$ et λ_1 et $\lambda_2 \in \mathbb{R}$. On désigne par $C[-1,1]$ l'espace vectoriel des fonctions continues sur $[-1,1]$ et à valeurs réelles et on définit

$$T : C[-1,1] \to \mathbb{R} \quad \text{par} \quad T(f) = \lambda_1 f(x_1) + \lambda_2 f(x_2).$$

(a) Quelles conditions doivent vérifier x_1, x_2, λ_1, λ_2 pour que T soit une méthode d'intégration sur $[-1,1]$ exacte pour

(α) Les fonctions constantes ?

(β) Les fonctions affines ?

(γ) Les polynômes de degré inférieur ou égal à 2 ?

(b) Parmi les méthodes exactes pour les polynômes de degré inférieur ou égal à 2, une seule vérifie $x_1 = -x_2$. Montrer que ce choix de x_1 et x_2 (et des λ_1 et λ_2 correspondants) fournit une méthode exacte pour les polynômes de degré inférieur ou égal à 3 et qu'il s'agit de la seule méthode d'intégration exacte pour les polynômes de degré inférieur ou égal à 3 qui soit du type étudié dans le problème. Quelle est cette méthode ?

6.2.

(a) Montrer que pour un polynôme trigonométrique de degré n

$$\sum_{p=-n}^{n} c_p e^{ipx},$$

la méthode des trapèzes de pas constant $h = \frac{2\pi}{n+1}$ est exacte sur l'intervalle $[0, 2\pi]$.

(b) Montrer que si f peut être approchée par un polynôme trigonométrique de degré n à moins de ε sur $[a,b]$, la méthode des trapèzes pour $h = \frac{2\pi}{n+1}$ fournit un erreur inférieure à $4\pi\varepsilon$ pour $\int_0^{2\pi} f(x)dx$.

(c) On considère $f(x) = \exp\left(\frac{1}{2} \sin x\right)$. Donner une majoration de l'erreur pour la méthode des trapèzes pour $\int_0^{2\pi} f(x)\,dx$ avec $h = \pi/2$, $h = \pi/4$. Que pensez-vous de ce dernier résultat ?

6.3. Soit $f : [-1,1] \to \mathbb{R}$ une fonction de classe C^n, où n sera supposé aussi grand que les besoins l'exigeront. On considère la méthode d'intégration numérique approchée donnée par

$$\text{(M)} \qquad \int_{-1}^{1} f(x)dx \simeq f(\omega) + f(-\omega) \quad \text{avec} \quad \omega \in [0,1].$$

(a) Calculer l'erreur

$$E(f) = \int_{-1}^{1} f(x) - (f(\omega) + f(-\omega))$$

pour $f(x) = 1, x, x^2$ respectivement. Déterminer l'ordre de la méthode (M) en fonction de ω.

(b) On se place ici dans le cas où la méthode (M) est d'ordre 1.

(α) Calculer le noyau de Peano $K_1(t)$, et tracer le graphe de K_1 pour $\omega = 5/8$. Pour quelles valeurs de ω le noyau K_1 est-il de signe constant ?

(β) Montrer que l'erreur vérifie une majoration

$$|E(f)| \leq C(\omega)\|f''\|_\infty$$

où $C(\omega)$ est une constante dont on déterminera la valeur optimale :

- lorsque K_1 est de signe constant ;
- lorsque $\omega = 5/8$.

(c) Calculer le noyau de Peano dans le cas où la méthode (M) est d'ordre 3 et vérifier que ce noyau est une fonction paire. En déduire qu'il existe $\xi \in]-1, 1[$ tel que

$$E(f) = \frac{1}{135}\ f^{(4)}(\xi).$$

(d) En utilisant le résultat du (c), estimer l'erreur obtenue par la méthode composée associée à la méthode (M) pour le calcul d'une intégrale

$$\int_a^b g(x)dx$$

avec une subdivision de $[a, b]$ de pas constant $h = (b-a)/k$, $k \in \mathbb{N}^\star$.

6.4. Soit p un entier naturel et soit $f(x) = x^p$. On note

$$S_{n,p} = \sum_{m=1}^{n} m^p.$$

On utilise la formule du développement asymptotique de $S_n(f)$ avec $\alpha = 0$.

(a) Montrer que pour k assez grand, le reste $R_{n,k}$ est nul. En déduire une expression de $S_{n,p}$; on calculera la valeur de la constante C en observant que $S_{0,p} = 0$.

(b) Donner une expression factorisée de $S_{n,p}$ pour $p = 2, 3, 4, 5$.

6.5. Soit β un réel > 1. On considère la fonction

$$f(x) = \frac{1}{x^\beta} \quad \text{et on note} \qquad \zeta(\beta) = \sum_{n=1}^{+\infty} \frac{1}{n^\beta}.$$

On utilise la formule du développement asymptotique de $S_n(f)$ avec $\alpha = 1$.

(a) Exprimer $\zeta(\beta)$ en fonction de la constante C de la formule ; pour cela, on fera tendre n vers $+\infty$.

(b) Déterminer le développement limité de $\zeta(\beta) - S_n(f)$ avec reste $R_{n,k}$. En prenant $n = 5$ et $k = 5$, donner un encadrement de $\zeta(3)$.

6.6. On applique ici la formule d'Euler-Maclaurin à la fonction $f(x) = e^{ax}$, $a \in \mathbb{C}$.

(a) Montrer l'égalité

$$\frac{a}{2}\,\frac{e^a+1}{e^a-1} = 1 + \sum_{m=1}^{k} \frac{b_{2m} a^{2m}}{(2m)!} - \frac{a^{2k+1}}{e^a-1} \int_0^1 \frac{B_{2k}(x)}{(2k)!}\, e^{ax} dx$$

(b) Montrer que le reste intégral est majoré pour tout $a \in \mathbb{C}$ par

$$\frac{|b_{2k}|}{(2k)!}\,\frac{e^{|\operatorname{Re} a|}-1}{|\operatorname{Re} a|}\,\frac{|a|}{|e^a-1|}\,|a|^{2k} \quad \text{si} \quad e^a \neq 1.$$

En déduire que $\frac{a}{2}\,\frac{e^a+1}{e^a-1} = 1 + \sum_{m=1}^{+\infty} \frac{b_{2m} a^{2m}}{(2m)!}$ sur le disque $|a| < 2\pi$, et que le rayon de convergence de la série est 2π.

(c) Lorsque a est réel, montrer que le reste intégral est majoré par $|b_{2k}| a^{2k}/(2k)!$, ainsi que par $2|b_{2k+2}| a^{2k+2}/(2k+2)!$.

Utiliser ceci pour trouver une valeur approchée de $(e+1)/(e-1)$ en prenant $k = 4$. Vérifier que l'erreur commise est inférieure à 10^{-7}.

6.7. On considère la fonction

$$f(x) = \frac{1}{1+x^2}, \quad x \in \mathbb{R}.$$

(a) A l'aide d'une décomposition en éléments simples, calculer la dérivée $f^{(m)}$ et montrer que $|f^{(m)}(x)| \leq m!\,(1+x^2)^{-(m+1)/2}$.

(b) Déterminer le développement asymptotique de la suite

$$S_n = \sum_{k=0}^{n} \frac{1}{1+k^2}.$$

(c) Calculer S_{10} et en déduire une valeur approchée à 10^{-6} près de la somme

$$\sum_{n=0}^{+\infty} \frac{1}{1+n^2}.$$

6.8. On se propose ici d'étudier une méthode d'intégration numérique analogue aux méthodes de Newton-Cotes.

(a) Soit g une fonction continue sur $[-1,2]$. Déterminer le polynôme $p(x)=\sum_{i=-1}^{2} g(i)\ell_i(x)$ de degré ≤ 3 qui interpole g aux points $-1,0,1,2$.
Exprimer l'erreur d'interpolation à l'aide du polynôme $\pi(x)=x(x+1)(x-1)(x-2)$.

(b) Calculer $\int_0^1 p(x)dx$ et $\int_{-1}^0 p(x)dx$ en fonction des valeurs $g(i)$, $-1\leq i\leq 2$. En déduire $\int_1^2 p(x)dx$.
Vérifier les formules pour $g(x)=1$ (resp. $g(x)=x$).

(c) Calculer $\int_0^1 |\pi(x)|dx$ et $\int_{-1}^0 |\pi(x)|dx$.

En déduire une majoration (la meilleure possible !) de $\int_i^{i+1} |g(x)-p(x)|dx$, $i=-1,0,1$, en fonction de la norme uniforme d'une dérivée convenable de g (g est supposée suffisamment dérivable).

(d) Soit f une fonction continue sur un intervalle $[a,b]$ avec $a<b$. On note

$$a=a_0<a_1<\ldots<a_{n-1}<a_n=b,\quad n\geq 8$$

la subdivision de pas constant $h=\frac{b-a}{n}$ et on pose $f_i=f(a_i)$.
On étudie la méthode d'intégration numérique

$$\int_a^b f(x)dx=\sum_{i=0}^{n-1}\int_{a_i}^{a_{i+1}} f(x)dx\simeq\sum_{i=0}^{n-1}\int_{a_{i+1}}^{a_i} p_i(x)dx$$

où p_i désigne le polynôme d'interpolation de Lagrange de f aux points a_{i-1}, a_i, a_{i+1}, a_{i+2} si $1\leq i\leq n-2$, avec la *convention d'écriture* $p_0=p_1$, $p_{n-1}=p_{n-2}$. Montrer que cette méthode s'écrit

$$\int_a^b f(x)dx\simeq h\sum_{i=0}^{n}\lambda_i f_i$$

pour des coefficients λ_i que l'on explicitera. Que peut-on dire de l'ordre de la méthode ?

(e) Majorer les erreurs $\int_{a_i}^{a_{i+1}} |f(x)-p_i(x)|dx$ et

$$E(f)=\int_a^b f(x)dx-h\sum_{i=0}^{n}\lambda_i f_i$$

en fonction de h, $b-a$, et de la norme uniforme d'une dérivée convenable de f.

6.9. On désigne par $\mathcal{C}$ l'espace des fonctions définies sur l'intervalle $[-1,1]$ à valeurs dans $\mathbb{R}$, muni de la norme uniforme.

(a) Montrer que pour tout $f \in \mathcal{C}$ l'intégrale

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx \quad \text{est convergente.}$$

(b) On note t_n le polynôme de Tchebychev de degré n.
On rappelle le résultat $\int_{-1}^{1} \frac{t_n(x)t_k(x)}{\sqrt{1-x^2}}\, dx = \frac{\pi}{2}\, \delta_{n,k}$ où $\delta_{n,k}$ est le symbole de Kronecker. Calculer $\int_{-1}^{1} \frac{x^n t_m(x)}{\sqrt{1-x^2}}\, dx$ pour $n < m$.

(c) On note x_0, x_1, x_2 les racines de t_3. Déterminer trois réels A_0, A_1, A_2 tels que pour tout polynôme P de degré ≤ 2 on ait

$$\int_{-1}^{1} \frac{P(x)}{\sqrt{1-x^2}}\, dx = A_0P(x_0) + A_1P(x_1) + A_2P(x_2).$$

Montrer que l'égalité est encore vérifiée si P est de degré ≤ 5.

(d) Montrer que l'intégrale $\int_0^1 \frac{x^4 dx}{\sqrt{x(1-x)}}$ est convergente et, à l'aide de (c), calculer sa valeur.

(e) Pour n fixé non nul, on désigne par x_k les racines de t_n et par A_k des nombres réels $(0 \leq k \leq n-1)$. Pour tout $f \in \mathcal{C}$ on note

$$S_n(f) = \sum_{k=0}^{n-1} A_k f(x_k) \quad \text{et} \quad R_n(f) = \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx - S_n(f).$$

(α) Montrer que l'on peut déterminer les A_k de manière unique de sorte que pour tout polynôme P de degré $\leq n-1$ on ait $R_n(P) = 0$.

(β) Montrer que pour $1 \leq p \leq n-1$ on a $\sum_{k=0}^{n-1} T_p(x_k) = 0$.
En déduire que $A_k = \frac{\pi}{n}$ pour tout k.

(γ) Montrer que $R_n(P) = 0$ pour tout polynôme de degré $\leq 2n-1$.

(f) Soient $f \in \mathcal{C}$ et P un polynôme. En supposant $\|f-p\| < \varepsilon$, donner un majorant de $|R_n(f)|$ lorsque $n \to +\infty$. En déduire

$$\lim_{n\to+\infty} S_n(f) = \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx.$$

6.10. Le but du problème est d'établir quelques résultats sur la formule approchée $\int_{-1}^{1} f(x)dx \simeq \int_{-1}^{1} P_n(x)dx$ où $P_n(x)$ est le polynôme d'interpolation de degré n de f aux points de Tchebychev $x_i = \cos\theta_i$, tels que $\theta_i = \frac{(2i+1)}{2n+2}\, \pi$, $0 \leq i \leq n$.

(a) Avec les notations de II 4.3, montrer que le polynômes l_i de Lagrange sont donnés par

$$l_i(x) = \frac{(-1)^i \sin\theta_i}{n+1}\, \frac{t_{n+1}(x)}{x - x_i}, \quad 0 \leq i \leq n.$$

(b) Pour $x \in [-1,1]$ et $n \in \mathbb{N}$ on note

$$a_n(x) = \int_{-1}^{1} \frac{\cos(n \operatorname{Arc}\cos x) - \cos(n \operatorname{Arc}\cos y)}{x-y}\, dy.$$

Montrer que a_n est un polynôme de degré n et que l'on a

$$\int_{-1}^{1} P_n(x)dx = \sum_{i=0}^{n} \omega_i f(x_i) \quad \text{avec} \quad \omega_i = \frac{(-1)^i \sin\,\theta_i}{n+1}\, a_{n+1}(x_i).$$

(c) Calculer $a_{n+1}(x) - a_{n-1}(x)$; en déduire la valeur de

$$a_{n+1}(x) - 2x a_n(x) + a_{n-1}(x).$$

(d) En distinguant deux cas suivant la parité de n, montrer l'égalité

$$\sin\,\theta\; a_n(\cos\,\theta) = 2\,\sin\, n\theta - 4 \sum_{1 \le q < \frac{n}{2}} \frac{1}{4q^2-1}\, \sin\,(n-2q)\theta.$$

(e) En déduire l'expression de ω_i :

$$\omega_i = \frac{\left[2 - 4 \sum\limits_{1 \le q \le \frac{n+1}{2}} \frac{1}{4q^2-1}\, \cos\, 2q\theta_i\right]}{n+1}.$$

Montrer l'inégalité $\omega_i > 0$.

(f) On fixe $n = 10$. Écrire un programme en langage informatique qui permet de calculer tous les coefficients ω_i.

CHAPITRE IV

MÉTHODES ITÉRATIVES POUR LA RÉSOLUTION D'ÉQUATIONS

Les méthodes itératives, et en particulier la méthode de Newton, figurent parmi les méthodes numériques les plus puissantes permettant la résolution approchée des équations de toute nature. L'idée de ces méthodes est de partir d'une valeur approchée grossière de la solution, et d'en améliorer la précision par une application itérée d'un algorithme bien choisi.

1. PRINCIPE DES MÉTHODES ITÉRATIVES

1.1. LE THÉORÈME DU POINT FIXE

Soit (E, d) un espace métrique complet et $\varphi : E \to E$ une application continue. On dit que $a \in E$ est un *point fixe* de φ si $\varphi(a) = a$. On dit que φ est *contractante* si φ est lipschitzienne de rapport $k < 1$, c'est-à-dire s'il existe $k < 1$ tel que

$$\forall x, y \in E, \quad d(f(x), f(y)) \leq k\, d(x, y).$$

Théorème – *Soit $\varphi : E \to E$ une application contractante d'un espace métrique complet dans lui-même. Alors φ admet un point fixe unique $a \in E$. De plus, pour tout point initial $x_0 \in E$, la suite itérée (x_p) définie par $x_{p+1} = \varphi(x_p)$ converge vers a.*

Unicité du point fixe. Si φ avait deux points fixes $a \neq b$, alors $d(\varphi(a), \varphi(b)) = d(a, b)$ et $d(a, b) \neq 0$, donc φ ne pourrait être contractante, contradiction.

Existence du point fixe. Soit $x_0 \in E$ un point initial quelconque et (x_p) la suite itérée associée. On a alors

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

d'où par récurrence $d(x_p, x_{p+1}) \leq k^p d(x_0, x_1)$. Pour tout entier $q > p$ il vient

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

avec $\displaystyle\sum_{l=p}^{q-1} k^l \leq \sum_{l=p}^{+\infty} k^l = \frac{k^p}{1-k}$. On a donc

$$d(x_p, x_q) \leq \frac{k^p}{1-k}\, d(x_0, x_1), \quad \forall p < q$$

ce qui montre que (x_p) est une suite de Cauchy. Comme (E, d) est complet, la suite (x_p) converge vers un point limite $a \in E$. L'égalité $x_{p+1} = \varphi(x_p)$ et la continuité de φ impliquent à la limite $a = \varphi(a)$. ■

Estimation de la vitesse de convergence – L'inégalité

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

implique par récurrence

$$d(x_p, a) \leq k^p d(x_0, a).$$

La convergence est donc exponentiellement rapide. Lorsque $E = \mathbb{R}^m$, on dit parfois qu'il s'agit d'une *convergence linéaire*, dans le sens où le nombre de décimales exactes de x_p croît au moins linéairement avec p.

Généralisation – *Le théorème précédent reste entièrement valable si on remplace l'hypothèse que φ est contractante par l'hypothèse que φ est continue et qu'il existe une certaine itérée $\varphi^m = \varphi \circ \ldots \circ \varphi$ qui soit contractante.*

En effet, dans ce cas, l'hypothèse que φ^m soit contractante implique que φ^m admet un unique point fixe a. On a donc $\varphi^m(a) = a$ et en appliquant φ à cette égalité on trouve

$$\varphi^m(\varphi(a)) = \varphi^{m+1}(a) = \varphi(\varphi^m(a)) = \varphi(a),$$

de sorte que $\varphi(a)$ est encore un point fixe de φ^m. L'unicité du point fixe de φ^m entraîne $\varphi(a) = a$. Par ailleurs, comme tout point fixe de φ est aussi un point fixe de φ^m, ce point fixe est nécessairement unique. Enfin, pour tout point initial x_0, la sous-suite $x_{mp} = \varphi^{mp}(x_0) = (\varphi^m)^p(x_0)$ (correspondant aux indices multiples de m) converge vers a. Il en résulte que $x_{mp+r} = \varphi^r(x_{mp})$ converge aussi vers $\varphi^r(a) = a$ pour $r = 0, 1, \ldots m-1$, et on en déduit $\lim_{q \to +\infty} x_q = a$. Voir aussi le problème 4.1 pour une autre démonstration de ces résultats. ■

1.2. Application à la résolution d'équations

Comme première application élémentaire du résultat précédent, soit à résoudre une équation $f(x) = 0$ d'une variable réelle x. Supposons qu'on ait une fonction différentiable $f : [a, b] \to \mathbb{R}$ telle que disons $f(a) < 0$, $f(b) > 0$, et f strictement croissante, $0 < m \leq f'(x) \leq M$ sur $[a, b]$ (dans le cas opposé $f(a) > 0$, $f(b) < 0$, et $-M \leq f'(x) < -m < 0$ il suffira de changer f en $-f$). Si on pose

$$\varphi(x) = x - Cf(x)$$

avec une constante $C \neq 0$, il est clair que l'équation $f(x) = 0$ équivaut à $\varphi(x) = x$ et donc la résolution de l'équation $f(x) = 0$ se ramène à rechercher les points fixes de φ. L'espace $E = [a, b]$ est complet, et il nous faut vérifier de plus

- que φ envoie bien E dans E,
- que φ est bien contractante sur E.

Or nous avons $\varphi'(x) = 1 - Cf'(x)$, donc $1 - CM \leq \varphi'(x) \leq 1 - Cm$, et pour le choix $C = 1/M$, la fonction φ est bien contractante dans le rapport $k = 1 - m/M$. De plus φ est croissante et on a $\varphi(a) > a$, $\varphi(b) < b$, donc $\varphi([a, b]) \subset [a, b]$. Il en résulte que toute suite itérative $x_{p+1} = \varphi(x_p)$ calculée à partir d'un point $x_0 \in [a, b]$ quelconque va converger vers l'unique solution de l'équation $f(x) = 0$.

La vitesse de convergence peut être estimée par la suite géométrique $(1 - m/M)^p$, et on voit qu'on a intérêt à ce que les bornes m et M de l'encadrement $m \leq f' \leq M$ soient proches, ce qui est toujours possible si f' est continue et si l'encadrement initial $[a, b]$ de la solution x cherchée est suffisamment fin. L'objet de ce chapitre est d'étudier et de généraliser ce type de techniques, pour des fonctions d'une ou plusieurs variables.

2. Cas des fonctions d'une variable

2.1. Points fixes attractifs et répulsifs

Notre objectif est ici d'étudier le comportement itératif d'une fonction au voisinage de ses points fixes. Soit I un intervalle fermé de $\mathbb{R}$ et $\varphi : I \to I$ une application de classe C^1. Soit $a \in I$ un point fixe de φ. On peut distinguer trois cas :

(1) $|\varphi'(a)| < 1$.

Soit k tel que $|\varphi'(a)| < k < 1$. Par continuité de φ', il existe un intervalle $E = [a - h, a + h]$ sur lequel $|\varphi'| \leq k$, donc φ est contractante de rapport k sur E ; on a nécessairement $\varphi(E) \subset E$ et par conséquent

$$\forall x_0 \in [a - h, a + h], \quad \lim_{p \to +\infty} x_p = a.$$

On dit que a est un *point fixe attractif*. Dans ce cas la convergence de la suite (x_p) est au moins exponentiellement rapide : $|x_p - a| \leq k^p |x_0 - a|$.

Cas particulier : $\varphi'(a) = 0$.

Supposons de plus que φ soit de classe C^2 et que $|\varphi''| \leq M$ sur E. La formule de Taylor donne

$$\begin{aligned}\varphi(x) &= \varphi(a) + (x-a)\varphi'(a) + \frac{(x-a)^2}{2!}\,\varphi''(c)\\ &= a + \frac{1}{2}\,\varphi''(c)(x-a)^2, \qquad c \in\,]a, x[,\end{aligned}$$

d'où $|\varphi(x) - a| \leq \frac{1}{2}\,M|x-a|^2$, soit encore $\frac{1}{2}\,M|\varphi(x)-a| \leq \left[\frac{1}{2}\,M|x-a|\right]^2$. Par récurrence, on en déduit successivement

$$\frac{1}{2}\,M|x_p - a| \leq \left[\frac{1}{2}\,M|x_0 - a|\right]^{2^p},$$

$$|x_p - a| \leq \tfrac{2}{M}\left[\tfrac{1}{2}\,M|x_0 - a|\right]^{2^p}.$$

En particulier si x_0 est choisi tel que $|x_0 - a| \leq \frac{1}{5M}$, on obtient

$$|x_p - a| \leq \frac{2}{M}\,10^{-2^p}\ ;$$

on voit donc que le nombre de décimales exactes double environ à chaque itération ; 10 itérations suffiraient ainsi théoriquement pour obtenir plus de 1000 décimales exactes ! La convergence est donc ici extraordinairement rapide.

Ce phénomène est appelé phénomène de *convergence quadratique*, et le point fixe a est alors appelé parfois point fixe *superattractif.*

(2) $|\varphi'(a)| > 1$.

Comme $\lim\limits_{x\to 0}\left|\dfrac{\varphi(x)-\varphi(a)}{x-a}\right| = |\varphi'(a)| > 1$, on voit qu'il existe un voisinage $[a-h, a+h]$ de a tel que

$$\forall x \in [a-h, a+h] \setminus \{a\}, \quad |\varphi(x) - a| > |x - a|.$$

On dit alors que le point fixe a est *répulsif.* Dans ce cas, la dérivée φ' est de signe constant au voisinage de a, donc il existe $h > 0$ tel que la restriction $\varphi|_{[a-h,a+h]}$ admette une application réciproque φ^{-1} définie sur $\varphi([a-h, a+h])$, qui est un intervalle contenant $\varphi(a) = a$. L'équation $\varphi(x) = x$ peut se récrire $x = \varphi^{-1}(x)$ au voisinage de a, et comme $(\varphi^{-1})'(a) = 1/\varphi'(a)$, le point a est un point fixe attractif pour φ^{-1}.

(3) $|\varphi'(a)| = 1$.

On est ici dans *un cas douteux*, comme le montrent les deux exemples suivants dans lesquels $a = 0$, $\varphi'(a) = 1$:

Exemple 1 – $\varphi(x) = \sin x$, $x \in \left[0, \frac{\pi}{2}\right]$. On a ici $\sin x < x$ pour tout $x \in \left]0, \frac{\pi}{2}\right]$. Pour tout $x_0 \in \left]0, \frac{\pi}{2}\right]$ la suite itérée (x_p) est strictement décroissante minorée, donc convergente. La limite l vérifie $l = \sin l$, donc $l = 0$.

Exemple 2 – $\varphi(x) = \sinh x$, $x \in [0, +\infty[$. Comme $\sinh x > x$ pour tout $x > 0$, on voit que le point fixe 0 est répulsif et que $\forall x_0 > 0$, $\lim\limits_{p \to +\infty} x_p = +\infty$.

2.2. Comportement graphique

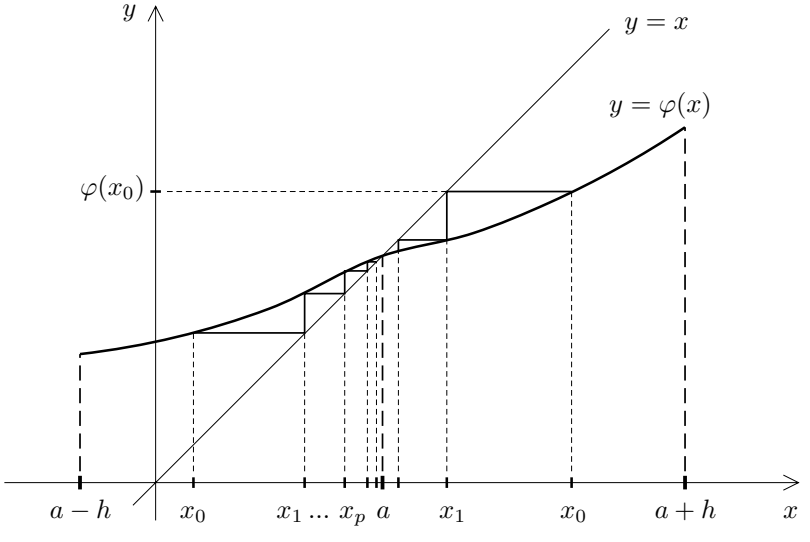
Nous voulons décrire ici un peu plus finement le comportement de la suite itérative $x_{p+1} = \varphi(x_p)$ au voisinage d'un point fixe attractif a. On suppose donc φ de classe C^1 et $|\varphi'(a)| < 1$. On peut de nouveau distinguer plusieurs cas.

(1) $\varphi'(a) > 0$.

Par continuité de φ' on va avoir $0 < \varphi'(x) < 1$ au voisinage de a, donc il existe un voisinage $[a-h, a+h]$ sur lequel $x \mapsto \varphi(x)$ et $x \mapsto x - \varphi(x)$ sont strictement croissantes, par suite

$$\begin{aligned} &x < \varphi(x) < \varphi(a) = a \quad \text{pour } x \in [a-h, a[\\ &a = \varphi(a) < \varphi(x) < x \quad \text{pour } x \in]a, a+h], \end{aligned}$$

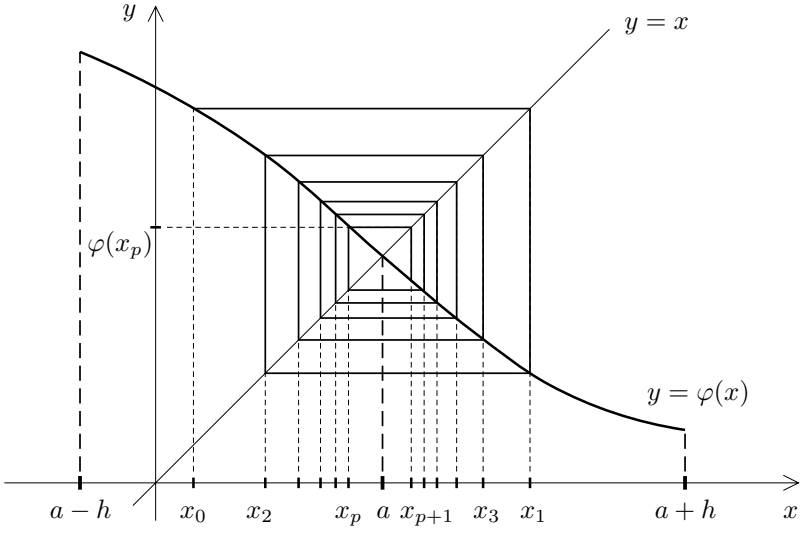
ce qui implique en particulier que $\varphi([a-h, a+h] \subset [a-h, a+h]$. Il est facile de voir dans ce cas que la suite itérative $x_{p+1} = \varphi(x_p)$ va être strictement croissante pour $x_0 \in [a-h, a[$ et strictement décroissante si $x_0 \in]a, a+h]$. On obtient alors typiquement un graphe « en escalier » :



(2) $\varphi'(a) < 0$.

Par continuité de φ' on va avoir $-1 < \varphi'(x) < 0$ sur un voisinage $[a-h, a+h]$ de a, sur lequel φ est donc strictement décroissante. Si $x < a$, alors $\varphi(x) > \varphi(a) = a$,

tandis que si $x > a$, on $\varphi(x) < \varphi(a) = a$. Comme $\varphi \circ \varphi$ est strictement croissante (de dérivée < 1) sur $[a-h, a+h]$, le cas (1) montre que les suites x_{2p} et x_{2p+1} sont monotones de limite Il s'agit donc de suites *adjacentes*, et on obtient un graphe « en escargot » :



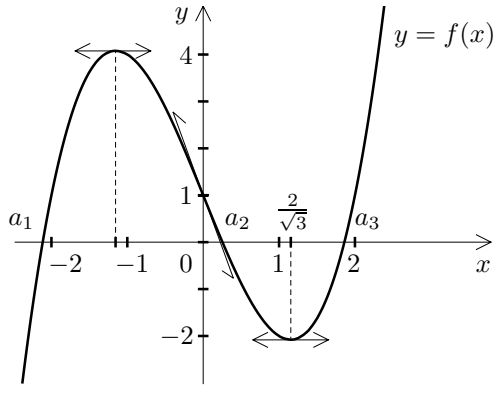
(3) $\varphi'(a) = 0$.

En général, on ne va rien pouvoir conclure. Cependant, si φ est de classe C^2 et $\varphi''(a) \neq 0$, alors le point a est un extremum local. On choisira h assez petit pour que φ'' ne change pas de signe et $|\varphi'| < 1$ sur $[a-h, a+h]$. Alors, si $\varphi''(a) < 0$ (resp. $\varphi''(a) > 0$), le point a est un maximum local (resp. un minimum local), et pour $x_0 \in [a-h, a+h] \setminus \{a\}$ quelconque, il est facile de voir que la suite (x_p) est strictement croissante (resp. décroissante), mis à part peut-être pour le terme initial x_0. On aboutit encore à un graphe en escalier.

2.3. *Exemple de résolution d'équation*

Soit à résoudre l'équation $f(x) = x^3 - 4x + 1 = 0$, $x \in \mathbb{R}$. On a $f'(x) = 3x^2 - 4$ et l'étude de f donne :

x	$-\infty$		$-\frac{2}{\sqrt{3}}$		$\frac{2}{\sqrt{3}}$		$+\infty$
$f'(x)$		$+$	0	$-$	0	$-$	
$f(x)$	$-\infty$	$\nearrow$	$1+\frac{16}{3\sqrt{3}}$	$\searrow$	$1-\frac{16}{3\sqrt{3}}$	$\nearrow$	$+\infty$



L'équation $f(x) = 0$ admet donc 3 racines réelles $a_1 < a_2 < a_3$. le calcul de quelques valeurs de f donne

$$-2,5 < a_1 < -2, \quad 0 < a_2 < 0,5, \quad 1,5 < a_3 < 2.$$

L'équation $f(x) = 0$ peut se récrire $x = \varphi(x)$ avec $\varphi(x) = \frac{1}{4}(x^3 + 1)$. On a $\varphi'(x) = \frac{3}{4}x^2$, d'où :

- sur $[-2,5\,;\,-2]$, $\quad \varphi' \geq \varphi'(2) = 3.$
- sur $[0\,;\,0,5]$, $\quad 0 \leq \varphi' \leq 0,1875.$
- sur $[1,5\,;\,2]$, $\quad \varphi' \geq \varphi'(1,5) = 1,6875.$

Seul a_2 est un point fixe attractif de φ. L'intervalle $[0\,;\,0,5]$ est nécessairement stable par φ puisqu'il contient un point fixe et que φ est contractante et croissante. Pour tout $x_0 \in [0\,;\,0,5]$ on aura donc $a_2 = \lim\limits_{p\to+\infty} x_p$.

Pour obtenir a_1 et a_3, on peut itérer la fonction $\varphi^{-1}(x) = \sqrt[3]{4x-1}$, qui est contractante au voisinage de ces points.

Il sera numériquement plus efficace de récrire l'équation sous la forme $x^2 - 4 + \frac{1}{x} = 0$, soit $x = \varphi_+(x)$ ou $x = \varphi_-(x)$ avec

$$\varphi_+(x) = \sqrt{4 - \frac{1}{x}}, \quad \varphi_-(x) = -\sqrt{4 - \frac{1}{x}},$$

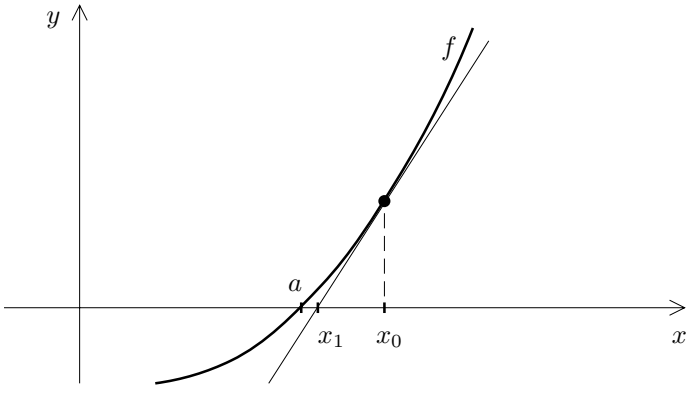
suivant que x est ≥ 0 ou ≤ 0. on a alors $\varphi'_\pm = \pm\frac{1}{2x^2}\left(4 - \frac{1}{x}\right)^{-1/2}$, de sorte que

$$\begin{aligned} &0 \leq \varphi'_+ \leq \varphi'_+(1,5) \simeq 0,122 \quad \text{sur} \quad [1,5\,;\,2], \\ &\varphi'_-(-2) \simeq -0,059 \leq \varphi' \leq 0 \quad \text{sur} \quad [-2,5\,;\,-2]\ ; \end{aligned}$$

La convergence sera donc assez rapide. Nous allons voir qu'il existe en fait une méthode générale plus efficace et plus systématique.

2.4. MÉTHODE DE NEWTON

On cherche à évaluer numériquement la racine a d'une équation $f(x) = 0$, en supposant qu'on dispose d'une valeur grossière x_0 de cette racine.



L'idée est de remplacer la courbe représentative de f par sa tangente au point x_0 :

$$y = f'(x_0)(x - x_0) + f(x_0).$$

L'abscisse x_1 du point d'intersection de cette tangente avec l'axe $y = 0$ est donnée par

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \ ;$$

x_1 est en général une meilleure approximation de a que x_0. On est donc amené à itérer la fonction

$$\varphi(x) = x - \frac{f(x)}{f'(x)}.$$

Supposons que f soit de classe C^2 et que $f'(a) \neq 0$. La fonction φ est alors de classe C^1 au voisinage de a et

$$\varphi'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2},$$

ce qui donne $\varphi(a) = a$, $\varphi'(a) = 0$. La racine a de $f(x) = 0$ est donc un point fixe superattractif de φ. Le résultat suivant donne une estimation de l'écart $|x_p - a|$.

Théorème – *On suppose que f est de classe C^2 sur l'intervalle $I = [a-r, a+r]$ et que $f' \neq 0$ sur I. Soit $M = \max\limits_{x\in I} \left|\dfrac{f''(x)}{f'(x)}\right|$ et $h = \min\left(r, \frac{1}{M}\right)$. Alors pour tout $x \in [a-h, a+h]$ on a $|\varphi(x)-a| \leq M|x-a|^2$, et pour tout point initial $x_0 \in [a-h, a+h]$*

$$|x_p - a| \leq \frac{1}{M}\left(M|x_0 - a|\right)^{2^p}.$$

Démonstration. Introduisons la fonction $u(x) = f(x)/f'(x)$. La fonction f est monotone sur I, nulle en a, donc f a même signe que $f'(a)(x-a)$, ce qui entraîne que $u(x)$ a même signe que $x-a$. De plus

$$u'(x) = 1 - \frac{f(x)f''(x)}{f'(x)^2} = 1 - \frac{f''(x)}{f'(x)}\,u(x),$$

donc $|u'(x)| \leq 1 + M|u(x)|$ sur I. Cette inégalité permet d'obtenir le

Lemme 1 – *On a* $|u(x)| \leq \frac{1}{M}\,(e^{M|x-a|} - 1)$ *sur* I.

Montrons-le par exemple pour $x \geq a$. Posons $v(x) = u(x)e^{-Mx}$. Il vient $u'(x) \leq 1 + Mu(x)$, d'où

$$v'(x) = (u'(x) - Mu(x))e^{-Mx} \leq e^{-Mx}.$$

Comme $v(a) = u(a) = 0$, on en déduit par intégration

$$v(x) \leq \frac{1}{M}\,(e^{-Ma} - e^{-Mx}),$$

soit encore $u(x) \leq \frac{1}{M}\,(e^{M(x-a)} - 1)$. Le lemme 1 est donc démontré. ■

Lemme 2 – *Pour* $|t| \leq 1$, $e^{|t|} - 1 \leq 2|t|$.

En effet la fonction exponentielle est convexe, donc sur tout intervalle la courbe est située sous sa corde. Sur l'intervalle $[0,1]$, ceci donne $e^t \leq 1 + (e-1)t$, d'où le lemme 2 puisque $e - 1 < 2$. ■

On peut maintenant écrire $\varphi'(x) = u(x)\,\frac{f''(x)}{f'(x)}$, et le lemme 1 implique

$$|\varphi'(x)| \leq M|u(x)| \leq e^{M|x-a|} - 1.$$

Grâce au lemme 2, on obtient $|\varphi'(x)| \leq 2M|x-a|$ pour $|x-a| \leq \min\left(r, \frac{1}{M}\right)$. Comme $\varphi(a) = a$, on voit par intégration que pour tout $x \in [a-h, a+h]$ on a

$$|\varphi(x) - a| \leq M|x-a|^2,$$

soit encore $M|\varphi(x) - a| \leq (M|x-a|)^2$. L'estimation $M|x_p - a| \leq (M|x_0 - a|)^{2p}$ s'en déduit aussitôt par récurrence.

Exemple – Pour la fonction $f(x) = x^3 - 4x + 1$ du § 2.3, on a

$$\varphi(x) = x - \frac{x^3 - 4x + 1}{3x^2 - 4} = \frac{2x^3 - 1}{3x^2 - 4}.$$

Par itération de φ, on obtient alors les valeurs suivantes :

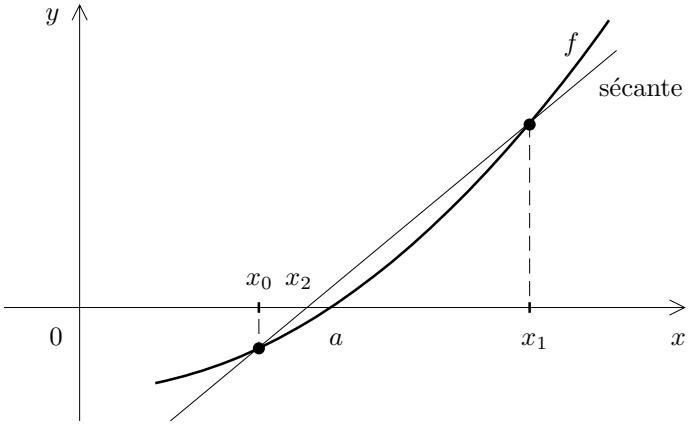
x_0	-2	0	2
x_1	$-2,125$	$0,25$	$1,875$
x_2	$-2,114975450$	$0,254098361$	$1,860978520$
x_3	$-2,114907545$	$0,254101688$	$1,860805877$
x_4	$-2,114907541$	$= x_3$	$1,860805853$
x_5	$= x_4$		$= x_4$

Ceci donne des valeurs approchées de a_1, a_2, a_3 à 10^{-9} près environ. Le nombre d'itérations nécessaires pour obtenir une précision de 10^{-9} par la méthode de Newton est typiquement 3 ou 4 ($10^{-2^3} = 10^{-8}$, $10^{-2^4} = 10^{-16}\ldots$). Le lecteur pourra vérifier que le nombre d'itérations requises avec les fonctions φ du § 2.3 est nettement plus élevé (de 8 à 20 suivant les cas).

2.4. *Méthode de la sécante*

Dans certaines situations, la dérivée f' est très compliquée ou même impossible à expliciter (c'est le cas par exemple si la fonction f est le résultat d'un algorithme complexe). On ne peut alors utiliser telle quelle la méthode de Newton.

L'idée est de remplacer f' par le taux d'accroissement de f sur un petit intervalle. Supposons qu'on dispose de deux valeurs approchées x_0, x_1 de la racine a de l'équation $f(x) = 0$ (fournies par un encadrement $x_0 < a < x_1$).



Le taux d'accroissement de f sur l'intervalle $[x_0, x_1]$ est

$$\tau_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

et l'équation de la sécante traversant le graphe de f aux points d'abscisse x_0 et x_1 est

$$y = \tau_1(x - x_1) + f(x_1).$$

On obtient ainsi une nouvelle approximation x_2 de a en calculant l'abscisse de l'intersection de la sécante avec l'axe Ox :

$$x_2 = x_1 - \frac{f(x_1)}{\tau_1}.$$

On va bien entendu itérer ce procédé à partir des nouvelles valeurs approchées x_1 et x_2, ce qui conduit à poser

$$\tau_p = \frac{f(x_p) - f(x_{p-1})}{x_p - x_{p-1}}, \quad x_{p+1} = x_p - \frac{f(x_p)}{\tau_p}.$$

La méthode est donc tout à fait analogue à celle de Newton, à ceci près que l'on a remplacé la dérivée $f'(x_p)$ par le taux d'accroissement τ_p de f sur l'intervalle $[x_p, x_{p-1}]$. On notera que l'algorithme itératif ne peut démarrer que si on dispose déjà de *deux* valeurs approchées x_0, x_1 de a.

Inconvénient de la méthode – Lorsque x_p et x_{p-1} sont trop voisins, le calcul de $f(x_p) - f(x_{p-1})$ et $x_p - x_{p-1}$ donne lieu à un phénomène de compensation et donc à une perte de précision sur le calcul de τ_p. Étudions l'erreur commise. La formule de Taylor-Lagrange à l'ordre 2 au point x_p donne

$$f(x_{p-1}) - f(x_p) = (x_{p-1} - x_p)f'(x_p) + \frac{1}{2}(x_{p-1} - x_p)^2 f''(c),$$

$$\tau_p - f'(x_p) = \frac{1}{2}(x_{p-1} - x_p)f''(c) = O(|x_p - x_{p-1}|)$$

après division de la première ligne par $x_{p-1} - x_p$.

Supposons par ailleurs que le calcul des $f(x_i)$ soit effectué avec une erreur d'arrondi de l'ordre de ε. Le calcul de τ_p est alors affecté d'une erreur absolue de l'ordre de $\frac{\varepsilon}{|x_p - x_{p-1}|}$. Il est inutile de continuer à calculer τ_p dès que cette erreur dépasse l'écart $|\tau_p - f'(x_p)|$, ce qui a lieu si $\frac{\varepsilon}{|x_p - x_{p-1}|} > |x_p - x_{p-1}|$ c'est-à-dire $|x_p - x_{p-1}| < \sqrt{\varepsilon}$.

Dans la pratique, si l'on dispose d'une précision absolue $\varepsilon = 10^{-10}$ par exemple, on arrête le calcul de τ_p dès que $|x_p - x_{p-1}| < \sqrt{\varepsilon} = 10^{-5}$; on poursuit alors les itérations avec $\tau_p = \tau_{p-1}$ jusqu'à l'obtention de la convergence (c'est-à-dire $|x_{p+1} - x_p| < \varepsilon$).

D'un point de vue théorique, la convergence de la suite est assurée par le résultat ci-dessous, qui donne simultanément une estimation précise pour $|x_p - a|$.

Théorème – *On suppose f de classe C^2 et de dérivée $f' \neq 0$ sur l'intervalle $I = [a - r, a + r]$. On introduit les quantités M_i, $i = 1, 2$, et les réels K, h tels que*

$$M_i = \max_{x \in I} |f^{(i)}(x)|, \qquad m_i = \min_{x \in I} |f^{(i)}(x)|,$$

$$K = \frac{M_2}{2m_1}\left(1 + \frac{M_1}{m_1}\right), \qquad h = \min\left(r, \frac{1}{K}\right).$$

Soit enfin (s_p) la suite de Fibonacci, définie par $s_{p+1} = s_p + s_{p-1}$ avec $s_0 = s_1 = 1$. Alors quel que soit le choix des points initiaux x_0, $x_1 \in [a - h, a + h]$ distincts, on a

$$|x_p - a| \leq \frac{1}{K}\left[K \max(|x_0 - a|, |x_1 - a|)\right]^{s_p}.$$

Remarque – On vérifie facilement que $s_p \sim \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^{p+1}$; ceci montre que le nombre de décimales exactes croît environ du facteur $\frac{1+\sqrt{5}}{2} \simeq 1,618$ à chaque itération. La convergence est donc tout juste un peu moins rapide que dans le § 2.4.

Démonstration.* Le lecteur pourra omettre cette démonstration sans compromettre la compréhension de la suite du chapitre. On considère le taux d'accroissement $\tau(x,y)$ de f sur $I \times I$ défini par

$$\begin{cases} \tau(x,y) = \frac{f(y)-f(x)}{y-x} & \text{si } y \neq x \\ \tau(x,y) = f'(x) & \text{si } y = x. \end{cases}$$

Pour tous $(x,y) \in I \times I$, on peut écrire

$$\tau(x,y) = \int_0^1 f'(x + t(y-x))dt$$

et le théorème de dérivation sous le signe somme montre que

$$\frac{\partial \tau}{\partial x}(x,y) = \int_0^1 (1-t) f''(x + t(y-x))dt,$$
$$\frac{\partial \tau}{\partial y}(x,y) = \int_0^1 t f''(x + t(y-x))dt.$$

Comme f' est de signe constant sur I et comme $\displaystyle\int_0^1 (1-t)dt = \frac{1}{2}$, on en déduit les inégalités

$$|\tau(x,y)| \geq m_1, \quad \left|\frac{\partial \tau}{\partial x}\right| \leq \frac{1}{2} M_2, \quad \left|\frac{\partial \tau}{\partial y}\right| \leq \frac{1}{2} M_2.$$

Il en résulte en particulier

$$|\tau(x,y) - f'(x)| = |\tau(x,y) - \tau(x,x)| = \left| \int_x^y \frac{\partial \tau}{\partial y}(x,t)dt \right| \leq \frac{1}{2} M_2 |y-x|.$$

La suite (x_p) est définie par la formule de récurrence $x_{p+1} = \psi(x_p, x_{p-1})$ où ψ est la fonction de classe C^1 telle que

$$\psi(x,y) = x - \frac{f(x)}{\tau(x,y)}.$$

Posons $h_p = x_p - a$. On a

$$h_{p+1} = \psi(x_p, x_{p-1}) - a = \psi(a + h_p, a + h_{p-1}) - \psi(a,a)$$

et en particulier $h_2 = \psi(a+h_1, a+h_0) - \psi(a,a)$. En intégrant sur $[0,1]$ la dérivée de la fonction $t \mapsto \psi(a + th_1, a + th_0)$, on trouve

$$h_2 = \int_0^1 \left[h_1 \frac{\partial \psi}{\partial x}(a + th_1, a + th_0) + h_0 \frac{\partial \psi}{\partial y}(a + th_1, a + th_0) \right] dt.$$

Des calculs immédiats donnent

$$\frac{\partial\psi}{\partial x} = 1 - \frac{f'(x)\tau(x,y) - f(x)\frac{\partial\tau}{\partial x}}{\tau(x,y)^2} = \frac{\tau(x,y) - f'(x)}{\tau(x,y)} + f(x)\,\frac{\frac{\partial\tau}{\partial x}}{\tau(x,y)^2},$$
$$\frac{\partial\psi}{\partial y} = f(x)\,\frac{\frac{\partial\tau}{\partial y}}{\tau(x,y)^2}.$$

Comme $|f(x)| = |f(x) - f(a)| \leq M_1|x-a|$, les inégalités ci-dessus impliquent

$$\left|\frac{\partial\psi}{\partial x}\right| \leq \frac{M_2|y-x|}{2m_1} + M_1|x-a|\,\frac{M_2}{2m_1^2},$$
$$\left|\frac{\partial\psi}{\partial y}\right| \leq M_1|x-a|\,\frac{M_2}{2m_1^2}.$$

Pour $(x,y) = (a+th_1, a+th_0)$, on a $|y-x| \leq (|h_0|+|h_1|)t$ et $|x-a| = |h_1|t$, d'où

$$\left|\frac{\partial\psi}{\partial x}\right| \leq \left[\frac{M_2}{2m_1}\,(|h_0|+|h_1|) + \frac{M_1M_2}{2m_1^2}\,|h_1|\right] t,$$
$$\left|\frac{\partial\psi}{\partial y}\right| \leq \frac{M_1M_2}{2m_1^2}\,|h_1|t,$$
$$|h_2| \leq \left(\frac{M_2}{2m_1} + \frac{M_1M_2}{2m_1^2}\right)|h_1|(|h_0|+|h_1|)\int_0^1 t\,dt$$
$$= \frac{K}{2}\,|h_1|(|h_0|+|h_1|) \leq K|h_1|\max(|h_0|,|h_1|).$$

Comme $|h_0|,\ |h_1| \leq h \leq \frac{1}{K}$, on voit que $|h_2| \leq |h_1|$. De même

$$|h_{p+1}| \leq K|h_p|\max(|h_p|,|h_{p-1}|),$$

et ceci entraîne par récurrence que la suite $(|h_p|)_{p\geq 1}$ est décroissante :
$|h_p| \leq |h_{p-1}| \leq \ldots \leq |h_1| \leq \frac{1}{K}$ implique $|h_{p+1}| \leq |h_p|$. On en déduit

$$|h_{p+1}| \leq K|h_p||h_{p-1}| \quad \text{pour} \quad p \geq 2.$$

L'inégalité $|h_p| \leq \frac{1}{K}\,[K\,\max(|h_0|,|h_1|)]^{s_p}$ est triviale si $p=0$ ou $p=1$, résulte de l'estimation déjà vue pour h_2 si $p=2$, et se généralise facilement par récurrence pour $p \geq 3$. Le théorème est démontré. ■

3. Cas des fonctions de $\mathbb{R}^m$ dans $\mathbb{R}^m$

3.1. Préliminaires : rayon spectral d'un endomorphisme

Soit E un espace vectoriel de dimension m sur $\mathbb{R}$ et u un endomorphisme de E.

Définition – *On appelle* spectre *de u la famille $(\lambda_1, \dots, \lambda_m)$ de ses valeurs propres réelles ou complexes, comptées avec multiplicités (= racines du polynôme caractéristique). On appelle* rayon spectral *de u, noté $\rho(u)$, la quantité*

$$\rho(u) = \max_{1 \le i \le m} |\lambda_i|.$$

Étant donné une norme N sur E, on peut d'autre part associer à u sa norme $|||u|||_N$ en tant qu'opérateur linéaire sur E :

$$|||u|||_N = \sup_{x \in E \setminus \{0\}} \frac{N(u(x))}{N(x)}.$$

On notera $\mathcal{N}$ (resp. $\mathcal{N}_e$) l'ensemble des normes (resp. des normes euclidiennes) définies sur E.

Théorème – *Soit u un endomorphisme quelconque de E. Alors*

(1) *Pour tout $N \in \mathcal{N}$, $\rho(u) \le |||u|||_N$.*

(2) $\rho(u) = \inf_{N \in \mathcal{N}} |||u|||_N = \inf_{N \in \mathcal{N}_e} |||u|||_N$.

(3) *Pour tout $N \in \mathcal{N}$, $\rho(u) = \lim_{p \to +\infty} |||u^p|||_N^{1/p}$.*

Remarque – Considérons l'espace $\mathbb{R}^2$ muni de sa norme euclidienne canonique et u_a l'endomorphisme de matrice $\begin{pmatrix} 0 & a \\ 0 & 1 \end{pmatrix}$, $a \in \mathbb{R}$. On a $\rho(u_a) = 1$ et pourtant la norme $|||u_a|||$ n'est pas bornée quand $|a| \to +\infty$; l'inégalité (1) n'a donc pas de réciproque. Pour $m = \dim E \ge 2$, le rayon spectral n'est pas une norme sur $\mathcal{L}(E, E)$; il ne vérifie d'ailleurs pas l'inégalité triangulaire, comme le montre l'exemple suivant : si u, v sont les endomorphismes de matrices

$$\mathrm{Mat}(u) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{et} \quad \mathrm{Mat}(v) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

on a $\rho(u) = \rho(v) = 0$, mais $\rho(u + v) = 1$.

Démonstration*. Une base de E étant fixée, on peut identifier E à $\mathbb{R}^m$ et u à une matrice A carrée $m \times m$. Observons d'abord que si $(\lambda_1, \dots, \lambda_m)$ est le spectre de A, alors le spectre de A^p est $(\lambda_1^p, \dots, \lambda_m^p)$. On a donc

$$\rho(A^p) = \rho(A)^p.$$

(1) Soit λ une valeur propre de A. Si $\lambda \in \mathbb{R}$, soit X un vecteur propre associé, $X \neq 0$. Les égalités $AX = \lambda X$, $N(AX) = |\lambda| N(X)$ entraînent bien $|\lambda| \le |||A|||_N$.

Supposons maintenant que $\lambda = \alpha + i\beta$ soit une valeur propre complexe non réelle de A. Soit $Z = X + iY$ un vecteur colonne complexe, propre pour la valeur propre λ.

L'égalité $AZ = \lambda Z = (\alpha + i\beta)(X + iY)$ donne pour les parties réelles et imaginaires les égalités $AX = \alpha X - \beta Y$, $AY = \beta X + \alpha Y$ d'où

$$\begin{cases} A(\alpha X + \beta Y) = (\alpha^2 + \beta^2)X \\ A(-\beta X + \alpha Y) = (\alpha^2 + \beta^2)Y. \end{cases}$$

Il s'ensuit

$$(\alpha^2 + \beta^2)N(X) \leq |||A|||_N N(\alpha X + \beta Y) \leq |||A|||_N(|\alpha|\, N(X) + |\beta|\, N(Y)),$$
$$(\alpha^2 + \beta^2)N(Y) \leq |||A|||_N N(-\beta X + \alpha Y) \leq |||A|||_N(|\alpha|\, N(Y) + |\beta|\, N(X)).$$

Après addition et simplification par $N(X) + N(Y)$ on obtient

$$|\lambda|^2 = \alpha^2 + \beta^2 \leq |||A|||_N(|\alpha| + |\beta|) \leq 2|\lambda|\, |||A|||_N,$$

d'où $|\lambda| \leq 2|||A|||_N$. Par conséquent $\rho(A) \leq 2|||A|||_N$. D'après la remarque initiale et le résultat précédent appliqué à A^p, il vient

$$\rho(A)^p = \rho(A^p) \leq 2|||A^p|||_N \leq 2|||A|||_N^p$$

soit $\rho(A) \leq 2^{1/p}|||A|||_N$. En faisant tendre p vers $+\infty$, la conclusion attendue $\rho(A) \leq |||A|||_N$ s'ensuit.

(2) D'après (1) et l'inclusion $\mathcal{N}_e \subset \mathcal{N}$ on obtient

$$\rho(A) \leq \inf_{N \in \mathcal{N}} |||A|||_N \leq \inf_{N \in \mathcal{N}_e} |||A|||_N.$$

Il suffit donc de voir que $\inf_{N \in \mathcal{N}_e} |||A|||_N \leq \rho(A)$.

Pour cela, considérons A comme un endomorphisme de $\mathbb{C}^m$, défini par une matrice à coefficients réels. Il existe une base $(e_1, \ldots, e_m)$ de $\mathbb{C}^m$ dans laquelle A devient triangulaire supérieure :

$$A' = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ & \ddots \quad a_{ij} & \vdots \\ 0 & & a_{mm} \end{pmatrix}, \quad j \geq i, \quad a_{ii} = \lambda_i.$$

Si on remplace la base $(e_1, \ldots, e_m)$ par la base

$$(\widetilde{e}_j) = (e_1, \varepsilon e_2, \varepsilon^2 e_3, \ldots, \varepsilon^{m-1} e_m)$$

avec $\varepsilon > 0$ petit, on voit que le coefficient a_{ij} de A est remplacé par $\varepsilon^{j-i} a_{ij}$. Pour les coefficients au-dessus de la diagonale on a $j > i$, donc ε^{j-i} est petit. Dans une base convenable $(\widetilde{e}_1, \widetilde{e}_2, \ldots, \widetilde{e}_m)$ de $\mathbb{C}^m$, la matrice A se transforme donc en une matrice

$$\widetilde{A} = D + T$$

où D est diagonale de valeurs propres $\lambda_1, \ldots, \lambda_m$ et T strictement triangulaire supérieure avec des coefficients $O(\varepsilon)$ arbitrairement petits. Soit N_h la norme

hermitienne sur $\mathbb{C}^m$ ayant $(\widetilde{e}_1, \ldots, \widetilde{e}_m)$ pour base orthonormée et N_e la norme euclidienne induite sur $\mathbb{R}^m$ (restriction de N_h à $\mathbb{R}^m$). On a alors

$$|||A|||_{N_e} \leq |||\widetilde{A}|||_{N_h} \leq |||D|||_{N_h} + |||T|||_{N_h}.$$

Comme $|||D|||_{N_h} = \rho(A)$ et comme $|||T|||_{N_h} = O(\varepsilon)$ peut être rendue arbitrairement petite, il vient

$$\inf_{N \in \mathcal{N}_e} |||A|||_{N_e} \leq \rho(A).$$

(3) On a d'une part $\rho(A) = \rho(A^p)^{1/p} \leq |||A^p|||_N^{1/p}$.
Inversement, étant donné $\varepsilon > 0$, on peut choisir grâce à (2) une norme euclidienne $N_\varepsilon \in \mathcal{N}_e$ telle que $|||A|||_{N_\varepsilon} \leq \rho(A) + \varepsilon$. Comme toutes les normes sur l'espace de dimension finie des matrices carrées $m \times m$ sont équivalentes, il existe une constante $C_\varepsilon \geq 1$ telle que $|||B|||_N \leq C_\varepsilon |||B|||_{N_e}$ pour toute matrice B. Pour $B = A^p$, on en déduit

$$\begin{aligned} |||A^p|||_N &\leq C_\varepsilon |||A^p|||_{N_\varepsilon} \leq C_\varepsilon |||A|||_{N_\varepsilon}^p \leq C_\varepsilon (\rho(A) + \varepsilon)^p, \\ |||A^p|||_N^{1/p} &\leq C_\varepsilon^{1/p} (\rho(A) + \varepsilon). \end{aligned}$$

Comme $\lim_{p \to +\infty} C_\varepsilon^{1/p} = 1$, il existe $p_\varepsilon \in \mathbb{N}$ tel que

$$p \geq p_\varepsilon \to |||A^p|||_N^{1/p} \leq \rho(A) + 2\varepsilon,$$

et le théorème est démontré. ■

3.2. Critère d'attractivité

Soit Ω un ouvert de $\mathbb{R}^m$ et $\varphi : \Omega \to \mathbb{R}^m$ une fonction de classe C^1. Soit $N = \| \ \|$ une norme fixée sur $\mathbb{R}^m$. On note $\varphi'(x) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$ l'application linéaire tangente au point $x \in \Omega$, de sorte que

$$\varphi(x+h) - \varphi(x) = \varphi'(x) \cdot h + \|h\| \varepsilon(h), \quad \lim_{h \to 0} \varepsilon(h) = 0.$$

Lemme

(1) *Si φ est k-lipschizienne sur Ω relativement à la norme N, alors $|||\varphi'(x)|||_N \leq k$ pour tout $x \in \Omega$.*

(2) *Si Ω est convexe et si $|||\varphi'(x)|||_N \leq k$ pour tout $x \in \Omega$, alors φ est k-lipschitzienne sur Ω relativement à N.*

Démonstration

(1) Pour tout $\delta > 0$, il existe $r > 0$ tel que $\|h\| \leq r \to \|\varepsilon(h)\| \leq \delta$. Par linéarité de $\varphi'(x)$, on a

$$|||\varphi'(x)|||_N = \sup_{\|h\| = r} \frac{\|\varphi'(x) \cdot h\|}{\|h\|} \ ;$$

or $\varphi'(x)\cdot h = \varphi(x+h) - \varphi(x) - \|h\|\varepsilon(h)$,

$$\begin{aligned}\|\varphi'(x)\cdot h\| &\leq \|\varphi(x+h)-\varphi(x)\| + \|h\|\,\|\varepsilon(h)\| \\ &\leq k\|h\| + \|h\|\,\|\varepsilon(h)\| \leq (k+\delta)\|h\|.\end{aligned}$$

Il vient donc $|||\varphi'(x)|||_N \leq k+\delta$, et ce quel que soit $\delta > 0$.

(2) Inversement, si Ω est convexe, on peut écrire

$$\varphi(y) - \varphi(x) = \psi(1) - \psi(0) = \int_0^1 \psi'(t)dt \qquad \text{avec}$$
$$\psi(t) = \varphi(x+t(y-x)), \quad \psi'(t) = \varphi'(x+t(y-x))\cdot(y-x).$$

On en déduit

$$\varphi(y) - \varphi(x) = \int_0^1 \varphi'(x+t(y-x))\cdot(y-x)dt,$$
$$\|\varphi(y)-\varphi(x)\| \leq \int_0^1 |||\varphi'(x+t(y-x))|||_N \|y-x\|dt \leq k\|y-x\|. \qquad \blacksquare$$

Théorème – *Soit $a \in \Omega$ un point fixe de φ. Alors les deux propriétés suivantes sont équivalentes :*

(i) *Il existe un voisinage fermé V de a tel que $\varphi(V) \subset V$ et une norme N sur $\mathbb{R}^n$ telle que $\varphi|_V$ soit contractante pour N.*

(ii) *$\rho(\varphi'(a)) < 1$.*

On dit alors que le point fixe a est attractif.

Démonstration. Si $\varphi|_V$ est contractante de rapport $k < 1$ alors d'après la partie (1) du lemme on a

$$\rho(\varphi'(a)) \leq |||\varphi'(a)|||_N \leq k < 1.$$

Inversement, si $\rho(\varphi'(a)) < 1$, il existe une norme euclidienne N telle que $|||\varphi'(a)|||_N < 1$. Par continuité de φ', il existe une boule fermée $V = \overline{B}(a,r)$, $r > 0$, telle que $\sup_V |||\varphi'|||_N = k < 1$. Comme V est convexe, φ est alors contractante de rapport k sur V ; en particulier $\varphi(V) \subset \overline{B}(a,kr) \subset V$. $\blacksquare$

Remarque – Si φ est de classe C^2 et si $\varphi'(a) = 0$, la formule de Taylor montre qu'il existe une constante $M \geq 0$ telle que

$$\|\varphi(x) - a\| \leq M\|x-a\|^2, \quad x \in \overline{B}(a,r).$$

Le phénomène de convergence quadratique a donc encore lieu ici.

3.3. MÉTHODE DE NEWTON-RAPHSON

Soit à résoudre une équation $f(x) = 0$ où $f : \Omega \to \mathbb{R}^m$ est une application de classe C^2 définie sur un ouvert $\Omega \subset \mathbb{R}^m$. On cherche à évaluer numériquement une solution a du système $f(x) = 0$, connaissant une valeur approchée grossière x_0 de a. Comme dans la méthode de Newton usuelle, l'idée est d'approximer f par sa partie linéaire au point x_0 :

$$f(x) = f(x_0) + f'(x_0) \cdot (x - x_0) + o(\|x - x_0\|)$$

On résout alors l'équation $f(x_0) + f'(x_0) \cdot (x - x_0) = 0$. Si $f'(x_0) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ est inversible, on a une solution unique x_1 telle que $x_1 - x_0 = -f'(x_0)^{-1} \cdot f(x_0)$, soit

$$x_1 = x_0 - f'(x_0)^{-1} \cdot f(x_0).$$

On va donc itérer ici l'application de classe C^1

$$\varphi(x) = x - f'(x)^{-1} \cdot f(x).$$

Théorème – *On suppose que f est de classe C^2, que $f(a) = 0$ et que l'application linéaire tangente $f'(a) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$ est inversible. Alors a est un point fixe superattractif de φ.*

Démonstration. Calculons un développement limité à l'ordre 2 de $\varphi(a+h)$ quand h tend vers 0. On a

$$\begin{aligned}
f(a+h) &= f'(a) \cdot h + \frac{1}{2} f''(a) \cdot (h)^2 + o(\|h\|^2) \\
&= f'(a) \cdot \Big[h + \frac{1}{2} f'(a)^{-1} \cdot (f''(a) \cdot (h)^2) + o(\|h\|^2) \Big]. \\
f'(a+h) &= f'(a) + f''(a) \cdot h + o(\|h\|) \\
&= f'(a) \circ \Big[\mathrm{Id} + f'(a)^{-1} \circ (f''(a) \cdot h) + o(\|h\|) \Big], \\
f'(a+h)^{-1} &= [\ \ ]^{-1} \circ f'(a)^{-1} \\
&= \Big[\mathrm{Id} - f'(a)^{-1} \circ (f''(a) \cdot h) + o(\|h\| \Big] \circ f'(a)^{-1},
\end{aligned}$$

$$\begin{aligned}
&f'(a+h)^{-1} \cdot f(a+h) \\
&\quad = \Big[\mathrm{Id} - f'(a)^{-1} \circ (f''(a) \cdot h) + o(\|h\|) \Big] \cdot \Big[h + \frac{1}{2} f'(a)^{-1} \cdot (f''(a) \cdot (h)^2) + o(\|h\|) \Big] \\
&\quad = h - \frac{1}{2} f'(a)^{-1} \cdot (f''(a) \cdot (h)^2) + o(\|h\|)^2,
\end{aligned}$$

d'où finalement

$$\begin{aligned}
\varphi(a+h) &= a + h - f'(a+h)^{-1} \cdot f(a+h) \\
&= a + \frac{1}{2} f'(a)^{-1} \cdot (f''(a) \cdot (h)^2) + o(\|h\|^2).
\end{aligned}$$

On en déduit $\varphi'(a) = 0$ et $\varphi''(a) = f'(a)^{-1} \circ f''(a)$. En particulier

$$\|\varphi(a+h) - a\| \leq \frac{1}{2}\,(M + \varepsilon(h))\|h\|^2$$

où $M = |||\varphi''(a)|||$. Le théorème est démontré. ■

Exemple – Soit à résoudre le système

$$\begin{cases} x^2 + xy - 2y^2 = 4 \\ xe^x + ye^y = 0. \end{cases}$$

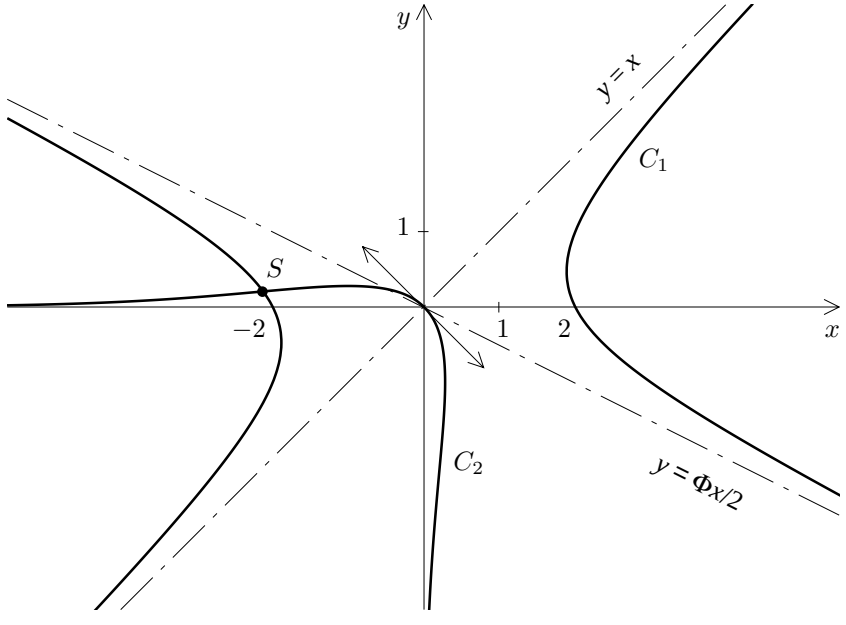
On commence par tracer les courbes $C_1 : x^2 + xy - 2y^2 = 4$ et $C_2 : xe^x + ye^y = 0$ de manière à obtenir graphiquement une approximation grossière des solutions.

C_1 est une hyperbole d'asymptotes $x^2 + xy - 2y^2 = (x-y)(x+2y) = 0$; cette hyperbole passe par les points $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$ et $\begin{pmatrix} -2 \\ 0 \end{pmatrix}$.

La courbe C_2 est symétrique par rapport à la droite $y = x$; de plus x, y sont nécessairement de signe opposés. Supposons par exemple $x \leq 0$, $y \geq 0$. Comme la fonction $y \mapsto ye^y$ est strictement croissante de $[0, +\infty[$ *sur* $[0, +\infty[$, à chaque point $x \leq 0$ correspond un unique point $y \geq 0$. Ce point y est la solution de $xe^x + ye^y = 0$, et peut par exemple s'obtenir par itération de la fonction

$$\varphi(x) = y - \frac{xe^x + ye^y}{(1+y)e^y} = \frac{y^2 - xe^{x-y}}{1+y},$$

fournie par la méthode de Newton appliquée à la variable y. Pour $x = 0$, on a $y = 0$; en décrémentant x par pas de $0,1$ avec comme valeur initiale de y_0 la valeur de la solution y trouvée pour la valeur x précédente, on obtient la courbe C_2.



On voit que le système précédent admet une solution $S\begin{pmatrix} a \\ b \end{pmatrix}$ unique, avec $\begin{pmatrix} a \\ b \end{pmatrix} \simeq \begin{pmatrix} -2 \\ 0,2 \end{pmatrix}$ très grossièrement. Pour obtenir une valeur approchée plus précise, on cherche à résoudre l'équation $f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ avec

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^2 + xy - 2y^2 - 4 \\ xe^x + ye^y \end{pmatrix}.$$

L'application linéaire tangente à f est donnée par

$$f'\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{pmatrix} = \begin{pmatrix} 2x + y & x - 4y \\ (x+1)e^x & (y+1)e^y \end{pmatrix}$$

La condition $f'(S)$ inversible signifie que les tangentes

$$\frac{\partial f_i}{\partial x}(S)(x-a) + \frac{\partial f_i}{\partial y}(S)(y-b) = 0, \quad i = 1, 2,$$

aux courbes C_1, C_2 au point S sont distinctes. C'est bien le cas ici. On obtient

$$\left[f'\begin{pmatrix} x \\ y \end{pmatrix}\right]^{-1} = \frac{1}{\Delta(x,y)} \begin{pmatrix} (x+1)e^y & -x + 4y \\ -(x+1)e^x & 2x + y \end{pmatrix}$$

avec $\Delta(x,y) = (2x+y)(y+1)e^y - (x-4y)(x+1)e^x$. On est alors amené à calculer les itérés $\begin{pmatrix} x_{p+1} \\ y_{p+1} \end{pmatrix} = \varphi\begin{pmatrix} x_p \\ y_p \end{pmatrix}$ avec

$$\varphi\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{\Delta(x,y)} \begin{pmatrix} (y+1)e^y & -x + 4y \\ -(x+1)e^x & 2x + y \end{pmatrix} \begin{pmatrix} x^2 + xy - 2y^2 - 4 \\ xe^x + ye^y \end{pmatrix}$$

Partant du point initial $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} -2 \\ 0,2 \end{pmatrix}$, on trouve

p	x_p	y_p
0	-2	$0,2$
1	$-2,130690999$	$0,205937784$
2	$-2,126935837$	$0,206277868$
3	$-2,126932304$	$0,206278156$
4	$-2,126932304$	$0,206278156$

d'où $S = \begin{pmatrix} a \\ b \end{pmatrix} \simeq \begin{pmatrix} -2,126932304 \\ 0,206278156 \end{pmatrix}$.

4. Le théorème des fonctions implicites

Nous allons ici exploiter le théorème du point fixe pour démontrer quelques résultats fondamentaux du calcul différentiel. Notre objectif est d'obtenir aussi des estimations quantitatives pour ces théorèmes, parce que ces estimations sont souvent nécessaires pour majorer les erreurs commises dans les calculs numériques.

4.1. Le théorème d'inversion locale

Nous commençons par un lemme de perturbation, qui s'applique dans un espace numérique $\mathbb{R}^m$ muni d'une norme N quelconque.

Lemme – *Soit $f : B(x_0, r) \to \mathbb{R}^m$ une application définie sur une boule de rayon r dans $\mathbb{R}^m$, telle que*

$$f(x) = x + u(x)$$

où u est une application contractante de rapport $k < 1$ (« petite perturbation de l'identité »). Alors

(1) *f est un homéomorphisme de $B(x_0, r)$ sur un ouvert $V = f(B(x_0, r))$ de $\mathbb{R}^m$;*

(2) *f est $(1+k)$-lipschitzienne et son application réciproque $f^{-1} : V \to B(x_0, r)$ est $(1-k)^{-1}$ lipschitzienne ;*

(3) *l'image V satisfait l'encadrement*

$$B(f(x_0), (1-k)r) \subset V \subset B(f(x_0), (1+k)r).$$

Démonstration. Quitte à remplacer f par $\widetilde{f}(x) = f(x + x_0) - f(x_0)$ et u par $\widetilde{u}(x) = u(x + x_0) - u(x_0)$, on peut supposer que $x_0 = 0$ et $f(0) = u(0) = 0$. On a de façon évidente

$$\|x_2 - x_1\| - \|u(x_2) - u(x_1)\| \leq \|f(x_2) - f(x_1)\| \leq \|x_2 - x_1\| + \|u(x_2) - u(x_1)\|,$$
$$(1-k)\|x_2 - x_1\| \leq \|f(x_2) - f(x_1)\| \leq (1+k)\|x_2 - x_1\|.$$

Il en résulte que f est injective et $(1+k)$-lipschitzienne, et que $\|f(x)\| \leq (1+k)\|x\|$. Par suite f est une bijection de $B(0, r)$ sur son image V, et

$$V = f(B(0, r)) \subset B(0, (1+k)r).$$

On considère maintenant un rayon $r' < r$ quelconque, et pour $y \in \mathbb{R}^m$ donné, on introduit l'application

$$\varphi(x) = y + x - f(x) = y - u(x).$$

De même que u, c'est une application k-lipschitzienne, et comme

$$\|\varphi(x)\| \leq \|y\| + k\|x\|,$$

on voit que φ envoie $\overline{B}(0, r')$ dans $\overline{B}(0, r')$ dès lors que $\|y\| \leq (1-k)r'$. Le théorème du point fixe appliqué à l'espace complet $E = \overline{B}(0, r')$ implique que φ possède un

point fixe $\varphi(x) = x$ unique, c'est-à-dire que l'équation $f(x) = y$ détermine un unique antécédent $x \in \overline{B}(0, r')$. On en déduit que $f(\overline{B}(0, r')) \supset \overline{B}(0, (1-k)r')$. Comme $r' < r$ est arbitraire, on a bien

$$V = f(B(0, r)) \supset B(0, (1-k)r).$$

Ceci démontre déjà (3). En se plaçant en un point $x_1 \in B(x_0, r)$ quelconque et en remplaçant r par $\varepsilon > 0$ petit, on voit que l'image $f(B(x_1, \varepsilon))$ contient un voisinage $B(f(x_1), (1-k)\varepsilon)$ de $f(x_1)$, par suite $V = f(B(x_0, r))$ est un ouvert. Enfin, l'inégalité de gauche dans l'encadrement de Lipschitz de f implique

$$(1-k)\|f^{-1}(y_2) - f^{-1}(y_1)\| \leq \|y_2 - y_1\|$$

pour tous $y_1, y_2 \in V$, ce qui entraîne que f^{-1} est continue $(1-k)^{-1}$-lipschitzienne et que f est un homéomorphisme de $B(x_0, r)$ sur V. Le lemme est démontré. ■

Théorème d'inversion locale – *Soit $f : \Omega \to \mathbb{R}^m$ une application de classe C^k définie sur un ouvert Ω de $\mathbb{R}^m$ avec $k \geq 1$. Étant donné un point $x_0 \in \Omega$, on suppose que l'application linéaire tangente $\ell = f'(x_0) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$ est inversible. Alors*

(1) *Il existe un voisinage $U = B(x_0, r)$ de x_0 tel que $V = f(U)$ soit un ouvert de $\mathbb{R}^m$ et f un difféomorphisme de classe C^k de U sur V.*

(2) *Pour tout $y = f(x)$, $x \in U$, la différentielle de $g = f^{-1}$ est donnée par $g'(y) = (f'(x))^{-1}$, soit $g'(y) = (f'(g(y)))^{-1}$ ou encore*

$$(f^{-1})'(y) = (f'(f^{-1}(y)))^{-1} \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m).$$

(3) *On suppose $k \geq 2$. Si $B(x_0, r_0) \subset \Omega$ et si M est un majorant de f'' sur $B(x_0, r_0)$, on peut choisir $U = B(x_0, r)$ avec $r = \min(r_0, \frac{1}{2}M^{-1}|||\ell^{-1}|||^{-1})$.*

Démonstration. L'application $u(x) = \ell^{-1}(f(x)) - x$ est de classe C^1 et on a $u'(x_0) = \ell^{-1} \circ f'(x_0) - \mathrm{Id} = 0$ dans $\mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$. Par continuité de u', il existe une boule $B(x_0, r)$ sur laquelle $|||u'(x)||| \leq k < 1$, ce qui entraîne que u est contractante de rapport k. Le lemme précédent montre alors que $\widetilde{f}(x) = \ell^{-1} \circ f(x) = x + u(x)$ définit un homéomorphisme lipschitzien de $U = B(x_0, r)$ sur $\widetilde{V} = \widetilde{f}(V)$, donc $f = \ell \circ \widetilde{f}$ est un homéomorphisme lipschitzien de U sur $V = \ell(\widetilde{V})$, la constante de Lipschitz de f étant majorée par $K = (1+k)|||l|||$ et celle de $g = f^{-1} = \widetilde{f}^{-1} \circ \ell^{-1}$ par $K' = (1-k)^{-1}|||\ell^{-1}|||$.

Maintenant, si f est de classe C^2 et $|||f''||| \leq M$ sur la boule $B(x_0, r_0) \subset \Omega$, nous avons $u'' = \ell^{-1} \circ f''$, donc $|||u''||| \leq M|||\ell^{-1}|||$ et le théorème des accroissements finis nous donne $|||u'(x)||| \leq M|||\ell^{-1}|||\|x - x_0\|$ sur $B(x_0, r_0)$. Pour $r = \min(r_0, \frac{1}{2}M^{-1}|||\ell^{-1}|||^{-1})$, on voit que $|||u'||| \leq k = \frac{1}{2}$ sur $B(x_0, r)$ et on peut appliquer ce qui précède.

Pour tous $y, \eta \in V$ et $x = g(y)$, $\xi = g(x) \in U = B(x_0, r)$, l'hypothèse de différentiablilité de f au point x implique

$$\eta - y = f(\xi) - f(x) = f'(x)(\xi - x) + \varepsilon(\xi - x)\|\xi - x\|$$

avec $\lim_{h\to 0}\varepsilon(h)=0$. Comme $\ell^{-1}\circ f'(x)=\mathrm{Id}+u'(x)$ où $|||u'(x)|||\le k<1$ sur $B(x_0,r)$, l'application linéaire $\ell^{-1}\circ f'(x)$ est bien inversible. Par conséquent $f'(x)$ l'est aussi, et nous pouvons écrire

$$\xi-x=f'(x)^{-1}\Big(\eta-y-\varepsilon(\xi-x)\|\xi-x\|\Big),$$

soit

$$g(\eta)-g(y)=f'(x)^{-1}(\eta-y)-f'(x)^{-1}\big(\varepsilon(g(\eta)-g(y))\|g(\eta)-g(y)\|$$

avec $\lim_{\eta\to y}f'(x)^{-1}\big(\varepsilon(g(\eta)-g(y)\big)=0$ et $\|g(\eta)-g(y)\|\le K'\|\eta-y\|$. On voit ainsi que $g=f^{-1}$ est bien différentiable au point y et que $g'(y)=f'(x)^{-1}=f'(g(y))^{-1}$. L'inversion d'une matrice étant une opération continue (et même indéfiniment différentiable), on en déduit que g' est continue sur V, c'est-à-dire que g est de classe C^1.

Si f est de classe C^k, $k\ge 2$, on vérifie par récurrence que g est aussi de classe C^k : f' est de classe C^{k-1}, g l'est aussi par hypothèse de récurrence, donc g' est de classe C^{k-1}, c'est-à-dire que g est de classe C^k. Le théorème est démontré. ∎

4.2.* *Le théorème des fonctions implicites et ses variantes*

Nous allons reformuler le théorème d'inversion locale pour en tirer différentes variantes et différentes conséquences géométriques. La variante la plus importante est le théorème des fonctions implicites.

Théorème des fonctions implicites – *Soit Ω un ouvert de $\mathbb{R}^p\times\mathbb{R}^m$ et $f:\Omega\to\mathbb{R}^m$ une application de classe C^k, $k\ge 1$. On se donne un point $(x_0,y_0)\in\Omega\subset\mathbb{R}^p\times\mathbb{R}^m$, et on suppose que*

(i) $f(x_0,y_0)=0$;

(ii) *la matrice des dérivées partielles* $f'_y(x_0,y_0)=\Big(\dfrac{\partial f_i}{\partial y_j}(x_0,y_0)\Big)_{1\le i,j\le m}$ *est inversible dans* $\mathcal{L}(\mathbb{R}^m,\mathbb{R}^m)$.

Alors il existe un voisinage $U\times V$ de (x_0,y_0) dans Ω sur lequel $f'_y(x,y)$ est inversible, et une application $g:U\to V$ de classe C^k telle que « l'équation implicite » $f(x,y)=0$ pour $(x,y)\in U\times V$ soit équivalente à $y=g(x)$, $x\in U$. La dérivée de g est donnée par la formule

$$g'(x)=-f'_y(x,g(x))^{-1}\circ f'_x(x,g(x)).$$

Autrement dit, le théorème des fonctions implicites dit que l'ensemble des solutions de l'équation $f(x,y)=0$ dans $U\times V$ peut être « explicité » comme le graphe $y=g(x)$ d'une fonction $g:U\to V$ de classe C^k, là où $f'_y(x,y)$ est inversible.

Remarque 1 – On voit immédiatement que le théorème des fonctions implicites contient le théorème d'inversion locale. Il suffit de poser $F(x,y)=x-f(y)$,

$(x,y) \in \widetilde{\Omega} = \mathbb{R}^m \times \Omega \subset \mathbb{R}^m \times \mathbb{R}^m$ pour obtenir l'existence de la fonction réciproque $y = g(x)$ de f au voisinage de tout point $y_0 \in \Omega$ où $f'(y_0)$ est inversible.

Démonstration. En sens inverse, nous allons montrer que le théorème d'inversion locale implique facilement le théorème des fonctions implicites (ce sont donc des théorèmes « équivalents »). Avec les hypothèses faites sur f, considérons

$$F : \Omega \to \mathbb{R}^p \times \mathbb{R}^m, \qquad (x,y) \mapsto F(x,y) = (x, f(x,y)).$$

Nous avons $F(x_0,y_0) = (x_0,0)$ et la matrice de la différentielle de F est donnée par

$$F'(x,y) = \begin{pmatrix} \mathrm{Id} & 0 \\ f'_x(x,y) & f'_y(x,y) \end{pmatrix}.$$

Ceci permet de voir aussitôt que $F'(x,y) \in \mathcal{L}(\mathbb{R}^{p+m}, \mathbb{R}^{p+m})$ est inversible en tout point où $f'_y(x,y) \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^p)$ l'est, avec

$$F'(x,y)^{-1} = \begin{pmatrix} \mathrm{Id} & 0 \\ -f'_y(x,y)^{-1} \circ f'_x(x,y) & f'_y(x,y)^{-1} \end{pmatrix}.$$

En particulier, $F'(x_0,y_0)$ est inversible par hypothèse.
Le théorème d'inversion locale montre que F est un difféomorphisme de classe C^k d'un voisinage $\widetilde{T} = U_1 \times V_1$ de (x_0,y_0) sur un voisinage $\widetilde{W}$ de $(x_0,0)$ dans $\mathbb{R}^p \times \mathbb{R}^m$ (Quitte à rétrécir $\widetilde{T}$ on peut supposer que $\widetilde{T}$ est un produit de boules ouvertes). L'application $H = F^{-1} : \widetilde{W} \to U_1 \times V_1$, $(u,v) \mapsto (x,y) = H(u,v)$ est la solution de l'équation $F(x,y) = (x, f(x,y)) = (u,v)$ pour $(x,y) \in \widetilde{U}_1 \times V_1$, donc $x = u$ et on voit que H est de la forme $H(u,v) = (u, h(u,v))$ pour une certaine fonction $h : \widetilde{W} \to V_1$. Pour $(x,y) \in U_1 \times V_1$ et $(x,v) \in \widetilde{W}$, nous avons l'équivalence

$$f(x,y) = v \Leftrightarrow y = h(x,v).$$

La fonction $g(x) = h(x,0)$ donne donc précisément $y = g(x)$ comme solution del'équation $f(x,y) = 0$ lorsque $(x,y) \in U_1 \times V_1$ et $(x,0) \in \widetilde{W}$. Définissons U comme l'ensemble des $x \in U_1$ tels que $(x,0) \in \widetilde{W}$ et $V = V_1$. Nous obtenons ainsi U, V qui sont des voisinages ouverts de x_0 et y_0 respectivement, et une application $g : U \to V$ de classe C^k qui répond à la question. De plus $g'(x) = h'_x(x,0)$ est la dérivée partielle en x de la composante h dans $H'(x,0) = F'(H(x,0))^{-1} = F'(x,g(x))^{-1}$, c'est-à-dire

$$g'(x) = -f'_y(x,g(x))^{-1} \circ f'_x(x,g(x)).$$

■

Remarque 2 – D'un point de vue numérique, le calcul de $y = g(x)$ au voisinage de x_0 pourra se faire en utilisant la méthode de Newton-Raphson appliquée à la fonction $y \mapsto f(x,y)$, c'est-à-dire en calculant les itérés successifs $y_{p+1} = y_p - f'_y(x,y_p)^{-1}(f(x,y_p))$ à partir de la valeur approchée y_0.

Nous abordons maintenant des énoncés plus géométriques.

Définition – *Soit Ω un ouvert de $\mathbb{R}^m$ et $f : \Omega \to \mathbb{R}^p$ une application de classe C^k, $k \geq 1$. On dit que*

(1) *f est une immersion en un point $x_0 \in \Omega$ si $f'(x_0) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^p)$ est injective (ce qui implique $m \leq p$) ;*

(2) *f est une submersion en un point $x_0 \in \Omega$ si $f'(x_0) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^p)$ est surjective (ce qui implique $m \geq p$) ;*

(3) *f est de rang constant si le rang de $f'(x)$ pour $x \in \Omega$ est un entier r constant (ce qui implique $r \leq \min(m, p)$).*

La structure locale de telles applications est décrite respectivement par les trois théorèmes suivants.

Théorème des immersions – *Si f est une immersion en un point $x_0 \in \Omega$, il existe un voisinage U de x_0 dans $\mathbb{R}^m$, un voisinage V de $y_0 = f(x_0)$ dans $\mathbb{R}^p$ et un C^k-difféomorphisme $\psi : V \to U \times T$ de V sur un voisinage $U \times T$ de $(x_0, 0)$ dans $\mathbb{R}^p = \mathbb{R}^m \times \mathbb{R}^{p-m}$ tel que $\psi \circ f(x) = (x, 0)$ sur U.*
Autrement dit, au difféomorphisme ψ près dans l'espace d'arrivée, $\widetilde{f} = \psi \circ f$ s'identifie à l'injection triviale $x \mapsto (x, 0)$ au voisinage de x_0.

Théorème des submersions – *Si f est une submersion en un point $x_0 \in \Omega$, il existe un voisinage U de x_0 dans $\mathbb{R}^m$, un voisinage V de $y_0 = f(x_0)$ dans $\mathbb{R}^p$ et un C^k-difféomorphisme $\varphi : U \to V \times S$ de U sur un voisinage $V \times S$ de $(y_0, 0)$ dans $\mathbb{R}^m = \mathbb{R}^p \times \mathbb{R}^{m-p}$ tel que $f \circ \varphi^{-1}(y, s) = y$ sur $V \times S$.*
Autrement dit, au difféomorphisme φ près dans l'espace de départ, $\widetilde{f} = f \circ \varphi^{-1}$ s'identifie à la projection triviale $(y, s) \mapsto y$ au voisinage de $(y_0, 0)$.

Théorème du rang – *Si f est de rang constant r sur Ω et si $x_0 \in \Omega$, il existe un voisinage U de x_0 dans $\mathbb{R}^m$, un voisinage V de $y_0 = f(x_0)$ dans $\mathbb{R}^p$, un C^k-difféomorphisme $\varphi : U \to Q \times S$ de U sur un voisinage $Q \times S$ de 0 dans $\mathbb{R}^m = \mathbb{R}^r \times \mathbb{R}^{m-r}$, et un C^k-difféomorphisme $\psi : V \to Q \times T$ de V sur un voisinage $Q \times T$ de 0 dans $\mathbb{R}^p = \mathbb{R}^p \times \mathbb{R}^{p-r}$, tels que*

$$\psi \circ f \circ \varphi^{-1}(x) = (x_1, \ldots, x_r, 0, \ldots, 0) \in \mathbb{R}^p \quad \text{sur } Q \times S.$$

Autrement dit, aux difféomorphismes φ, ψ près à la fois au départ et à l'arrivée, $\widetilde{f} = \psi \circ f \circ \varphi^{-1}$ s'identifie à l'application linéaire canonique de rang r

$$(*) \qquad (x_1, \ldots, x_r, x_{r+1}, \ldots, x_m) \in \mathbb{R}^m \mapsto (x_1, \ldots, x_r, 0, \ldots, 0) \in \mathbb{R}^p$$

au voisinage de 0, et on a le diagramme commutatif

$$\begin{array}{ccc} U & \xrightarrow{f} & V \\ \simeq \downarrow \varphi & & \simeq \downarrow \psi \\ Q \times S & \xrightarrow[\widetilde{f}]{} & Q \times T \end{array}$$

où $\widetilde{f}$ est l'application linéaire $(*)$.

Démonstration du théorème des immersions. Choisissons des vecteurs linéairement indépendants $(a_1, \dots a_{p-m})$ de $\mathbb{R}^p$ de sorte qu'on ait une décomposition en somme directe

$$\mathbb{R}^p = \operatorname{Im} f'(x_0) \oplus \bigoplus_{1 \le i \le p-m} \mathbb{R} a_i.$$

C'est possible puisque $\dim \operatorname{Im} f'(x_0) = m$ d'après l'injectivité de $f'(x_0)$. Nous définissons

$$\Psi : \Omega \times \mathbb{R}^{p-m} \to \mathbb{R}^p, \qquad \Psi(x,t) = f(x) + \sum_{1 \le i \le p-m} t_i a_i.$$

Comme $\partial \Psi(x,t)/\partial t_i = a_i$, il est clair que $\operatorname{Im} \Psi'(x_0, 0)$ contient à la fois $\operatorname{Im} f'(x_0)$ et les a_i, donc $\Psi'(x_0, 0)$ est surjective. Mais comme Ψ est une application de $\mathbb{R}^p$ dans $\mathbb{R}^p$, ceci entraîne que $\Psi'(x_0, 0)$ est inversible. Par conséquent Ψ définit un C^k-difféomorphisme d'un voisinage $U \times T$ de $(x_0, 0)$ sur un voisinage V de $y_0 = f(x_0)$. Nous avons $\Psi(x, 0) = f(x)$ par définition de Ψ, donc $\psi = \Psi^{-1}$ répond à la question. ∎

Démonstration du théorème des submersions. Soit $(a_1, \dots, a_m)$ une base de $\mathbb{R}^m$ choisie de telle sorte que $(a_{p+1}, \dots, a_m)$ soit une base de $\operatorname{Ker} f'(x_0)$. Nous définissons

$$\varphi : \Omega \to \mathbb{R}^p \times \mathbb{R}^{m-p}, \qquad \varphi(x) = (f(x), s_{p+1}, \dots, s_m)$$

où $(s_1, \dots, s_m)$ désignent les coordonnées de x dans la base (a_i), c'est-à-dire $x = \sum s_i a_i$. Le noyau $\operatorname{Ker} \varphi'(x_0)$ est l'intersection du noyau $\operatorname{Ker} f'(x_0)$ avec le sous-espace $s_{p+1} = \dots = s_m = 0$ engendré par $(a_1, \dots, a_p)$, et comme ces espaces sont supplémentaires on a $\operatorname{Ker} \varphi'(x_0) = \{0\}$. Ceci montre que $\varphi'(x_0)$ est injective, et donc bijective. Par conséquent φ est un C^k-difféomorphisme d'un voisinage U de x_0 sur un voisinage $V \times S$ de $(y_0, 0) = (f(x_0), 0)$ (quitte à rétrécir ces voisinages, on peut toujours supposer que le voisinage d'arrivée est un produit). On voit alors que $f = \pi \circ \varphi$ où π est la projection sur les p-premières coordonnées, de sorte que $f \circ \varphi^{-1} = \pi : (y, s_{p+1}, \dots, s_m) \mapsto y$. Le théorème est démontré. ∎

Démonstration du théorème du rang. On construit des difféomorphismes φ_i entre ouverts de $\mathbb{R}^m$ et ψ_i entre ouverts de $\mathbb{R}^p$ de façon à « simplifier » progressivement f :

$$\begin{array}{ccc} U & \xrightarrow{f} & V \\ \downarrow \varphi_1 & & \downarrow \psi_1 \\ U_1 & \xrightarrow{f_1} & V_1 \\ \downarrow \varphi_2 & & \downarrow \psi_2 \\ U_2 & \xrightarrow{f_2} & V_2. \end{array}$$

Le premier niveau (φ_1, ψ_1) consiste simplement en des changements affines de coordonnées : on choisit respectivement x_0 et $y_0 = f(x_0)$ comme nouvelles origines

dans $\mathbb{R}^m$ et $\mathbb{R}^p$, ainsi que de nouvelles bases $(a_1,\dots,a_m)$ et $(b_1,\dots,b_p)$ de sorte que $(a_{r+1},\dots,a_m)$ soit une base du noyau de $\operatorname{Ker} f'(x_0)$ et $b_j = f'(x_0)(a_j)$, $1 \le j \le r$. Dans ces nouveaux repères, la matrice de $f'(x_0)$ devient par construction la matrice de rang r

$$\begin{pmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \dots & 1 & & \dots & \\ 0 & \dots & & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & & 0 & \dots & 0 \end{pmatrix}.$$

Ceci est précisément la matrice de la dérivée $f_1'(0)$ de $f_1 = \psi_1 \circ f \circ \varphi_1^{-1}$ à l'origine. En particulier, si $\pi : \mathbb{R}^p \to \mathbb{R}^r$ est la projection sur les r-premières coordonnées, alors $\pi \circ f_1$ est une submersion de $\mathbb{R}^m$ dans $\mathbb{R}^r$ en 0, et d'après le théorème des submersions on peut trouver un C^k-difféomorphisme φ_2 de $\mathbb{R}^m$ à l'origine tel que

$$\pi \circ f_1 \circ \varphi_2^{-1}(x_1,\dots,x_m) = (x_1,\dots,x_r)$$

près de 0. Par conséquent nous avons une écriture

$$f_1 \circ \varphi_2^{-1}(x_1,\dots,x_m) = (x_1,\dots,x_r,h_{r+1}(x),\dots,h_p(x))$$

au voisinage de 0. Or, par hypothèse, $f_1 \circ \varphi_2^{-1} = \psi_1 \circ f \circ \varphi_1^{-1} \circ \varphi_2^{-1}$ est une application de rang constant r, et la matrice de son application linéaire tangente

$$\begin{pmatrix} 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \dots & 1 & & \dots & \\ 0 & \dots & & \partial h_{r+1}/\partial x_{r+1} & \dots & \partial h_{r+1}/\partial x_m \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & & \partial h_p/\partial x_{r+1} & \dots & \partial h_p/\partial x_m \end{pmatrix}$$

ne peut être de rang r que si $\partial h_i/\partial x_j = 0$ pour $i,j > r$, ce qui signifie que $h_{r+1},\dots,h_p$ sont en fait des fonctions des seules variables $x_1,\dots,x_r$. On a donc une application de rang constant r

$$(x_1,\dots,x_r) \mapsto (x_1,\dots,x_r,h_{r+1}(x_1,\dots,x_r),\dots,h_p(x_1,\dots,x_r))$$

au voisinage de 0. En d'autres termes, c'est une immersion de $\mathbb{R}^r$ dans $\mathbb{R}^p$ en 0, et d'après le théorème des immersions il existe un C^k-difféomorphisme ψ_2 de $\mathbb{R}^p$ en 0 tel que

$$\psi_2(x_1,\dots,x_r,h_{r+1}(x_1,\dots,x_r),\dots,h_p(x_1,\dots,x_r)) = (x_1,\dots,x_r,0,\dots,0)$$

près de 0. Il s'ensuit que

$$\psi_2 \circ f_1 \circ \varphi_2^{-1}(x_1,\dots,x_m) = (x_1,\dots,x_r,0,\dots,0)$$

au voisinage de 0. L'existence de petits voisinages $U_2 = Q \times S$ et $V_2 = Q \times T$ est alors immédiate, et le théorème du rang constant est démontré avec $\varphi = \varphi_2 \circ \varphi_1$, $\psi = \psi_2 \circ \psi_1$, $U = \varphi^{-1}(U_2)$, $V = \psi^{-1}(V_2)$. ∎

5. PROBLÈMES

5.1. Soit (E, d) un espace métrique et $\varphi : E \to E$ une application continue telle que l'itérée $\varphi^m = \varphi \circ \ldots \circ \varphi$ soit contractante, avec constante de Lipschitz $k \in]0, 1[$ et $m \in \mathbb{N}^\star$.

(a) On convient de noter $\varphi^0 = \mathrm{Id}_E$. Vérifier que la formule

$$d'(x, y) = \max_{0 \leq i \leq m-1} k^{-i/m} d(\varphi^i(x), \varphi^i(y))$$

définit une distance sur E, topologiquement équivalente à la distance d (c'est-à-dire que les ouverts pour d et d' sont les mêmes). Montrer que (E, d') est complet dès que (E, d) est complet.

(b) Montrer que φ est lipschitzienne de rapport $k^{1/m}$ pour d'. En déduire que φ admet un point fixe unique a, et que, pour tout point initial $x_0 \in E$, il existe une constante C telle que la suite des itérés $x_p = \varphi(x_{p-1})$ vérifie $d(x_p, a) \leq C\, k^{p/m}$.

5.2. On considère la fonction f telle que

$$f(x) = x \ln(x), \quad x \in [1, +\infty[.$$

On se propose d'étudier des algorithmes itératifs permettant de calculer l'image réciproque $f^{-1}(a)$ pour un réel $a \in [0, +\infty[$ fixé quelconque.

(a) Montrer que f est une bijection de $[1, +\infty[$ sur $[0, +\infty[$.

(b) On pose $\varphi(x) = \frac{a}{\ln(x)}$. Pour $a = e$, calculer explicitement $f^{-1}(a)$.

Pour quelles valeurs de $a \neq e$ le procédé itératif $x_{p+1} = \varphi(x_p)$ converge-t-il lorsque la valeur initiale x_0 est choisie assez voisine de $f^{-1}(a)$?

(c) Soit $x_{p+1} = \psi(x_p)$ l'algorithme itératif fourni par la méthode de Newton pour la résolution de l'équation $x \ln(x) - a = 0$.

(α) Étudier les variations de la fonction ψ et tracer sommairement la courbe représentative de ψ.

(β) Étudier la convergence de la suite (x_p) pour $a \in [0, +\infty[$ et $x_0 \in [1, +\infty[$ quelconques.

(γ) Évaluer $f^{-1}(2)$ par ce procédé à l'aide d'une calculette. On donnera les approximations successives obtenues.

5.3. On considère la fonction

$$f(x) = \exp(\exp(-\cos(\sin(x + e^x)))) + x^3.$$

On donne $f(0, 1) \simeq 1,737$; $f(0, 2) \simeq 1,789$ à 10^{-3} près. Écrire un programme permettant de résoudre l'équation $f(x) = 7/4$ à la précision $\varepsilon = 10^{-10}$, ceci au

moyen d'une méthode itérative adaptée (qui permet d'éviter des calculs formels trop compliqués). La justification de la convergence n'est pas demandée.

5.4. On se propose d'étudier le comportement des itérés d'une fonction au voisinage d'un point fixe, dans le cas critique où la dérivée vaut 1 en ce point.

Soit $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ une fonction de classe C^1. On suppose que $\varphi(0) = 0$, $\varphi'(0) = 1$, et que φ admet un développement limité

$$\varphi(x) = x - ax^k + x^k\varepsilon(x)$$

avec

$$a > 0, \quad k > 1, \quad \lim_{x \to 0_+} \varepsilon(x) = 0.$$

(a) Montrer qu'il existe $h > 0$ tel que pour tout $x_0 \in]0, h]$ la suite itérée $x_{p+1} = \varphi(x_p)$ converge vers 0.

(b) On pose $u_p = x_p^m$ où $m \in \mathbb{R}$. Déterminer un équivalent de $u_{p+1} - u_p$ en fonction de x_p.

(c) Montrer qu'il existe une valeur de m pour laquelle $u_{p+1} - u_p$ possède une limite finie non nulle. En déduire un équivalent de x_p.

(d) Pour $\varphi(x) = \sin x$ et $x_0 = 1$, estimer le nombre d'itérations nécessaires pour atteindre $x_p < 10^{-5}$.

5.5. Dans tout ce problème, on travaille sur un intervalle $[a, b]$ fixé.

(a) Soit $g : [a, b] \to \mathbb{R}$ une fonction de classe C^2 telle que $g(a) = g(b) = 0$, $g''(x) > 0$ pour tout x dans $]a, b[$. Démontrer

- que $g(x)$ ne s'annule en aucun x de $]a, b[$,
- puis que $g(x) < 0$ pour tout x dans $]a, b[$.

[Raisonner par l'absurde et utiliser le théorème de Rolle.]

(b) Soit $f : [a, b] \to \mathbb{R}$ une fonction de classe C^2 telle que $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$ et $f''(x) > 0$ pour tout x dans $]a, b[$.

Démontrer

(α) qu'il existe c (unique) dans $]a, b[$ tel que $f(c) = 0$;

(β) qu'il existe m_1 et m_2 tels que

$$0 < m_1 \leq f'(x), \quad 0 < f''(x) \leq m_2 \quad \text{pour tout } x \text{ dans } [a, b[.$$

(c) On conserve désormais les hypothèses de la question (b), et on se propose de « calculer » c. Soit p le polynôme de degré 1 tel que $p(a) = f(a)$, $p(b) = f(b)$, et soit c_1 dans $]a, b[$ tel que $p(c_1) = 0$.

(α) Démontrer que $a < c_1 < c$ [appliquer la question (a) à $g(x) = f(x) - p(x)$].

(β) Établir la majoration

$$|f(c_1)| \leq \frac{1}{2}\, m_2 |(c_1 - a)(c_1 - b)|.$$

(d) Soit (c_n), $n \geq 0$, la suite récurrente définie de la façon suivante :

• on pose $c_0 = a$;

• pour tout $n \geq 0$ (et c_n étant déjà définie) on note p_n l'unique polynôme de degré 1 tel que $p_n(c_n) = f(c_n)$, $p_n(b) = f(b)$; et on définit c_{n+1} par $p_n(c_{n+1}) = 0$.

(α) Mettre explicitement cette récurrence sous la forme $c_{n+1} = \varphi(c_n)$.

(β) Démontrer que (c_n) est une suite strictement croissante contenue dans l'intervalle $[a, c]$.

(γ) Démontrer que la suite (c_n) converge vers c, et que, pour tout $n \geq 0$, on a

$$|c_n - c| \leq \frac{f(c_n)}{m_1}.$$

(e) *Programmation d'un exemple.* On pose $f(x) = x^4 + x - 1$.

Démontrer que l'équation $f(x) = 0$ admet une racine c et une seule dans l'intervalle $[0, 1]$. Écrire un programme permettant de calculer c à 10^{-8} près.

5.6. On considère l'application $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$ définie par

$$\varphi \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix}, \quad \begin{cases} X = -x + \dfrac{3}{2}y + \dfrac{5}{4} \\ Y = -\dfrac{1}{2}x + y^2 + \dfrac{3}{4} \end{cases}$$

(a) Déterminer les points fixes de φ.

(b) Ces points fixes sont-ils attractifs ?

(c) Soit B le point fixe non attractif de φ. Montrer que φ admet une application réciproque $\psi : V \to W$ de classe C^∞, où V, W sont des voisinages de B. Le point B est-il attractif pour ψ ?

5.7. On cherche à résoudre numériquement le système d'équations

$$\text{(S)} \quad \begin{cases} y \ln y - x \ln x = 3 \\ x^4 + xy + y^3 = a \end{cases}$$

où $x, y > 0$ et où a est un paramètre réel.

(a) Étudier sommairement les variations de la fonction $x \mapsto x \ln x$ et montrer que pour $a \geq 31$ le système (S) n'a pas de solution (x, y) telle que $x < 1$. Montrer

que pour $x \geq 1$ la solution y de la première équation est fonction croissante de x et en déduire que le système (S) admet une solution (x, y) unique pour $a \geq 31$.

(b) Montrer que la solution (x, y) est telle que

$$y = x + \frac{3}{1 + \ln c} \quad \text{øù} \quad c \in]x, y[.$$

En déduire un équivalent de x et y en fonction de a quand a tend vers $+\infty$. Pouvez-vous raffiner cet équivalent et donner un développement plus précis ?

(c) Écrire l'algorithme permettant de résoudre le système (S) au moyen de la méthode de Newton. On prendra $a = 10^4$.

5.8. Soit $\mathcal{A}$ une algèbre unitaire normée de dimension finie sur $\mathbb{R}$, par exemple l'algèbre des matrices carrées $m \times m$. Soit $u \in \mathcal{A}$ un élément inversible.

(a) Montrer qu'il existe des réels α, β tels que l'application $\varphi(x) = \alpha x + \beta x u x$ admette $x = u^{-1}$ comme point fixe superattractif.

(b) Pour les valeurs de α, β trouvées au (a), montrer que l'on a l'inégalité $|||\varphi'(x)||| \leq 2\|u\| \, \|x - u^{-1}\|$. En déduire que la suite itérée $x_{p+1} = \varphi(x_p)$ converge vers u^{-1} dès que $x_0 \in \overline{B}(u^{-1}, r)$ avec $r < \frac{1}{2\|u\|}$.

(c) On suppose $u = e - v$ avec $e =$ élément unité de $\mathcal{A}$ et $\lambda = \|v\| < 1$. Montrer que u est inversible et que $u^{-1} = \sum_{k=0}^{+\infty} v^k$. Déterminer un entier $n \in \mathbb{N}$ tel que l'algorithme du (b) converge pour $x_0 = e + v + \ldots + v^n$.

(d) On suppose ici que $\mathcal{A}$ est *commutative* (exemple : $\mathcal{A} = \mathbb{R}$ ou $\mathcal{A} = \mathbb{C}$). Chercher un algorithme permettant de déterminer une racine carrée de u (s'il en existe), en utilisant uniquement additions et multiplications (*Indication :* considérer $\psi(x) = \alpha x + \beta u x^3$).

Si $\mathcal{A} = \mathbb{R}$, comment peut-on choisir x_0 pour être assuré d'obtenir la convergence ?

5.9*. On suppose $f : \Omega \to \mathbb{R}^p$ de classe C^k, $k \geq 2$, où Ω est un ouvert de $\mathbb{R}^m \times \mathbb{R}^p$. Soit $(x_0, y_0) \in \Omega$ un point tel que $f(x_0, y_0) = 0$ et $\ell = f'_y(x_0, y_0)$ inversible. Donner des estimations précises de la taille des voisinages U, V intervenant dans le théorème des fonctions implicites en fonction de $|||\ell|||$, $|||\ell^{-1}|||$ et de bornes sur les dérivées premières et secondes de f sur un voisinage $B(x_0, r_0) \times B(y_0, r'_0) \subset \Omega$.

Chapitre V

Équations différentielles
Résultats fondamentaux

Le but de ce chapitre est de démontrer les théorèmes généraux d'existence et d'unicité des solutions pour les équations différentielles ordinaires. Il s'agit du chapitre central de la théorie, de ce fait nécessairement assez abstrait. Sa bonne compréhension est indispensable en vue de la lecture des chapitres ultérieurs.

1. Définitions. Solutions maximales et globales

1.1. Équation différentielle ordinaire du premier ordre

Soit U un ouvert de $\mathbb{R} \times \mathbb{R}^m$ et

$$f : U \to \mathbb{R}^m$$

une application *continue*. On considère l'équation différentielle

$$y' = f(t, y), \quad (t, y) \in U, \quad t \in \mathbb{R}, \quad y \in \mathbb{R}^m. \tag{E}$$

Définition – *Une solution de (E) sur un intervalle $I \subset \mathbb{R}$ est une fonction dérivable $y : I \to \mathbb{R}^m$ telle que*

(i) $(\forall t \in I) \qquad (t, y(t)) \in U$

(ii) $(\forall t \in I) \qquad y'(t) = f(t, y(t))$.

L'« inconnue » de l'équation (E) est donc en fait une *fonction*. Le qualificatif « ordinaire » pour l'équation différentielle (E) signifie que la fonction inconnue y dépend d'*une seule* variable t (lorsqu'il y a plusieurs variables t_i et plusieurs dérivées $\partial y/\partial t_i$, on parle d'équations aux dérivées partielles).

Écriture en coordonnées – Écrivons les fonctions à valeurs dans $\mathbb{R}^m$ en termes de leurs fonctions composantes, c'est-à-dire

$$y = (y_1, \dots, y_m), \quad f = (f_1, \dots, f_m).$$

L'équation (E) apparaît comme un *système différentiel* du premier ordre à m fonctions inconnues $y_1, \dots, y_m$:

$$\text{(E)} \quad \begin{cases} y_1'(t) = f_1(t, y_1(t), \dots, y_m(t)) \\ \dots \\ y_m'(t) = f_m(t, y_1(t), \dots, y_m(t)). \end{cases}$$

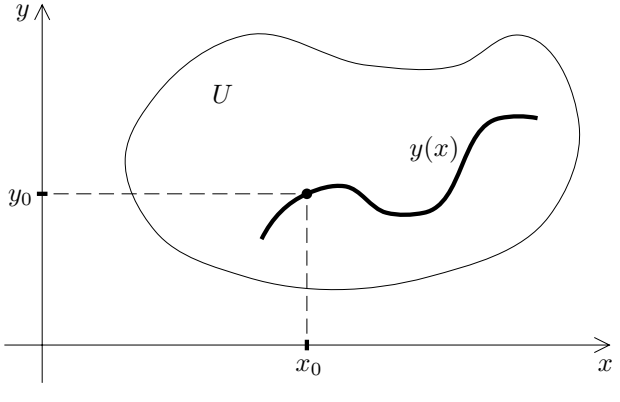
Problème de Cauchy – Étant donné un point $(t_0, y_0) \in U$, le problème de Cauchy consiste à trouver une solution $y : I \to \mathbb{R}^m$ de (E) sur un intervalle I contenant t_0 dans son intérieur, telle que $y(t_0) = y_0$.

Interprétation physique – Dans de nombreuses situations concrètes, la variable t représente le temps et $y = (y_1, \dots, y_m)$ est une famille de paramètres décrivant l'état d'un système matériel donné. L'équation (E) traduit physiquement la loi d'évolution du système considéré en fonction du temps et de la valeur des paramètres. Résoudre le problème de Cauchy revient à prévoir l'évolution du système au cours du temps, sachant qu'en $t = t_0$ le système est décrit par les paramètres $y_0 = (y_{0,1}, \dots, y_{0,m})$. On dit que (t_0, y_0) sont les *données initiales* du problème de Cauchy.

1.2. CAS DE LA DIMENSION UN $(m = 1)$

Si on note $x = t$, l'équation (E) se récrit

$$\text{(E)} \qquad y' = \frac{dy}{dx} = f(x, y), \quad (x, y) \in U \subset \mathbb{R} \times \mathbb{R}.$$

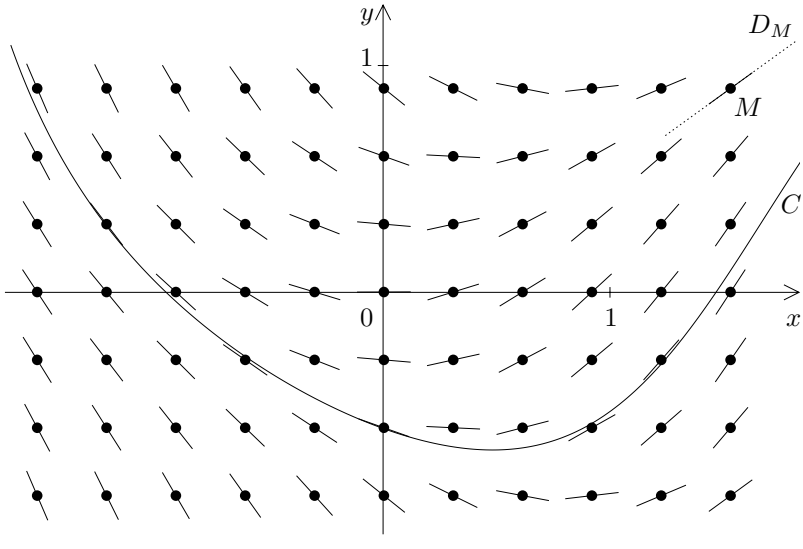


Résoudre le problème de Cauchy revient à trouver une « *courbe intégrale* » de (E) passant par un point donné $(x_0, y_0) \in U$.

Champ des tangentes – A 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 $M \to D_M$ est appelée *champ des tangentes* associé à l'équation (E). Une courbe intégrale de (E) est une courbe différentiable C qui a pour tangente en chaque point $M \in C$ la droite D_M du champ des tangentes. L'exemple ci-dessous correspond à l'équation $y' = f(x, y) = x - y^2$.



***Lignes isoclines de* (E)** – Par définition, ce sont les courbes

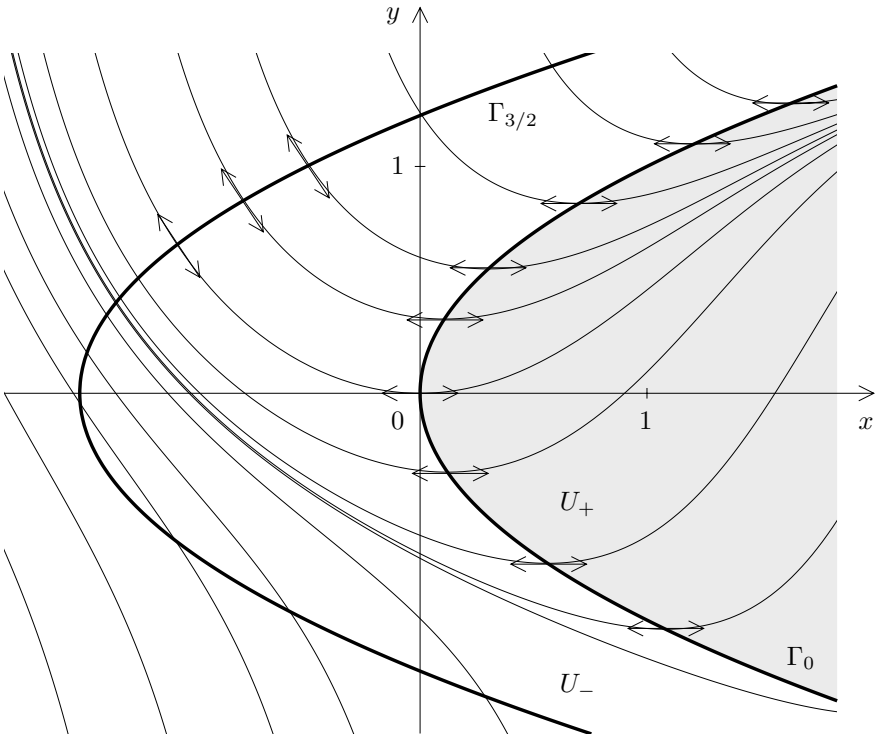
$$\Gamma_p : f(x, y) = p$$

correspondant à l'ensemble des points M où la droite D_M a une pente donnée p. La courbe Γ_0 joue un rôle intéressant. On a en effet un régionnement de U :

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

Les courbes intégrales sont croissantes dans U_+, décroissantes dans U_-, stationnaires (souvent extrêmales) sur Γ_0.

Exemple – Les lignes isoclines de l'équation $y' = f(x, y) = x - y^2$ sont les paraboles $x = y^2 + p$.



1.3. Solutions maximales

Nous introduisons d'abord le concept de prolongement d'une solution. L'expression *solution maximale* est alors entendue implicitement au sens de la relation d'ordre fournie par le prolongement des solutions.

Définition 1 – *Soient $y : I \to \mathbb{R}^m$, $\widetilde{y} : \widetilde{I} \to \mathbb{R}^m$ des solutions de* (E)*. On dit que $\widetilde{y}$ est un prolongement de y si $\widetilde{I} \supset I$ et $\widetilde{y}|_I = y$.*

Définition 2 – *On dit qu'une solution $y : I \to \mathbb{R}^m$ est* maximale *si y n'admet pas de prolongement $\widetilde{y} : \widetilde{I} \to \mathbb{R}^m$ avec $\widetilde{I} \supsetneq I$.*

Théorème – Toute solution y se prolonge en une solution maximale $\widetilde{y}$ (pas nécessairement unique).

Démonstration.* Supposons que y soit définie sur un intervalle $I = |a, b|$ (cette notation désigne un intervalle ayant pour bornes a et b, incluses ou non dans I). Il suffira de montrer que y se prolonge en une solution $\widetilde{y} : |a, \widetilde{b}| \to \mathbb{R}^m$ ($\widetilde{b} \geq b$) maximale à droite, c'est-à-dire qu'on ne pourra plus prolonger $\widetilde{y}$ au delà de $\widetilde{b}$. Le même raisonnement s'appliquera à gauche.

Pour cela, on construit par récurrence des prolongements successifs $y_{(1)}, y_{(2)} \ldots$ de y avec $y_{(k)} : |a, b_k[\to \mathbb{R}^m$. On pose $y_{(1)} = y$, $b_1 = b$. Supposons $y_{(k-1)}$ déjà construite

pour un indice $k \geq 1$. On pose alors

$$c_k = \sup\{c\,;\, y_{(k-1)} \text{ se prolonge sur } |a,c[\,\}.$$

On a $c_k \geq b_{k-1}$. Par définition de la borne supérieure, il existe b_k tel que $b_{k-1} \leq b_k \leq c_k$ et un prolongement $y_{(k)} : |a,b_k[\to \mathbb{R}^m$ de $y_{(k-1)}$ avec b_k arbitrairement voisin de c_k ; en particulier, on peut choisir

$$\begin{aligned} c_k - b_k &< \tfrac{1}{k} \qquad \text{si} \quad c_k < +\infty, \\ b_k &> k \qquad\quad \text{si} \quad c_k = +\infty. \end{aligned}$$

La suite (c_k) est décroissante, car l'ensemble des prolongements de $y_{(k-1)}$ contient l'ensemble des prolongements de $y_{(k)}$; au niveau des bornes supérieures on a donc $c_k \geq c_{k+1}$. Si $c_k < +\infty$ à partir d'un certain rang, les suites

$$b_1 \leq b_2 \leq \ldots \leq b_k \leq \ldots \leq c_k \leq c_{k-1} \leq \ldots \leq c_1$$

sont adjacentes, tandis que si $c_k = +\infty$ quel que soit k on a $b_k > k$. Dans les deux cas, on voit que

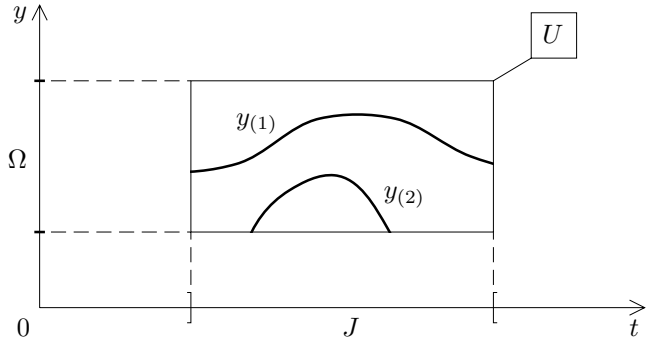
$$\widetilde{b} = \lim_{k\to+\infty} b_k = \lim_{k\to+\infty} c_k.$$

Soit $\widetilde{y} : |a,\widetilde{b}| \to \mathbb{R}^m$ le prolongement commun des solutions $y_{(k)}$, éventuellement prolongé au point $\widetilde{b}$ si cela est possible. Soit $z : |a,c| \to \mathbb{R}^m$ un prolongement de $\widetilde{y}$. Alors z prolonge $y_{(k-1)}$ et par définition de c_k il s'ensuit $c \leq c_k$. A la limite il vient $c \leq \widetilde{c}$, ce qui montre que la solution $\widetilde{y}$ est maximale à droite. ■

1.4. *Solutions globales*

On suppose ici que l'ouvert U est de la forme $U = J \times \Omega$ où J est un intervalle de $\mathbb{R}$ et Ω un ouvert de $\mathbb{R}^m$.

Définition – *Une solution globale est une solution définie sur l'intervalle J tout entier.*



Attention : toute solution globale est maximale, mais la réciproque est fausse.

Sur le schéma ci-dessus par exemple, $y_{(1)}$ est globale tandis que $y_{(2)}$ est maximale mais non globale.

Donnons un exemple explicite de cette situation.

Exemple – (E) $y' = y^2$ sur $U = \mathbb{R} \times \mathbb{R}$.

Cherchons les solutions $t \to y(t)$ de (E).

- On a d'une part la solution $y(t) = 0$.
- Si y ne s'annule pas, (E) s'écrit $\frac{y'}{y^2} = 1$, d'où par intégration

$$-\frac{1}{y(t)} = t + C, \quad y(t) = -\frac{1}{t+C}.$$

Cette formule définit en fait deux solutions, définies respectivement sur $]-\infty, -C[$ et sur $]-C, +\infty[$; ces solutions sont maximales mais non globales. Dans cet exemple $y(t) = 0$ est la seule solution globale de (E).

1.5. Régularité des solutions

Rappelons qu'une fonction de plusieurs variables est dite de classe C^k si elle admet des dérivées partielles continues jusqu'à l'ordre k.

Théorème – *Si $f : \mathbb{R} \times \mathbb{R}^m \supset U \to \mathbb{R}^m$ est de classe C^k, toute solution de* (E) $y' = f(t, y)$ *est de classe C^{k+1}.*

Démonstration. On raisonne par récurrence sur k.

- $k = 0$: f continue.

Par hypothèse $y : I \to \mathbb{R}^m$ est dérivable, donc continue.

Par conséquent $y'(t) = f(t, y(t))$ est continue, donc y est de classe C^1.

- Si le résultat est vrai à l'ordre $k-1$, alors y est au moins de classe C^k. Comme f est de classe C^k, il s'ensuit que y' est de classe C^k comme composée de fonctions de classe C^k, donc y est de classe C^{k+1}.

Calcul des dérivées successives d'une solution y – On suppose pour simplifier $m = 1$. En dérivant la relation $y'(x) = f(x, y(x))$ il vient

$$\begin{aligned} y''(x) &= f'_x(x, y(x)) + f'_y(x, y(x))y'(x), \\ y'' &= f'_x(x, y) + f'(x, y)f(x, y) = f^{[1]}(x, y) \end{aligned}$$

avec $f^{[1]} = f'_x + f'_y f$. Notons de manière générale l'expression de la dérivée k-ième $y^{(k)}$ en fonction de x, y sous la forme

$$y^{(k)} = f^{[k-1]}(x, y) \ ;$$

d'après ce qui précède $f^{[0]} = f$, $f^{[1]} = f'_x + f'_y f$. En dérivant une nouvelle fois, on trouve

$$\begin{aligned} y^{(k+1)} &= (f^{[k-1]})'_x(x,y) + (f^{[k-1]})'_y(x,y)\, y' \\ &= (f^{[k-1]})'_x(x,y) + (f^{[k-1]})'_y(x,y)\, f(x,y). \end{aligned}$$

On obtient donc les relations de récurrence

$$\begin{aligned} y^{(k+1)} &= f^{[k]}(x,y) \\ f^{[k]} &= (f^{[k-1]})'_x + (f^{[k-1]})'_y\, f, \quad \text{avec} \quad f^{[0]} = f. \end{aligned}$$

En particulier, le lieu des points d'inflexion des courbes intégrales est contenu dans la courbe $f^{[1]}(x,y) = 0$.

2. Théorème d'existence des solutions

Dans tout ce paragraphe, on considère une équation différentielle

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

où $f : U \to \mathbb{R}^m$ est continue et U est un ouvert de $\mathbb{R} \times \mathbb{R}^m$.

2.1. Équivalence du problème de Cauchy avec la résolution d'une équation intégrale

Le lemme très simple ci-dessous montre que la résolution de (E) est équivalente à la résolution d'une équation intégrale :

Lemme – *Une fonction $y : I \to \mathbb{R}^m$ est une solution du problème de Cauchy de données initiales (t_0, y_0) si et seulement si*

(i) *y est continue et $(\forall t \in I)\ (t, y(t)) \in U$,*

(ii) $(\forall t \in I) \quad y(t) = y_0 + \displaystyle\int_{t_0}^{t} f(u, y(u))du.$

En effet si y vérifie (i) et (ii) alors y est différentiable et on a $y(t_0) = y_0$, $y'(t) = f(t, y(t))$. Inversement, si ces deux relations sont satisfaites, (ii) s'en déduit par intégration. ∎

2.2. Cylindres de sécurité

Pour résoudre l'équation différentielle (E), on va plutôt chercher à construire des solutions de l'équation intégrale 2.1 (ii), et en premier lieu, on va montrer qu'une solution passant par un point $(t_0, y_0) \in U$ ne peut s'éloigner « trop vite » de y_0.

On note $\|\ \|$ une norme quelconque sur $\mathbb{R}^m$ et $B(x,r)$ (resp. $\overline{B}(x,r)$) la boule ouverte (resp. fermée) de centre x et de rayon r dans $\mathbb{R}^m$. Comme U est supposé ouvert, il existe un *cylindre*

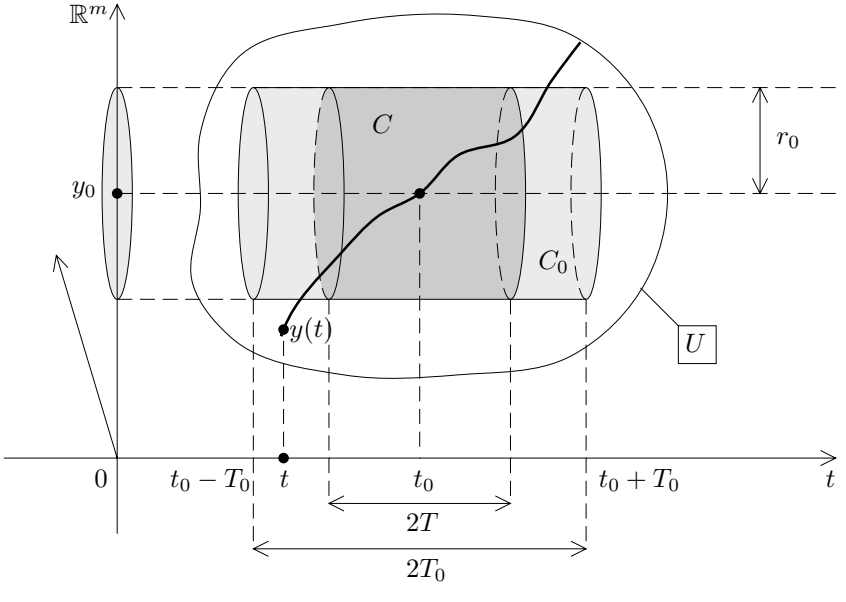
$$C_0 = [t_0 - T_0, t_0 + T_0] \times \overline{B}(y_0, r_0)$$

de longueur $2T_0$ et de rayon r_0 assez petit, tel que $C_0 \subset U$. L'ensemble C_0 est fermé borné dans $\mathbb{R}^{m+1}$, donc compact. Ceci entraîne que f est bornée sur C_0, c'est-à-dire

$$M = \sup_{(t,y)\in C_0} \|f(t,y)\| < +\infty.$$

Soit $C = [t_0 - T, t_0 + T] \times \overline{B}(y_0, r_0) \subset C_0$ un cylindre de même diamètre que C_0 et de demi-longueur $T \leq T_0$.

Définition – *On dit que C est un cylindre de sécurité pour l'équation* (E) *si toute solution $y : I \to \mathbb{R}^m$ du problème de Cauchy $y(t_0) = y_0$ avec $I \subset [t_0 - T, t_0 + T]$ reste contenue dans $\overline{B}(y_0, r_0)$.*



Sur le schéma ci-dessus, C est un cylindre de sécurité mais C_0 n'en est pas un : la solution y « s'échappe » de C_0 avant le temps $t_0 + T_0$.

Supposons que la solution y s'échappe de C sur l'intervalle $[t_0, t_0 + T]$. Soit τ le premier instant où cela se produit :

$$\tau = \inf \{t \in [t_0, t_0 + T]\,;\ \|y(t) - y_0\| > r_0\}.$$

Par définition de τ on a $\|y(t) - y_0\| \leq r$ pour $t \in [t_0, \tau[$, donc par continuité de y on obtient $\|y(\tau) - y_0\| = r_0$. Comme $(t, y(t)) \in C \subset C_0$ pour $t \in [t_0, \tau]$, il vient $\|y'(t)\| = \|f(t, y(t))\| \leq M$ et

$$r_0 = \|y(\tau) - y_0\| = \Big\| \int_{t_0}^{\tau} y'(u)du \Big\| \leq M(\tau - t_0)$$

donc $\tau - t_0 \geq r_0/M$. Par conséquent si $T \leq r_0/M$, aucune solution ne peut s'échapper de C sur $[t_0 - T, t_0 + T]$.

Corollaire – *Pour que C soit un cylindre de sécurité, il suffit de prendre*

$$T \leq \min\Big(T_0, \frac{r_0}{M}\Big).$$

Le choix $T = \min\big(T_0, \frac{r_0}{M}\big)$ convient par exemple.

Remarque – Si $C \subset C_0$ est un cylindre de sécurité, toute solution du problème de Cauchy $y : [t_0 - T, t_0 + T] \to \mathbb{R}^m$ vérifie $\|y'(t)\| \leq M$, donc y est lipschitzienne de rapport M.

2.3. Solutions approchées. Méthode d'Euler

On cherche à construire une solution approchée de (E) sur un intervalle $[t_0, t_0 + T]$. On se donne pour cela une subdivision

$$t_0 < t_1 < t_2 \ldots < t_{N-1} < t_N = t_0 + T.$$

Les pas successifs sont notés

$$h_n = t_{n+1} - t_n, \quad 0 \leq n \leq N - 1,$$

et on pose

$$h_{\max} = \max(h_0, \ldots, h_{N-1}).$$

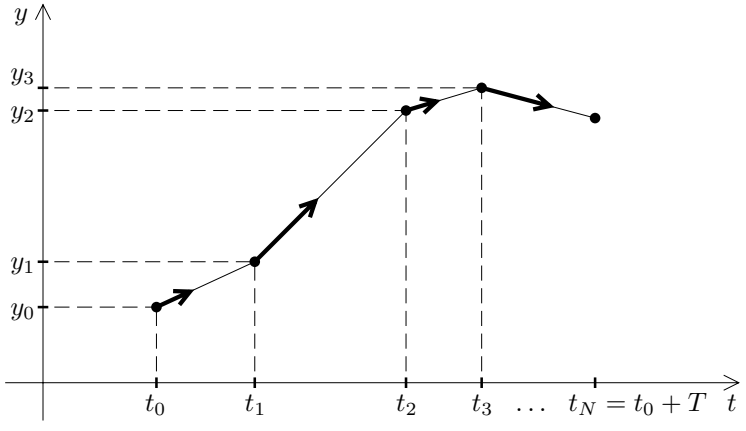
La méthode d'Euler (ou méthode de la tangente) consiste à construire une solution approchée y affine par morceaux comme suit. Soit $y_n = y(t_n)$. On confond la courbe intégrale sur $[t_n, t_{n+1}]$ avec sa tangente au point (t_n, y_n) :

$$y(t) = y_n + (t - t_n)f(t_n, y_n), \quad t \in [t_n, t_{n+1}].$$

Partant de la donnée initiale y_0, on calcule donc y_n par récurrence en posant

$$\begin{cases} y_{n+1} = y_n + h_n f(t_n, y_n) \\ t_{n+1} = t_n + h_n, \quad 0 \leq n \leq N - 1. \end{cases}$$

La solution approchée y s'obtient graphiquement en traçant pour chaque n les segments joignant les points (t_n, y_n), (t_{n+1}, y_{n+1}).



On construit de même une solution approchée sur $[t_0 - T, t_0]$ en prenant des pas $h_n < 0$.

Proposition 1 – *Si $C = [t_0 - T, t_0 + T] \times \overline{B}(y_0, r_0)$ est un cylindre de sécurité tel que $T \leq \min\left(T_0, \frac{r_0}{M}\right)$, toute solution approchée y donnée par la méthode d'Euler est contenue dans la boule $\overline{B}(y_0, r_0)$.*

Démonstration. On vérifie par récurrence sur n que

$$\begin{cases} y([t_0, t_n]) \subset \overline{B}(y_0, r_0) \\ \|y(t) - y_0\| \leq M(t - t_0) \quad \text{pour} \quad t \in [t_0, t_n]. \end{cases}$$

C'est trivial pour $n = 0$. Si c'est vrai pour n, alors on a en particulier $(t_n, y_n) \in C$, donc $\|f(t_n, y_n)\| \leq M$, et par conséquent

$$\|y(t) - y_n\| = (t - t_n)\|f(t_n, y_n)\| \leq M(t - t_n)$$

pour $t \in [t_n, t_{n+1}]$. Par hypothèse de récurrence

$$\|y_n - y_0\| = \|y(t_n) - y_0\| \leq M(t_n - t_0).$$

L'inégalité triangulaire entraîne alors $\forall t \in [t_n, t_{n+1}]$:

$$\|y(t) - y_0\| \leq M(t - t_n) + M(t_n - t_0) \leq M(t - t_0).$$

En particulier $\|y(t) - y_0\| \leq MT \leq r_0$, d'où

$$y([t_0, t_{n+1}]) \subset \overline{B}(y_0, r_0). \qquad \blacksquare$$

Définition – *Soit $y : [a, b] \to \mathbb{R}^m$ une fonction de classe C^1 par morceaux (ceci signifie qu'il existe une subdivision $a = a_0 < a_1 < \ldots < a_N = b$ de $[a, b]$ telle que pour tout n la restriction $y_{[a_n, a_{n+1}]}$ soit de classe C^1 ; on suppose donc seulement la continuité et l'existence d'une dérivée à droite et à gauche de y aux points a_n). On dit que y est une solution ε-approchée de* (E) *si*

(i) $(\forall t \in [a, b]) \quad (t, y(t)) \in U$;

(ii) $(\forall n)$, $(\forall t \in]a_n, a_{n+1}[) \quad \|y'(t) - f(t, y(t))\| \leq \varepsilon$.

Autrement dit, y est une solution ε-approchée si y vérifie (E) avec une erreur $\leq \varepsilon$.

Majoration de l'erreur pour les solutions approchées d'Euler – Soit ω_f le module de continuité de f sur C, défini par

$$\omega_f(u) = \max\{\|f(t_1,y_1) - f(t_2,y_2)\|\,;\ |t_1-t_2| + \|y_1-y_2\| \leq u\}$$

où $u \in [0,+\infty[$ et où les points (t_1,y_1), (t_2,y_2) parcourent C. Comme C est compact, la fonction f est uniformément continue sur C, par conséquent

$$\lim_{u\to 0_+} \omega_f(u) = 0.$$

On suppose dans la suite que $C = [t_0-T, t_0+T] \times \overline{B}(y_0, r_0)$ est un cylindre de sécurité tel que $T \leq \min\left(T_0, \frac{r_0}{M}\right)$.

Proposition 2 – *Soit $y : [t_0-T, t_0+T] \to \mathbb{R}^m$ une solution approchée construite par la méthode d'Euler avec pas maximum $h_{\max}$. Alors l'erreur ε vérifie $\varepsilon \leq \omega_f((M+1)h_{\max})$.*

En particulier, l'erreur ε tend vers 0 quand $h_{\max}$ tend vers 0.

Démonstration. Majorons par exemple $\|y'(t) - f(t,y(t))\|$ pour $t \in [t_0, t_0+T]$, où y est la solution approchée associée à la subdivision $t_0 < t_1 < \ldots < t_N = t_0+T$. Pour $t \in]t_n, t_{n+1}[$, on a $y'(t) = f(t_n, y_n)$ et

$$\begin{aligned} \|y(t) - y_n\| &= (t-t_n)\|f(t_n,y_n)\| \leq Mh_n, \\ |t-t_n| &\leq h_n. \end{aligned}$$

Par définition de ω_f, il vient

$$\begin{aligned} \|f(t_n,y_n) - f(t,y(t))\| &\leq \omega_f(Mh_n + h_n), \\ \|y'(t) - f(t,y(t))\| &\leq \omega_f((M+1)h_{\max}. \end{aligned}$$

■

Montrons finalement un résultat sur la convergence des solutions approchées.

Proposition 3 – *Soit $y_{(p)} : [t_0-T, t_0+T] \to \mathbb{R}^m$ une suite de solutions ε_p-approchées contenues dans le cylindre de sécurité C, telles que $y_{(p)}(t_0) = y_0$ et $\lim_{p\to+\infty} \varepsilon_p = 0$. On suppose que $y_{(p)}$ converge uniformément sur $[t_0-T, t_0+T]$ vers une fonction y. Alors y est une solution exacte du problème de Cauchy pour l'équation* (E).

Démonstration. Comme $\|y'_{(p)}(t) - f(t, y_{(p)}(t)\| \leq \varepsilon_p$, il vient après intégration

$$\|y_{(p)}(t) - y_0 - \int_{t_0}^{t} f(u, y_{(p)}(u))du\| \leq \varepsilon_p |t-t_0|.$$

Si $\delta_p = \max_{[t_0-T,t_0+T]} \|y - y_{(p)}\|$, on voit que

$$\|f(u, y_{(p)}(u)) - f(u, y(u))\| \leq \omega_f(\delta_p)$$

tend vers 0, d'où, grâce à la convergence uniforme :

$$y(t) - y_0 - \int_{t_0}^{t} f(u, y(u))du = 0, \quad \forall t \in [t_0 - T, t_0 + T].$$

Comme la limite uniforme y est continue, le lemme du début du § 2 entraîne que y est une solution exacte de (E).

2.4. Théorème d'Ascoli

Il s'agit d'un résultat préliminaire de nature topologique que nous allons formuler dans le cadre général des espaces métriques. Si (E, δ) et (F, δ') sont des espaces métriques, rappelons que par définition une suite d'applications $\varphi_{(p)} : E \to F$ converge uniformément vers $\varphi : E \to F$ si la distance uniforme

$$d(\varphi_{(p)}, \varphi) = \sup_{x \in E} \delta'(\varphi_{(p)}(x), \varphi(x))$$

tend vers 0 quand p tend vers $+\infty$.

Théorème (Ascoli) – *On suppose que E, F sont des espaces métriques* compacts. *Soit $\varphi_{(p)} : E \to F$ une suite d'applications k-lipschitziennes, où $k \geq 0$ est une constante donnée. Alors on peut extraire de $\varphi_{(p)}$ une sous-suite $\varphi_{(p_n)}$ uniformément convergente, et la limite est une application k-lipschitzienne.*

Soit $\mathrm{Lip}_k(E, F)$ l'ensemble des applications $E \to F$ lipschitziennes de rapport k. Une autre manière d'exprimer le théorème d'Ascoli est la suivante.

Corollaire – *Si E, F sont compacts, alors $(\mathrm{Lip}_k(E, F), d)$ est un espace métrique compact.*

Démonstration. On construit par récurrence des parties infinies

$$S_0 = \mathbb{N} \supset S_1 \supset \ldots \supset S_{n-1} \supset S_n \supset \ldots$$

telles que la sous-suite $(\varphi_{(p)})_{p \in S_n}$ ait des oscillations de plus en plus faibles.

Supposons S_{n-1} construite, $n \geq 1$. Comme E, F sont compacts, il existe des recouvrements finis de E (resp. de F) par des boules ouvertes $(B_i)_{i \in I}$, resp. $(B'_j)_{j \in J}$, de rayon $\frac{1}{n}$. Notons $I = \{1, 2, \ldots, N\}$ et x_i le centre de B_i. Soit p un indice fixé. Pour tout $i = 1, \ldots, N$ il existe un indice $j = j(p, i)$ tel que $\varphi_{(p)}(x_i) \in B'_{j(p,i)}$.

On considère l'application

$$S_{n-1} \longrightarrow J^N, \qquad p \longmapsto (j(p, 1), \ldots, j(p, N)).$$

Comme S_{n-1} est infini et que J^N est fini, l'un des éléments $(l_1, \dots, l_N) \in J^N$ admet pour image réciproque une partie infinie de S_{n-1} : on note S_n cette partie. Ceci signifie que pour tout $p \in S_n$ on a $(j(p,1), \dots, j(p,N)) = (l_1, \dots, l_N)$ et donc $\varphi_{(p)}(x_i) \in B'_{l_i}$. En particulier

$$(\forall p, q \in S_n) \quad \delta'(\varphi_{(p)}(x_i), \varphi_{(q)}(x_i)) \leq \operatorname{diam} B'_{l_i} \leq \frac{2}{n}.$$

Soit $x \in E$ un point quelconque. Il existe $i \in I$ tel que $x \in B_i$, d'où $\delta(x, x_i) < \frac{1}{n}$. L'hypothèse que les $\varphi_{(p)}$ sont k-lipschitziennes entraîne

$$\delta'(\varphi_{(p)}(x), \varphi_{(p)}(x_i)) < \frac{k}{n}, \quad \delta'(\varphi_{(q)}(x), \varphi_{(q)}(x_i)) < \frac{k}{n}.$$

L'inégalité triangulaire implique alors $(\forall p, q \in S_n)$

$$\delta'(\varphi_{(p)}(x), \varphi_{(q)}(x)) \leq \frac{2}{n} + 2\,\frac{k}{n} = \frac{2k+2}{n}.$$

Désignons par p_n le n-ième élément de S_n. Pour $N \geq n$ on a $p_N \in S_N \subset S_n$, donc

$$\delta'(\varphi_{(p_n)}(x), \varphi_{(p_N)}(x)) \leq \frac{2k+2}{n}. \qquad (*)$$

Ceci entraîne que $\varphi_{(p_n)}(x)$ est une suite de Cauchy dans F pour tout $x \in E$. Comme F est compact, F est aussi complet, donc $\varphi_{(p_n)}(x)$ converge vers une limite $\varphi(x)$. Quand $N \to +\infty$, $(*)$ implique à la limite $d(\varphi_{(p_n)}, \varphi) \leq \frac{2k+2}{n}$. On voit donc que $\varphi_{(p_n)}$ converge uniformément vers φ. Il est facile de voir que $\varphi \in \operatorname{Lip}_k(E, F)$. ■

Exercice – *On pose $E = [0, \pi]$, $F = [-1, 1]$, $\varphi_p(x) = \cos px$. Calculer*

$$\int_0^\pi (\varphi_p(x) - \varphi_q(x))^2 dx$$

et en déduire que $d(\varphi_p, \varphi_q) \geq 1$ si $p \neq q$. L'espace

$$\operatorname{Lip}(E, F) = \bigcup\nolimits_{k \geq 0} \operatorname{Lip}_k(E, F)$$

est-il compact ?

2.5. Théorème d'existence (Cauchy-Peano-Arzela)

L'idée est d'utiliser le théorème d'Ascoli pour montrer l'existence d'une sous-suite uniformément convergente de solutions approchées. On obtient ainsi le

Théorème – *Soit $C = [t_0 - T, t_0 + T] \times \overline{B}(y_0, r_0)$ avec $T \leq \min\left(T_0, \frac{r_0}{M}\right)$ un cylindre de sécurité pour l'équation* (E) *: $y' = f(t, y)$. Alors il existe une solution $y : [t_0 - T, t_0 + T] \to \overline{B}(y_0, r_0)$ de* (E) *avec condition initiale $y'(t_0) = y_0$.*

Démonstration. Soit $y_{(p)}$ la solution approchée donnée par la méthode d'Euler en utilisant la subdivision avec pas constant $h = T/p$ des intervalles $[t_0, t_0 + T]$ et $[t_0 - T, t_0]$. Cette solution est ε_p-approchée avec erreur $\varepsilon_p \leq \omega_f((M+1)\,T/p)$ tendant vers 0. Chaque application $y_{(p)} : [t_0 - T, t_0 + T] \to \overline{B}(y_0, r_0)$ est lipschitzienne de rapport M, donc d'après le théorème d'Ascoli on peut extraire de $(y_{(p)})$ une sous-suite $(y_{(p_n)})$ convergeant uniformément vers une limite y. D'après la proposition 3 du § 2.3, y est une solution exacte de l'équation (E). ■

Corollaire – *Par tout point $(t_0, y_0) \in U$, il passe au moins une solution maximale $y : I \to \mathbb{R}^m$ de* (E)*. De plus, l'intervalle de définition I de toute solution maximale est ouvert* (*mais en général, il n'y a pas unicité de ces solutions maximales*)*.*

On vient de voir en effet qu'il existe une solution locale z définie sur un intervalle $[t_0 - T, t_0 + T]$. D'après le théorème du § 1.3, z se prolonge en une solution maximale $y = \widetilde{z} : |a, b| \to \mathbb{R}^m$. Si y était définie au point b, il existerait une solution $y_{(1)} : [b - \varepsilon, b + \varepsilon] \to \mathbb{R}^m$ du problème de Cauchy avec donnée initiale $(b, y(b)) \in U$. La fonction $\widetilde{y} : |a, b + \varepsilon] \to \mathbb{R}^m$ coïncidant avec y sur $|a, b]$ et avec $y_{(1)}$ sur $[b, b + \varepsilon]$ serait alors un prolongement strict de y, ce qui est absurde. ■

Exemple – Pour donner un exemple de non unicité, il suffit de considérer l'équation $y' = 3|y|^{2/3}$. Le problème de Cauchy de condition initiale $y(0) = 0$ admet alors au moins 2 solutions maximales :

$$y_{(1)}(t) = 0, \quad y_{(2)}(t) = t^3, \quad t \in \mathbb{R}.$$

2.6. CRITÈRE DE MAXIMALITÉ DES SOLUTIONS

Nous allons voir ici une condition géométrique nécessaire et suffisante permettant d'affirmer qu'une solution est maximale.

Théorème – *U un ouvert de $\mathbb{R} \times \mathbb{R}^m$ et $y : I = [t_0, b[\to \mathbb{R}^m$ une solution de l'équation* (E) *$y' = f(t, y)$, où f est une fonction continue sur U. Alors $y(t)$ peut se prolonger au delà de b si et seulement si il existe un compact $K \subset U$ tel que la courbe $t \mapsto (t, y(t))$, $t \in [t_0, b[$, reste contenue dans K.*

Autrement dit, y est non prolongeable au delà du temps b si et seulement si $(t, y(t))$ s'échappe de tout compact K de U quand $t \to b_-$. La conséquence suivante est immédiate.

Critère de maximalité – *Une solution $y :]a, b[\to \mathbb{R}^m$ de* (E) *est maximale si et seulement si $t \mapsto (t, y(t))$ s'échappe de tout compact K de U quand $t \to a_+$ ou quand $t \to b_-$. Puisque les compacts sont les parties fermées bornées, ceci signifie encore que $(t, y(t))$ s'approche du bord de U ou tend vers ∞, c'est-à-dire $|t| + \|y(t)\| + 1/d\big((t, y(t)), \partial U\big) \to +\infty$ quand $t \to a_+$ ou $t \to b_-$.*

Démonstration du théorème. La condition de prolongement est évidemment nécessaire, puisque si $y(t)$ se prolonge à $[t_0, b]$, alors l'image du compact $[t_0, b]$ par l'application continue $t \mapsto (t, y(t))$ est un compact $K \subset U$.

Inversement, supposons qu'il existe un compact K de U tel que $(t, y(t)) \in K$ pour tout $t \in [t_0, b[$. Posons

$$M = \sup_{(t,y)\in K} \|f(t,y)\| < +\infty$$

qui est fini par continuité de $|f\|$ et compacité de K. Ceci entraîne que $t \mapsto y(t)$ est lipschitzienne sur $[t_0, b[$, donc uniformément continue, et le critère de cauchy montre que la limite $\ell = \lim_{t\to b_-} y(t)$ existe. Nous pouvons prolonger y par continuité en b en posant $y(b) = \ell$, et nous avons $(b, y(b)) \in K \subset U$ puisque K est fermé. La relation $y'(t) = f(t, y(t))$ montre alors que y est de classe C^1 sur $[t_0, b]$. Maintenant, le théorème d'existence locale des solutions implique qu'il existe une solution locale z d problème de Cauchy de donnée initiale $z(b) = \ell = y(b)$ sur un intervalle $[b-\varepsilon, b+\varepsilon]$. On obtient alors un prolongement $\widetilde{y}$ de y sur $[t_0, b+\varepsilon]$ en posant $\widetilde{y}(t) = z(t)$ pour $t \in [b, b+\varepsilon]$. Le théorème est démontré. ∎

3. Théorème d'existence et d'unicité de Cauchy-Lipschitz

Reprenons les notations du début du § 2. On suppose ici en outre que f est *localement lipschitzienne* en y : cela signifie que pour tout point $(t_0, y_0) \in U$ il existe un cylindre $C_0 = [t_0 - T_0, t_0 + T_0] \times \overline{B}(y_0, r_0) \subset U$ et une constante $k = k(t_0, y_0) \geq 0$ tels que f soit k-lipschitzienne en y sur C_0 :

$$\Big(\forall (t, y_1), (t, y_2) \in C_0\Big) \quad \|f(t, y_1) - f(t, y_2)\| \leq k\|y_1 - y_2\|.$$

Remarque – Pour que f soit localement lipschitzienne en y sur U, il suffit que f admette des dérivées partielles $\frac{\partial f_i}{\partial y_j}$, $1 \leq i, j \leq m$, *continues* sur U. Soit en effet

$$A = \max_{1\leq i,j\leq m} \sup_{(t,y)\in C_0} \left|\frac{\partial f_i}{\partial y_j}(t, y)\right|.$$

Le nombre A est fini puisque C_0 est compact. Le théorème des accroissement finis appliqués à f_i sur C_0 donne

$$f_i(t, y_1) - f_i(t, y_2) = \sum_j \frac{\partial f_i}{\partial y_j}\,(t, \xi)(y_{1,j} - y_{2,j})$$

avec $\xi \in]y_1, y_2[$. On a donc

$$\max_i |f_i(t, y_1) - f(t, y_2)| \leq mA \cdot \max_j |y_{1,j} - y_{2,j}|. \qquad ∎$$

Sous ces hypothèses sur f, nous allons montrer que la solution du problème de Cauchy est nécessairement unique, et que de plus toute suite de solutions ε-approchées avec ε tendant vers 0 converge nécessairement vers la solution exacte. Compte tenu de l'importance de ces résultats, nous donnerons ensuite une deuxième démonstration assez différente basée sur le théorème du point fixe (chapitre IV, § 1.1).

3.1. LEMME DE GRONWALL. CONVERGENCE ET UNICITÉ LOCALES

Soit $C_0 = [t_0 - T_0, t_0 + T_0] \times \overline{B}(y_0, r_0) \subset U$ un cylindre sur lequel f est k-lipschitzienne en y et soit $M = \sup_{C_0} \|f\|$. On se donne $\varepsilon > 0$ et on considère des solutions $y_{(1)}$ et $y_{(2)}$ respectivement ε_1-approchée et ε_2-approchée du problème de Cauchy de donnée initiale (t_0, y_0), avec $\varepsilon_1, \varepsilon_2 \leq \varepsilon$.

On a alors $\|y'_{(i)}(t)\| \leq M + \varepsilon$, et un raisonnement analogue à celui du § 2.1 montre que les graphes de $y_{(1)}, y_{(2)}$ restent contenus dans le cylindre

$$C = [t_0 - T, t_0 + T] \times \overline{B}(y, r_0) \subset C_0$$

dès que $T \leq \min\left(T_0, \frac{r_0}{M+\varepsilon}\right)$, ce qu'on suppose désormais.

Lemme de Gronwall – *Sous les hypothèses précédentes, on a*

$$\|y_{(2)}(t) - y_{(1)}(t)\| \leq (\varepsilon_1 + \varepsilon_2)\, \frac{e^{k|t-t_0|} - 1}{k}, \quad \forall t \in [t_0 - T, t_0 + T].$$

Démonstration. Quitte à changer l'origine du temps on peut supposer $t_0 = 0$ et, par exemple, $t \in [0, T]$. Posons alors

$$v(t) = \int_0^t \|y_{(2)}(u) - y_{(1)}(u)\| du.$$

Comme $y_{(i)}$ satisfait l'équation différentielle à ε_i près, on obtient par soustraction

$$\begin{aligned} \|y'_{(2)}(t) - y'_{(1)}(t)\| &\leq \|f(t, y_{(2)}(t) - f(t, y_{(1)}(t)\| + \varepsilon_1 + \varepsilon_2 \\ &\leq k\|y_{(2)}(t) - y_{(1)}(t)\| + \varepsilon_1 + \varepsilon_2, \end{aligned}$$

en utilisant l'hypothèse que f est k-lipschitzienne en y. De plus

$$y_{(2)}(t) - y_{(1)}(t) = \int_0^t (y'_{(2)}(u) - y'_{(1)}(u)) du$$

puisque $y_{(2)}(0) = y_{(1)}(0) = y_0$. On en déduit

$$\|y_{(2)}(t) - y_{(1)}(t)\| \leq k \int_0^t \|y_{(2)}(u) - y_{(1)}(u)\| du + (\varepsilon_1 + \varepsilon_2)t \qquad (*)$$

c'est-à-dire

$$v'(t) \leq kv(t) + (\varepsilon_1 + \varepsilon_2)t.$$

Après soustraction de $kv(t)$ et multiplication par e^{-kt}, on trouve

$$(v'(t) - kv(t))e^{-kt} = \frac{d}{dt}\left(v(t)e^{-kt}\right) \leq (\varepsilon_1 + \varepsilon_2)te^{-kt}.$$

Grâce à une nouvelle intégration (noter que $v(0) = 0$), il vient

$$v(t)e^{-kt} \leq \int_0^t (\varepsilon_1 + \varepsilon_2) u e^{-ku} du = (\varepsilon_1 + \varepsilon_2)\ \frac{1 - (1+kt)e^{-kt}}{k^2},$$
$$v(t) \leq (\varepsilon_1 + \varepsilon_2)\ \frac{e^{kt} - (1+kt)}{k^2},$$

tandis que la première inégalité intégrée $(*)$ donne

$$\|y_{(2)}(t) - y_{(1)}(t)\| \leq kv(t) + (\varepsilon_1 + \varepsilon_2)t \leq (\varepsilon_1 + \varepsilon_2)\ \frac{e^{kt} - 1}{k}.$$

Le cas où $t \in [-T, 0]$ s'obtient par un changement de variable $t \mapsto -t$. ■

Théorème (Cauchy-Lipschitz) – *Si $f : U \to \mathbb{R}^m$ est localement lipschitzienne en y, alors pour tout cylindre de sécurité $C = [t_0 - T, t_0 + T] \times \overline{B}(y_0, r_0)$ comme ci-dessus, le problème de Cauchy avec condition initiale (t_0, y_0) admet une unique solution exacte $y : [t_0 - T, t_0 + T] \to U$. De plus, toute suite $y_{(p)}$ de solutions ε_p-approchées avec ε_p tendant vers 0 converge uniformément vers la solution exacte y sur $[t_0 - T, t_0 + T]$.*

Existence. Soit $y_{(p)}$ une suite quelconque de solutions ε_p approchées avec $\lim \varepsilon_p = 0$, par exemple celles fournies par la méthode d'Euler. Le lemme de Gronwall montre que

$$d(y_{(p)}, y_{(q)}) \leq (\varepsilon_p + \varepsilon_q)\ \frac{e^{kT} - 1}{k} \quad \text{sur} \quad [t_0 - T, t_0 + T],$$

par conséquent $y_{(p)}$ est une suite de Cauchy uniforme. Comme les fonctions $y_{(p)}$ sont toutes à valeurs dans $\overline{B}(y_0, r_0)$ qui est un espace complet, $y_{(p)}$ converge vers une limite y. Cette limite y est une solution exacte de l'équation (E) d'après la proposition 3 du § 2.3.

Unicité. Si $y_{(1)}, y_{(2)}$ sont deux solutions exactes, le lemme de Gronwall avec $\varepsilon_1 = \varepsilon_2 = 0$ montre que $y_{(1)} = y_{(2)}$. ■

3.2. Autre démonstration (par le théorème du point fixe)*

Soit $C = [t_0 - T, t_0 + T] \times \overline{B}(y_0, r_0) \subset C_0$ avec $T \leq \min\left(T_0, \frac{r_0}{M}\right)$ un cylindre de sécurité pour (E).

Notons $\mathcal{F} = \mathcal{C}([t_0 - T, t_0 + T], \overline{B}(y_0, r_0))$ l'ensemble des applications continues de $[t_0 - T, t_0 + T]$ dans $\overline{B}(y_0, r_0)$, muni de la distance d de la convergence uniforme.

A toute fonction $y \in \mathcal{F}$, associons la fonction $\phi(y)$ définie par

$$\phi(y)(t) = y_0 + \int_{t_0}^t f(u, y(u)) du, \quad t \in [t_0 - T, t_0 + T].$$

D'après le lemme du § 2.1, y est une solution de (E) si et seulement si y est un point fixe de ϕ. On va donc essayer d'appliquer le théorème du point fixe. Observons que

$$\|\phi(y)(t) - y_0\| = \Big\| \int_{t_0}^{t} f(u, y(u))du \Big\| \leq M|t - t_0| \leq MT \leq r_0,$$

donc $\phi(y) \in \mathcal{F}$. L'opérateur ϕ envoie donc $\mathcal{F}$ dans $\mathcal{F}$. Soient maintenant $y, z \in F$ et $y_{(p)} = \phi^p(y)$, $z_{(p)} = \phi^p(z)$. On a

$$\begin{aligned}\|y_{(1)}(t) - z_{(1)}(t)\| &= \Big\| \int_{t_0}^{t} (f(u, y(u)) - f(u, z(u)))\, du \Big\| \\ &\leq \Big| \int_{t_0}^{t} k\|y(u) - z(u)\|du \Big| \leq k|t - t_0|\, d(y, z).\end{aligned}$$

De même

$$\begin{aligned}\|y_{(2)}(t) - z_{(2)}(t)\| &\leq \Big| \int_{0}^{t} k\|y_1(u) - z_1(u)\|\ du \Big| \\ &\leq \Big| \int_{t_0}^{t} k \cdot k|u - t_0|d(y, z)du \Big| = k^2\, \frac{|t - t_0|^2}{2}\ d(y, z).\end{aligned}$$

Par récurrence sur p, on vérifie aussitôt que

$$\|y_{(p)}(t) - z_{(p)}(t)\| \leq k^p\, \frac{|t - t_0|^p}{p!}\ d(y, z),$$

en particulier

$$d(\phi^p(y), \phi^p(z)) = d(y_{(p)}, z_{(p)}) \leq \frac{k^p T^p}{p!}\ d(y, z) \qquad (*)$$

et ϕ^p est lipschitzienne de rapport $\frac{k^p T^p}{p!}$ sur $\mathcal{F}$. Comme $\lim_{p\to+\infty} \frac{k^p T^p}{p!} = 0$, il existe p assez grand tel que $\frac{k^p T^p}{p!} < 1$; pour une telle valeur de p, ϕ^p est une application contractante de $\mathcal{F}$ dans $\mathcal{F}$. Par ailleurs, $\mathcal{F}$ est un espace métrique complet. Le théorème du point fixe démontré au chapitre IV (dans sa version généralisée au cas d'applications dont une itérée est contractante) montre alors que ϕ admet un point fixe unique y. Nous avons donc bien redémontré le théorème de Cauchy-Lipschitz affirmant l'existence et d'unicité de la solution du problème de Cauchy. ■

Remarque – D'après $(*)$, on voit que pour toute fonction $z \in \mathcal{F}$ la suite itérée $z_{(p)} = \varphi^p(z)$ converge uniformément vers la solution exacte y du problème de Cauchy.

3.3. Unicité globale

Le théorème d'unicité locale entraîne facilement un résultat d'unicité globale, au moyen d'un « raisonnement de connexité ».

Théorème – *Soient $y_{(1)}$, $y_{(2)} : I \to \mathbb{R}^m$ deux solutions de (E), avec f localement lipschitzienne en y. Si $y_{(1)}$ et $y_{(2)}$ coïncident en un point de I, alors $y_{(1)} = y_{(2)}$ sur I.*

Démonstration. Supposons $y_{(1)}(t_0) = y_{(2)}(t_0)$ en un point $t_0 \in I$. Montrons par exemple que $y_{(1)}(t) = y_{(2)}(t)$ pour $t \geq t_0$. S'il n'en est pas ainsi, considérons le premier instant $\widetilde{t}_0$ où $y_{(1)}$ et $y_{(2)}$ bifurquent :

$$\widetilde{t}_0 = \inf\{t \in I\,;\ t \geq t_0 \quad \text{et} \quad y_{(1)}(t) \neq y_{(2)}(t)\}$$

On a par définition $y_{(1)}(t) = y_{(2)}(t)$ pour $t \in [t_0, \widetilde{t}_0[$ et par continuité il s'ensuit que $y_{(1)}(\widetilde{t}_0) = y_{(2)}(\widetilde{t}_0)$. Soit $\widetilde{y}_0$ ce point et soit $\widetilde{C} = [\widetilde{t}_0 - \widetilde{T}, \widetilde{t}_0 + \widetilde{T}] \times \overline{B}(\widetilde{y}_0, \widetilde{r}_0)$ un cylindre de sécurité de centre $(\widetilde{t}_0, \widetilde{y}_0)$. Le théorème d'unicité locale implique que $y_{(1)} = y_{(2)}$ sur $[\widetilde{t}_0 - \widetilde{T}, \widetilde{t}_0 + \widetilde{T}]$, ce qui contredit la définition de $\widetilde{t}_0$. L'unicité est démontrée. ■

Corollaire – *Si f est localement lipschitzienne en y sur U, pour tout point $(t_0, y_0) \in U$ il passe une solution maximale $y : I \to \mathbb{R}^m$ et une seule.*

Interprétation géométrique – *Le théorème d'unicité signifie géométriquement que des courbes intégrales distinctes ne peuvent se couper.*

Exemple – $y' = 3|y|^{2/3}$ sur $U = \mathbb{R} \times \mathbb{R}$.

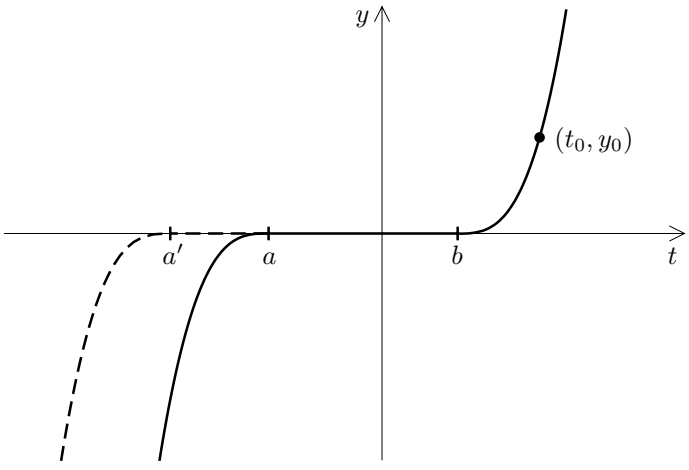
Déterminons l'ensemble des solutions maximales. On a ici $f(t, y) = 3|y|^{2/3}$, $\frac{\partial f}{\partial y} = \text{signe}\,(y) \times 2|y|^{-1/3}$ pour $y \neq 0$. La dérivée $y \neq 0$ la dérivée $\frac{\partial f}{\partial y}$ est continue sur les demi-plans $y > 0$ et $y < 0$, mais discontinue en $y = 0$. La fonction f est localement lipschitzienne en y sur $\{y > 0\}$ et $\{y < 0\}$, mais il est facile de voir qu'elle ne l'est pas au voisinage de tout point $(t_0, 0) \in \mathbb{R} \times \{0\}$ (on a vu d'ailleurs qu'il n'y a pas d'unicité locale en ces points). Sur $\{y > 0\}$ (resp. sur $\{y < 0\}$) l'équation équivaut à

$$\frac{1}{3}\, y' y^{-\frac{2}{3}} = 1 \quad (\text{resp.} \quad -\frac{1}{3}\, y'(-y)^{-\frac{2}{3}} = -1)$$

d'où $y^{\frac{1}{3}} = t + C_1$ (resp. $(-y)^{-\frac{1}{3}} = -(t + C_2)$) soit $y(t) = (t + C_i)^3$. Si y est une solution maximale dans $U = \mathbb{R} \times \mathbb{R}$, alors $y' \geq 0$, donc y est croissante. Notons

$$a = \inf\{t, y(t) = 0\}, \quad b = \sup\{t\,;\ y(t) = 0\}.$$

Si $a \neq -\infty$, on a $y(a) = 0$ et $y(t) < 0$ pour $t < a$, donc $y(t) = (t - a)^3$. De même $y(t) = (t - b)^3$ pour $t > b$ si $b \neq +\infty$.



On voit que pour tout point (t_0, y_0) il passe une infinité de solutions maximales : si $y_0 > 0$, $b = t_0 - y_0^{1/3}$ est imposé, mais le choix de $a \in [-\infty, b]$ est arbitraire. Noter que ce phénomène se produit bien qu'on ait unicité locale au point (t_0, y_0) !

3.4. Conditions suffisantes d'existence de solutions globales

Nous donnons ici des conditions suffisantes d'existence pour les solutions globales, reposant sur des hypothèses de croissance de $f(t, y)$ lorsque $\|y\|$ tend vers $+\infty$. On peut cependant obtenir des conditions suffisantes nettement plus faibles (voir l'exercice (b) ci-dessous, ainsi que le problème 5.9).

Théorème – *Soit $f : U \to \mathbb{R}^m$ une application continue sur un ouvert produit $U = J \times \mathbb{R}^m$, où $J \subset \mathbb{R}$ est un intervalle ouvert. On fait l'une ou l'autre des deux hypothèses suivantes* :

(1) *Il existe une fonction continue $k : J \to \mathbb{R}_+$ telle que pour tout $t \in J$ fixé, l'application $y \mapsto f(t, y)$ soit lipschitzienne de rapport $k(t)$ sur $\mathbb{R}^m$.*

(2) *Il existe des fonctions $c,\ k : J \to \mathbb{R}_+$ continues telles que l'application $y \mapsto f(t, y)$ satisfasse une croissance linéaire à l'infini du type*

$$\|f(t, y)\| \leq c(t) + k(t)\|y\|.$$

Alors toute solution maximale de l'équation différentielle $y' = f(t, y)$ est globale (c'est-à-dire définie sur J tout entier).

Démonstration. Il est évident que l'hypothèse (1) entraîne l'hypothèse (2) (avec $c(t) = \|f(t, 0)\|$), il suffirait donc de donner la preuve pour (2). Cependant, il y a une démonstration sensiblement plus simple sous l'hypothèse (1).

Démonstration sous l'hypothèse (1). Soit $(t_0, y_0) \in J \times \mathbb{R}^m$, et $[t_0 - T, t_0 + T']$ un intervalle compact quelconque contenu dans J. Reprenons la démonstration du théorème de Cauchy-Lipschitz.

Comme $U = J \times \mathbb{R}^m$, on peut choisir un cylindre de sécurité de rayon $r_0 = +\infty$. L'application ϕ définie au § 3.2 opère donc sur l'espace complet

$$\mathcal{F} = \mathcal{C}([t_0 - T, t_0 + T'], \mathbb{R}^m).$$

Soit

$$K = \max_{t \in [t_0 - T, t_0 + T']} k(t).$$

L'application f est par hypothèse K-lipschitzienne en y sur $[t_0 - T, t_0 + T'] \times \mathbb{R}^m$. D'après le raisonnement du § 3.2, l'application ϕ^p est lipschitzienne de rapport $\frac{1}{p!}\ K^p(\max(T, T'))^p$ sur $\mathcal{F}$, donc contractante pour p assez grand. Ceci implique que la solution (unique) du problème de Cauchy est définie sur tout intervalle $[t_0 - T, t_0 + T'] \subset J$. ∎

Démonstration sous l'hypothèse (2). L'idée est d'utiliser le critère de maximalité des solutions démontré au 2.6. Supposons qu'on ait une solution $y : [t_0, b[\to \mathbb{R}^m$ avec $t_0, b \in J$ (autrement dit, telle que b *ne soit pas* la borne supérieure de J). Posons $C = \sup_{t \in [t_0, b]} c(t)$ et $K = \sup_{t \in [t_0, b]} k(t)$. Nous obtenons

$$\|y'(t)\| = \|f(t, y(t))\| \leq C + K\|y(t)\|.$$

On utilise alors un raisonnement de type lemme de Gronwall pour majorer la norme $\|y(t)\|$. Nous avons $y(t) = y(t_0) + \int_{t_0}^t y'(u)\, du$, donc

$$\|y(t)\| \leq v(t) = \|y(t_0)\| + \int_{t_0}^t \|y'(u)\|\, du \qquad \text{avec}$$
$$v'(t) = \|y'(t)\| \leq C + K\|y(t)\| \leq C + Kv(t).$$

Ceci donne la majoration

$$\frac{d}{dt}\big(v(t)e^{-K(t-t_0)}\big) = \big(v'(t) - K\, v(t)\big)e^{-K(t-t_0)} \leq Ce^{-K(t-t_0)}.$$

Par intégration sur $[t_0, t]$, on obtient

$$v(t)e^{-K(t-t_0)} - v(t_0) \leq \frac{C}{K}(1 - e^{-K(t-t_0)}),$$

et comme $v(t_0) = \|y(t_0)\|$, il vient

$$\sup_{t \in [t_0, b[} \|y(t)\| \leq \sup_{t \in [t_0, b[} v(t) \leq R = \frac{C}{K}\big(e^{K(b-t_0)} - 1\big) + \|y(t_0)\|e^{K(b-t_0)}.$$

Par conséquent $(t, y(t))$ décrit une partie compacte $K = [t_0, b] \times \overline{B}(0, R)$ dans $U = J \times \mathbb{R}^m$, et y ne peut être une solution maximale. Toute solution maximale est donc globale. [Le lecteur pourra étudier l'exercice 5.9 pour un généralisation à une hypothèse de croissance plus faible que (2), tenant compte uniquement de la « direction radiale » du vecteur $f(t, y)$]. ∎

Exercices

(a) *Montrer que toute solution maximale de l'équation différentielle* $y' = t\sqrt{t^2+y^2}$, $(t,y) \in \mathbb{R}\times\mathbb{R}$, *est globale.*

(b) *On définit* $f : \mathbb{R} \to \mathbb{R}$ *par* $f(y) = e$ *si* $y \le e$ *et* $f(y) = y \ln y$ *si* $y \ge e$. *Montrer que* f *n'est pas lipschitzienne au voisinage de* 0. *Déterminer explicitement les solutions maximales de l'équation* $y' = f(y)$. *Les conditions suffisantes du théorème précédent sont-elles nécessaires ?*

4. ÉQUATIONS DIFFÉRENTIELLES D'ORDRE SUPÉRIEUR À UN

4.1. DÉFINITIONS

Un système différentiel d'ordre p dans $\mathbb{R}^m$ est une équation de la forme

$$\text{(E)} \qquad y^{(p)} = f(t, y, y', \dots, y^{(p-1)})$$

où $f : U \to \mathbb{R}^m$ est une application continue définie sur un ouvert $U \subset \mathbb{R}\times(\mathbb{R}^m)^p$.

Une *solution* de (E) sur un intervalle $I \subset \mathbb{R}$ est une application $y : I \to \mathbb{R}^m$ p-fois dérivable, telle que

(i) $(\forall t \in I) \quad (t, y(t), y'(t), \dots, y^{(p-1)}(t)) \in U$,

(ii) $(\forall t \in I) \quad y^{(p)}(t) = f(t, y(t), y(t'), \dots, y^{(p-1)}(t))$.

Le résultat suivant se démontre par récurrence d'une manière entièrement analogue à celle utilisée pour les équations différentielles d'ordre 1. Le détail de l'argument est laissé au lecteur.

Régularité des solutions – Si f est de classe C^k, les solutions y sont de classe C^{k+p}.

4.2. SYSTÈME DIFFÉRENTIEL D'ORDRE UN ASSOCIÉ

Il est clair que le système (E) est équivalent au système différentiel d'ordre 1

$$\text{(E}_1\text{)} \qquad \begin{cases} \frac{dY_0}{dt} = Y_1 \\ \frac{dY_1}{dt} = Y_2 \\ \qquad \dots \\ \frac{dY_{p-2}}{dt} = Y_{p-1} \\ \frac{dY_{p-1}}{dt} = f(t, Y_0, Y_1, \dots, Y_{p-1}) \end{cases}$$

si l'on pose $Y_0 = y$, $Y_1 = y'$, ... Le système (E$_1$) peut encore s'écrire

$$\text{(E}_1\text{)} \qquad Y' = F(T, Y)$$

avec

$$\begin{aligned} &Y = (Y_0, Y_1, \dots, Y_{p-1}) \in (\mathbb{R}^m)^p \\ &F = (F_0, F_1, \dots, F_{p-1}) : U \to (\mathbb{R}^m)^p \\ &F_0(t,Y) = Y_1, \dots, F_{p-2}(t,Y) = Y_{p-1}, \\ &F_{p-1}(t,Y) = f(t,Y). \end{aligned}$$

Tout système différentiel (E) d'ordre p dans $\mathbb{R}^m$ est donc équivalent à un système différentiel (E_1) d'ordre 1 dans $(\mathbb{R}^m)^p$. Il en résulte que les théorèmes d'existence et d'unicité démontrés pour les systèmes d'ordre 1 sont encore vrais pour les systèmes d'ordre p, avec des preuves qui sont des transpositions directes du cas d'ordre 1. En voici les principaux énoncés :

4.3. Théorème d'existence

Pour tout point $(t_0, y_0, y_1, \ldots, y_{p-1}) \in U$ le problème de Cauchy de conditions initiales

$$y(t_0) = y_0, \quad y'(t_0) = y_1, \ldots, y^{(p-1)}(t_0) = y_{p-1}$$

admet au moins une solution maximale $y : I \to \mathbb{R}^m$, définie sur un intervalle ouvert.

Remarque très importante – On voit ainsi que pour un système d'ordre p, la condition initiale requiert non seulement la donnée de la valeur y_0 de y au temps t_0, mais également la donnée de ses $(p-1)$ premières dérivées.

4.4. Théorème d'existence et d'unicité

Si de plus f est localement lipschitzienne en $(y_0, \ldots, y_{p-1})$ sur U, c'est-à-dire si $\forall (t_0, y_0, \ldots, y_{p-1}) \in U$ il existe un voisinage $[t_0 - T_0, t_0 + T_0] \times \overline{B}(y_0, r_0) \times \ldots \times \overline{B}(y_{p-1}, r_{p-1})$ contenu dans U sur lequel

$$\|f(t, z_0, \ldots, z_{p-1}) - f(t, w_0, \ldots, w_{p-1})\| \leq k(\|z_0 - w_0\| + \ldots + \|z_{p-1} - w_{p-1}\|),$$

alors le problème de Cauchy 4.3 admet une solution maximale et une seule.

4.5. Solutions globales

Si $U = J \times (\mathbb{R}^m)^p$ et s'il existe une fonction $k : J \to \mathbb{R}_+$ continue telle que $(\forall t \in J)$

$$\|f(t, z_0, \ldots, z_{p-1}) - f(t, w_0, \ldots, w_{p-1})\| \leq k(t)(\|z_0 - w_0\| + \ldots + \|z_{p-1} - w_{p-1}\|),$$

alors les solutions maximales sont définies sur J tout entier.

5. Problèmes

5.1. On considère l'équation différentielle $y' = y^2 - x$.

(a) Quelles sont les lignes isoclines ?

On notera I_0 l'isocline correspondant à la pente nulle.

Soit $\mathcal{P}^-$ l'ensemble des points du plan où la pente des solutions est strictement négative. Décrire $\mathcal{P}^-$. Montrer que si une solution entre dans $\mathcal{P}^-$, alors elle y

reste (c'est-à-dire : si une solution $y(x)$ a un point $(x_0, y(x_0))$ dans $\mathcal{P}^-$, alors si $x_1 > x_0$, $(x_1, y(x_1)) \in \mathcal{P}^-$).

(b) Étudier et tracer le graphe de la courbe $\mathcal{I}$ ensemble des points d'inflexion des solutions de l'équation différentielle. Quelles sont les régions du plan où $y'' > 0$, respectivement $y'' < 0$?

On notera $\mathcal{I}_1$ la partie de $\mathcal{I}$ extérieure à $\mathcal{P}^-$, et $\mathcal{I}_2$ la partie de $\mathcal{I}$ qui se trouve dans $\mathcal{P}^-$.

(c) Soit $\mathcal{C}$ une courbe solution rencontrant $\mathcal{I}_1$ en un point (x, y).

(α) Montrer qu'en ce point, la pente de $\mathcal{I}_1$ est strictement inférieure à la pente de $\mathcal{C}$.

(β) En déduire que $\mathcal{C}$ ne coupe $\mathcal{I}_1$ qu'en ce point, que $\mathcal{C}$ ne rencontre pas $\mathcal{P}^-$, et que $\mathcal{C}$ n'a qu'un point d'inflexion.

(γ) Montrer que $\mathcal{C}$ possède 2 branches infinies à direction asymptotique verticale.

(δ) Soit (x_0, y_0) un point de $\mathcal{C}$. Comparer en ce point, la pente de $\mathcal{C}$ et la pente de la solution de l'équation différentielle $y' = \frac{y^2}{2}$. En déduire que les branches infinies de $\mathcal{C}$ correspondent à des asymptotes verticales.

(d) Soit $\mathcal{D}$ une courbe solution rencontrant I_0.

(α) Montrer que $\mathcal{D}$ possède une asymptote verticale.

(β) Montrer que $\mathcal{D}$ a un point d'inflexion et un seul.

(γ) Montrer que lorsque $x \to \infty$, $\mathcal{D}$ est asymptote à I_0.

(e) Soit A (resp. B) l'ensemble des points de l'axe Oy par où passe une courbe solution qui rencontre $\mathcal{I}_1$ (resp. $\mathcal{I}_0$).

(α) Montrer qu'il existe a tel que $A = \{0\} \times]a, +\infty[$.

(β) Montrer qu'il existe b tel que $B = \{0\} \times]-\infty, b[$.

(γ) Montrer que $a = b$. Quelle est l'allure de la solution passant par le point de coordonnées $(0, a)$?

5.2. On considère l'équation différentielle $y' = f(t, y)$, où f et $\frac{\partial f}{\partial y}$ sont continues. Soit α une fonction réelle définie sur un intervalle $[t_0, t_1[$ où t_1 peut éventuellement être infini ; on suppose α continue et dérivable par morceaux.

On dit que α est une barrière inférieure [respectivement : supérieure] pour l'équation différentielle si $\alpha'(t) < f(t, \alpha(t))$ [resp : $\alpha'(t) > f(t, \alpha(t))$] pour tout t tel que $\alpha'(t)$ existe, et, aux points où α n'est pas dérivable, pour la dérivée à gauche et pour la dérivée à droite.

(a) Montrer que si α est une barrière inférieure pour $t_0 \leq t \leq t_1$ et si u est une solution de l'équation différentielle vérifiant $\alpha(t_0) \leq u(t_0)$, alors $\alpha(t) < u(t)$ pour tout $t \in]t_0, t_1[$. Montrer un résultat analogue pour une barrière supérieure.

(b) On suppose que α est une barrière inférieure sur $[t_0, t_1[$, que β est une barrière supérieure sur $[t_0, t_1[$, et que $\alpha(t) < \beta(t)$ pour tout $t \in [t_0, t_1[$. L'ensemble des points (t, x) tels que $t_0 \leq t \leq t_1$ et $\alpha(t) \leq x \leq \beta(t)$ est appelé entonnoir.

(α) Montrer que si une solution u de l'équation différentielle est telle que $(s, u(s))$ soit dans l'entonnoir pour un $s \in [t_0, t_1[$, alors $(t, u(t))$ est dans l'entonnoir pour tout $t \in [s, t_1[$.

(β) Si α est une barrière inférieure et β une barrière supérieure, et si $\alpha(t) > \beta(t)$ pour $t \in [t_0, t_1[$, on dit que l'ensemble des (t, x) tels que $t_0 \leq t \leq t_1$ et $\alpha(t) \geq x \geq \beta(t)$ est un anti-entonnoir.

Montrer qu'il existe une solution $u(t)$ de l'équation différentielle, telle que $\beta(t) \leq u(t) \leq \alpha(t)$ pour tout $t \in [t_0, t_1[$.

(c) Dans la suite du problème, on prend $f(t, y) = \sin(ty)$. On se restreindra aux solutions vérifiant $y > 0$.

(α) Déterminer les isoclines correspondant aux pentes -1, 0, 1.

(β) Pour quelles valeurs de t ces isoclines sont-elles des barrières inférieures ? supérieures ? Quels sont les entonnoirs formés par ces isoclines ?

(γ) Soit u une solution de l'équation différentielle ; soit γ la fonction continue, dérivable par morceaux, définie pour $t \geq 0$ par : $\gamma(0) = u(0) > 0$; γ est affine de pente 1 depuis $t = 0$ jusqu'à ce que son graphe rencontre la première isocline de pente 0, puis γ est affine de pente 0 jusqu'à l'isocline de pente 0 suivante, puis γ est affine de pente 1 jusqu'à l'isocline de pente 0 suivante, et ainsi de suite. Montrer que le graphe de γ rencontre la droite $y = t$.

(δ) Montrer que γ est une barrière supérieure.

(ε) En déduire que toute solution de l'équation différentielle rencontre la droite $y = t$, puis reste dans un entonnoir.

(ζ) Dessiner l'allure des solutions de l'équation différentielle $y' = \sin(ty)$.

5.3. On considère l'équation (appelée équation de Van der Pol) :

$$\text{(E)} \qquad \begin{cases} x'(t) = y(t) - x^3(t) + x(t), \\ y'(t) = -x(t), \end{cases} \quad t \in \mathbb{R}.$$

(a) Montrer que le problème de Cauchy correspondant admet une solution globale unique (on pourra utiliser le résultat de l'exercice 5.9).

(b) On appelle trajectoire associée à une solution de (E), l'ensemble parcouru dans le plan Euclidien par le point de coordonnées $(x(t), y(t))$ lorsque t parcourt $\mathbb{R}$. Montrer que les trajectoires associées à deux solutions distinctes de (E) coïncident ou n'ont aucun point commun ; montrer que par chaque point du plan passe une trajectoire et une seule ; montrer que si une trajectoire a un point double (c'est-à-dire correspondant à deux valeurs distinctes de t), les solutions associées de (E) sont périodiques (et tous les points sont alors doubles). Quelles sont les trajectoires réduites à un point ?

(c) Montrer que la courbe symétrique d'une trajectoire par rapport à $(0,0)$ est encore une trajectoire.

(d) On considère maintenant les sous-ensembles du plan

$$\begin{aligned}
&D^+ = \{(0,y)\,;\ y>0\}; \quad D^- = \{(0,y)\,;\ y<0\}\ ;\\
&E_1 = \{(x,y)\,;\ x>0 \quad \text{et} \quad y>x^3-x)\}; \quad \Gamma_+ = \{(x,x^3-x)\,;\ x>0)\}\ ;\\
&E_2 = \{(x,y)\,;\ x>0 \quad \text{et} \quad y<x^3-x\}\ ;\\
&E_3 = \{(x,y)\,;\ x<0 \quad \text{et} \quad y<x^3-x\}; \quad \Gamma_- = \{(x,x^3-x)\,;\ x<0\}\ ;\\
&E_4 = \{(x,y)\,;\ x<0 \quad \text{et} \quad y>x^3-x\}.
\end{aligned}$$

Soit $(x(t),y(t))$ une solution de (E) ; montrer que, si $(x(t_0),y(t_0)) \in D^+$, il existe $t_4 > t_3 > t_2 > t_1 > t_0$ tels que $(x(t),y(t)) \in E_i$ pour $t \in]t_{i-1},t_i[$, $i=1,2,3,4$, et $(x(t_1),y(t_1)) \in \Gamma^+$, $(x(t_2),y(t_2)) \in D^-$, $(x(t_3),y(t_3)) \in \Gamma^-$; $(x(t_4),y(t_4)) \in D^+$.

(e) Soit $y_0 > 0$ et $t_0 \in \mathbb{R}$; il existe une solution de (E) telle que $(x(t_0),y(t_0)) = (0,y_0)$; on pose $\sigma(y_0) = y(t_2)$; montrer que $\sigma(y_0)$ ne dépend que de y_0 (et non de t_0) et que σ est une application monotone continue de $\mathbb{R}^+$ dans $\mathbb{R}^-$.

(f) En utilisant le (c), montrer que $(0,y_0)$ appartient à la trajectoire d'une solution périodique si et seulement si $\sigma(y_0) = -y_0$.

(g) Soit $\beta > 0$ tel que pour la solution de (E) vérifiant $(x(t_0),y(t_0)) = (0,\beta)$ on ait $(x(t_1),y(t_1)) = (1,0)$. Montrer que pour $y_0 < \beta$, on a $\sigma(y_0)^2 - y_0^2 > 0$ (regarder $\displaystyle\int_{t_0}^{t_2} \frac{d}{dt}\,[x(t)^2+y(t)^2]dt$).

(h) Soit y_0 grand. Soit C la courbe formée des arcs suivants :

- le segment $(0,y_0)$, $(1,y_0)$;
- l'arc de cercle de centre O passant par $(1,y_0)$ et coupant $(y=x^3-x)$ en (x_1,y_1) avec $x_1>1$.
- le segment (x_1,y_1), $(x_1,0)$.
- l'arc de cercle de centre O passant par $(x_1,0)$ et coupant $(x=1)$ en (x_1',y_1').
- la tangente en (x_1',y_1') à cet arc de cercle qui recoupe Oy en $(0,y_2)$.

Montrer que la solution de (E) passant par $(0,y_0)$ est à l'intérieur de C. En déduire que $\sigma(y_0)^2 - y_0^2 < 0$.

(i) En déduire qu'il existe une trajectoire et une seule correspondant à des solutions périodiques de (E). Montrer que les trajectoires non réduites à $(0,0)$ convergent asymptotiquement vers cette trajectoire quand t tend vers $+\infty$.

5.4. Soit t une variable réelle ≥ 0. On considère le problème de Cauchy

$$y' = ty, \quad y(0) = 1.$$

(a) Démontrer que pour tout $T > 0$, ce problème admet une solution et une seule sur $[0, T]$, et indiquer comment la méthode d'Euler permet d'en trouver une approximation.

(b) Déduire de ce qui précède la formule

$$y(t) = \lim_{N \to +\infty} P_N(t) \quad \text{avec} \quad P_N(t) = \prod_{n=0}^{N-1} \left(1 + \frac{nt^2}{N^2}\right)$$

(c) Pour $\alpha > 0$, étudier les variations de la fonction $f(x) = x \ln(1 + \alpha/x)$ sur $]0, +\infty[$; on montrera que $f''(x) < 0$.

En déduire l'encadrement

$$\left(1 + \frac{t^2}{N}\right)^{\frac{n}{N}} \leq 1 + \frac{nt^2}{N^2} \leq \left(1 + \frac{t^2}{N^2}\right)^n \quad \text{si } 0 \leq n \leq N - 1.$$

(d) Calculer la limite du (b), et en déduire $y(t)$.

5.5. On considère l'équation différentielle

$$y' = |y|^{-3/4} y + t \sin\left(\frac{\pi}{t}\right) = f(t, y)$$

où le second membre est défini sur $\mathbb{R}^2$ à l'aide de prolongements par continuité. On note $Y(t)$ la solution approchée définie sur $\mathbb{R}$, obtenue par la méthode d'Euler pour le pas $h = \frac{1}{n+1/2}$ où $n \in \mathbb{N}^*$, et vérifiant $Y(0) = 0$. On suppose dans un premier temps que n est pair.

(a) Calculer $Y(h)$, $Y(2h)$ et $Y(3h)$.

Démontrer les inégalités $Y(3h) > \frac{h^{3/2}}{2} > \frac{(3h)^{3/2}}{16}$.

(b) Déterminer $c > 0$ tel que $0 < t < c$ on ait $\frac{1}{2}\, t^{3/8} - t > \frac{1}{10}\, t^{3/8}$. En supposant de plus $h \leq t$ et c assez petit vérifier $\frac{(t+h)^{3/2} - t^{3/2}}{h} < \frac{8}{5}\, t^{3/8}$ (on pourra utiliser la formule de Taylor).

(c) On suppose que pour $m \in \mathbb{N}^*$ on a $mh < c$ et $Y(m, h) > \frac{(mh)^{3/2}}{16}$.

Démontrer les inégalités

$$f(mh, Y(mh)) > Y(mh)^{1/4} - mh > \frac{1}{2}\, (mh)^{3/8} - mh > \frac{1}{10}\, (mh)^{3/8}.$$

En déduire $Y((m+1)h) > \frac{((m+1)h)^{3/2}}{16}$.

Montrer que si p entier vérifie $0 < ph \leq c$, on a

$$Y(ph) > \frac{(ph)^{3/2}}{16}.$$

(d) On suppose ici que n est impair. Calculer $Y(h)$, $Y(2h)$ et $Y(3h)$. Montrer l'inégalité $Y(3h) < -\frac{(3h)^{3/2}}{16}$.

On suppose que pour $mh < c$ on a $Y(mh) < -\frac{(mh)^{3/2}}{16}$; montrer comme ci-dessus que $Y((m+1)h) < -\frac{((m+1)h)^{3/2}}{16}$, puis que $Y(ph) < -\frac{(ph)^{3/2}}{16}$ pour tout entier p tel que $0 < ph \leq c$.

(e) Pour $0 < t < c$, montrer que les solutions approchées $Y(t)$ ne tendent vers aucune limite n tend vers $+\infty$.

5.6. Soit le système différentiel dans $\mathbb{R}^2$ défini par

$$\text{(S)} \qquad \begin{cases} \dfrac{dx}{dt} = 2(x - ty) \\ \dfrac{dy}{dt} = 2y. \end{cases}$$

(a) Déterminer la courbe intégrale qui passe par le point (x_0, y_0) au temps $t = 0$.

(b) On utilise la méthode d'Euler avec pas constant h, démarrant au temps $t_0 = 0$. Soit (x_n, y_n) le point atteint au temps $t_n = nh$ $(n \in \mathbb{N})$.

(α) Écrire la relation qui lie (x_{n+1}, y_{n+1}) à (x_n, y_n).

(β) Calculer explicitement (x_n, y_n) en fonction de n, h, x_0, y_0.

(γ) Sans utiliser les théorèmes généraux du cours, vérifier que la solution approchée qui interpole linéairement les points (x_n, y_n) converge sur $\mathbb{R}_+$ vers la solution exacte de (S).

5.7. Soit $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ une fonction continue et lipschitzienne de rapport k en sa deuxième variable. On définit une suite de fonctions $y_n : [a, b] \to \mathbb{R}$ en posant $y_0(t) = \lambda$ et

$$y_{n+1}(t) = \lambda + \int_a^t f(u, y_n(u))du, \quad n \in \mathbb{N}.$$

On sait d'après V 3.2 que y_n converge uniformément vers la solution exacte de l'équation $y' = f(t, y)$ telle que $y(a) = \lambda$. On étudie ici le cas particulier de l'équation

$$\frac{dy}{dt} = -2y + t, \quad t \in [0, +\infty[.$$

(a) Montrer que y_n peut s'écrire sous la forme

$$y_n(t) = \lambda P_n(t) + Q_n(t)$$

où P_n, Q_n sont des polynômes que l'on explicitera.

(b) Calculer $\lim_{n\to+\infty} P_n$ et $\lim_{n\to+\infty} Q_n$. Vérifier ce résultat en résolvant directement l'équation.

5.8. Soit T un réel positif et $f : [0,T] \times \mathbb{R} \to \mathbb{R}$ une application continue lipschitzienne de rapport k en la deuxième variable. On considère l'équation différentielle

$$\text{(E)} \qquad y' = f(t,y).$$

Soit un réel $h \in]0,T[$. On dira que z est une solution retardée de retard h si z est une fonction continue sur $[0,T]$, dérivable sur $]h,T]$ et si

$$z'(t) = f(t, z(t-h)), \quad \forall t \in]h,T].$$

(a) Soit y_0 un réel fixé. Montrer que (E) admet une solution retardée de retard h et une seule, notée z_h, telle que $z_h(t) = y_0$ pour tout $t \in [0,h]$.

(b) Soit z une solution retardée de retard h. On pose

$$A = \max_{t\in[0,T]} |f(t,0)|, \qquad m(t) = \max_{u\in[0,t]} |z(u)|.$$

(α) Montrer que pour tout $t \in [h,T]$ on a

$$m(t) \le m(h) + \int_h^t (A + km(u))du.$$

(β) En déduire que

$$m(t) \le \Big(\frac{A}{k} + m(h)\Big)e^{k(t-h)} - \frac{A}{k}, \quad \forall t \in [h,T].$$

[*Indication* : étudier la dérivée de la fonction $M(t) = e^{-kt}\int_h^t (A + km(u))du$.]

(γ) Montrer qu'il existe une constante B indépendante de h, que l'on explicitera, telle que $\|z_h\|_\infty \le B$ pour tout $h > 0$, si z_h désigne la solution retardée du (a).

(c) On se propose ici d'étudier la convergence de z_h quand h tend vers 0.

(α) Montrer que les fonctions z_h sont C-lipschitziennes avec une constante C indépendante de h.

(β) Soit y la solution exacte (non retardée) de (E) telle que $y(0) = y_0$. On pose

$$\delta(t) = \max_{u\in[0,t]} |z_h(u) - y(u)|.$$

Montrer que δ vérifie l'inégalité intégrale

$$\delta(t) \le \delta(h) + \int_h^t (kCh + k\delta(u))du.$$

où C est la constante de la question (c) α).

(γ) En déduire une majoration de $\|\delta\|_\infty$ et conclure.

(d) On construit maintenant une méthode de résolution approchée de (E) utilisant les solutions retardées z_h. Pour tout entier $n \in \mathbb{N}$, $n \leq T/h$, on pose

$$t_n = nh, \quad z_n = z_h(t_n) \ ;$$

dans la formule

$$z_{n+1} = z_n + \int_{t_n}^{t_{n+1}} f(t, z_h(t-h))dt$$

on remplace la valeur exacte de l'intégrale par sa valeur approchée calculée au moyen de la méthode des trapèzes élémentaires.

(α) Écrire la relation de récurrence définissant la suite (z_n).

(β) Exprimer l'erreur de consistance relative à une solution exacte y ; en calculer un développement limité à l'ordre 2 en fonction de h et des dérivées partielles de f au point (t, y). Quel est l'ordre de la méthode ? (voir chapitre VIII pour les définitions).

5.9. Soit J un intervalle ouvert de $\mathbb{R}$ et $f : J \times \mathbb{R}^m \to \mathbb{R}^m$ une application continue. On se propose de démontrer que toute solution maximale de l'équation différentielle $y' = f(t, y)$ est globale si f vérifie l'hypothèse suivante :

(H) Il existe des fonctions $a, b : I \to \mathbb{R}_+$ continues telles que

$$\langle f(t,y), y\rangle \leq a(t)\|y\|^2 + b(t), \quad \forall (t,y) \in J \times \mathbb{R}^m,$$

où $\langle \ , \ \rangle$ et $\| \ \|$ désignent respectivement le produit scalaire et la norme euclidienne standards sur $\mathbb{R}^m$.

(a) Soit $y : [t_0, t_1[\to \mathbb{R}^m$ une solution maximale à droite passant par un point (t_0, y_0) et soit $r(t) = \|y(t)\|^2$. Montrer que $r'(t) \leq 2a(t)r(t) + 2b(t)$.

En déduire que $\|y(t)\|^2 \leq \rho(t)$ où $\rho : J \to \mathbb{R}$ est la solution (toujours globale) de l'équation linéaire $\rho' = 2a(t)\rho + 2b(t)$, telle que $\rho(t_0) = \|y_0\|^2$.

[*Indication :* soit $A(t)$ une primitive de $a(t)$; étudier le signe de la dérivée de $(r(t) - \rho(t))e^{-2A(t)}$.

(b) Déterminer un majorant explicite de $\|y(t)\|$ lorsque a et b sont des constantes.

(c) On suppose que $t_1 < \sup J$. Montrer que $y(t)$, $y'(t)$ sont bornées sur $[t_0, t_1[$ et que ces fonctions se prolongent par continuité en t_1. Montrer que ceci conduit à une contradiction. Conclure.

CHAPITRE VI

MÉTHODES DE RÉSOLUTION EXPLICITE DES ÉQUATIONS DIFFÉRENTIELLES

On se propose d'étudier un certain nombre de types classiques d'équations différentielles du premier et du second ordre pour lesquelles on sait ramener le calcul des solutions à des calculs de primitives. Ceci fournira l'occasion d'illustrer les résultats généraux du chapitre V par des exemples.

1. ÉQUATIONS DU PREMIER ORDRE

1.1. REMARQUES GÉNÉRALES

On considère une équation différentielle

$$\text{(E)} \qquad \frac{dy}{dx} = f(x, y)$$

où $f : U \to \mathbb{R}$ est une fonction continue sur un ouvert $U \subset \mathbb{R}^2$, localement lipschitzienne en y.

Les différentes solutions de l'équation (E) s'écrivent en général sous la forme

$$y = \varphi(x, \lambda)$$

où λ est un paramètre réel : on dit parfois que la solution « générale » dépend d'un seul paramètre. Pour comprendre ce phénomène, il suffit d'appliquer le théorème de Cauchy-Lipschitz : si on cherche les solutions définies au voisinage d'un point x_0, on sait qu'il existe une solution y et une seule telle que $y(x_0) = y_0$; on peut donc choisir $\lambda = y_0$ pour paramétrer les solutions. Dans la pratique, le paramètre λ *apparaît souvent comme constante d'intégration.*

Il arrive parfois qu'en plus de la solution générale on ait des solutions particulières $y = \psi_0(x)$, $y = \psi_1(x)$, ... qui ne s'obtiennent pour aucune valeur de λ : on dit que ce sont des *solutions singulières* (ou *courbes intégrales singulières*) de (E).

On va maintenant décrire une situation un peu plus générale qui se ramène au cas d'une équation du type considéré ci-dessus.

Systèmes différentiels autonomes dans un ouvert $U \subset \mathbb{R}^2$ – On suppose donné un champ de vecteurs dans U, c'est-à-dire une application continue

$$M \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \overrightarrow{V}(M) \begin{pmatrix} a(x,y) \\ b(x,y) \end{pmatrix}, \quad M \in U.$$

On appelle système autonome associé au champ de vecteurs $\overrightarrow{V}(M)$ le système différentiel

$$\text{(S)} \qquad \frac{d\overrightarrow{M}}{dt} = \overrightarrow{V}(M) \Leftrightarrow \begin{cases} \dfrac{dx}{dt} = a(x,y) \\ \dfrac{dy}{dt} = b(x,y) \end{cases}.$$

Si $\overrightarrow{V}(M)$ représente un champ de vecteurs vitesse (associé par exemple à l'écoulement d'une nappe de fluide sur une surface plane), résoudre (S) revient à chercher la trajectoire et la loi du mouvement des particules de fluide en fonction du temps. Le mot « autonome » signifie que le champ de vecteurs ne dépend pas du temps t (cas d'un écoulement stationnaire).

Si $t \mapsto M(t)$ est solution, toute fonction $t \mapsto M(t+T)$ obtenue par un décalage dans le temps est encore solution. Dans l'ouvert $U' = \{M(x,y)\,;\ a(x,y) \neq 0\}$ on a (S) $\Rightarrow$ (E) où

$$\text{(E)} \qquad \frac{dy}{dx} = \frac{b(x,y)}{a(x,y)} = f(x,y).$$

Résoudre (E) permet de trouver la trajectoire des particules (mais pas la loi du mouvement en fonction du temps).

1.2. Équations à variables séparées

Ce sont les équations dans lesquelles on peut regrouper x, dx d'une part et y, dy d'autre part. Nous allons examiner 3 cas.

a) Équations $y' = f(x)$, avec $f : I \to \mathbb{R}$ continue.

Les solutions sont données par

$$y(x) = F(x) + \lambda, \quad \lambda \in \mathbb{R},$$

où F est une primitive de f sur I. Les courbes intégrales se déduisent les unes des autres par translations dans la direction Oy.

b) Équations $y' = g(y)$, avec $g : J \to \mathbb{R}$ continue.

L'équation peut se récrire $\frac{dy}{dx} = g(y)$, ou encore $\frac{dy}{g(y)} = dx$ à condition que $g(y) \neq 0$.

- Notons y_j les racines de $g(y) = 0$ dans l'intervalle J. Alors $y(x) = y_j$ est une solution (singulière) évidente de l'équation.
- Dans l'ouvert $U = \{(x, y) \in \mathbb{R} \times J\,;\ g(y) \neq 0\}$, on a

$$(\mathrm{E}) \Leftrightarrow \frac{dy}{g(y)} = dx.$$

Les solutions sont données par

$$G(y) = x + \lambda, \quad \lambda \in \mathbb{R}$$

où G est une primitive quelconque de $\frac{1}{g}$ sur chacun des intervalles ouverts $[y_j, y_{j+1}[$ délimités par les racines de g. Dans chaque bande $\mathbb{R} \times]y_j, y_{j+1}[$, les courbes intégrales se déduisent les unes des autres par translations dans la direction Ox ; ceci est à relier au fait que les lignes isoclines sont les droites $y = m =$ constante.

Comme $G' = \frac{1}{g}$ et que g est de signe constant sur $]y_j, y_{j+1}[$, on en déduit que G est une application strictement monotone bijective

$$G :]y_j, y_{j+1}[\to]a_j, b_j[$$

avec $a_j \in [-\infty, +\infty[$, $b_j \in]-\infty, +\infty]$. On peut donc (au moins théoriquement) exprimer y en fonction de x :

$$y = G^{-1}(x + \lambda), \quad \lambda \in \mathbb{R}.$$

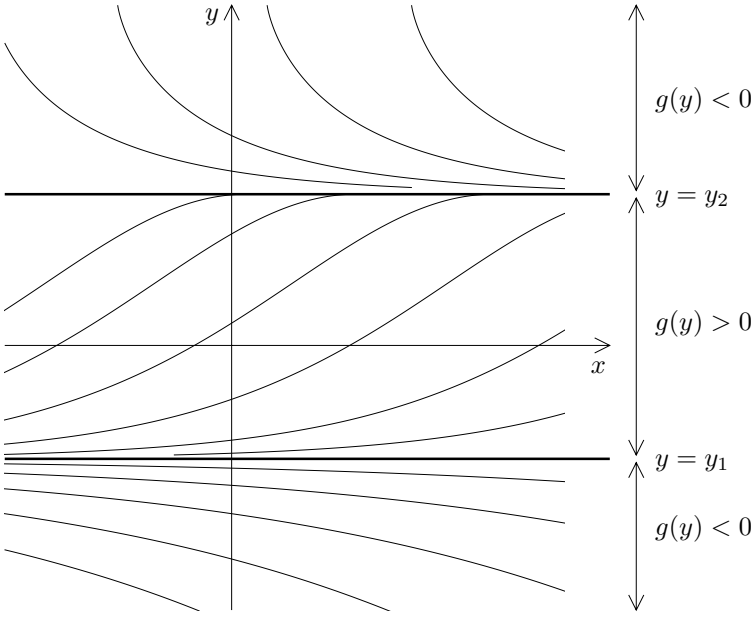
Supposons par exemple $g > 0$, et par suite G croissante sur $]y_j, y_{j+1}[$.

• Si $\int_{y_j}^{y_j+\varepsilon} \frac{dy}{g(y)}$ diverge, on a $a_j = -\infty$, par conséquent $x = G(y) - \lambda \to -\infty$ quand $y \to y_j + 0$. Dans ce cas, la courbe est asymptote à la droite $y = y_j$.

• Si $\int_{y_j}^{y_j+\varepsilon} \frac{dy}{g(y)}$ converge, alors $a_j \in \mathbb{R}$ et $x \to a_j - \lambda$ quand $y \to y_j + 0$, avec de plus $y' = g(y) \to 0$; la courbe vient rejoindre la droite $y = y_j$ au point $(a_j - \lambda, y_j)$ et admet la droite $y = y_j$ pour tangente en ce point. Cette situation montre qu'*il n'y a pas unicité* du problème de Cauchy en cas de convergence de l'intégrale.

Exercice – *Vérifier que $\int \frac{dy}{g(y)}$ est bien toujours divergente en tout point y_j tel que $g(y_j) = 0$, lorsque g est localement lipschitzienne.*

L'allure des courbes intégrales est la suivante (dans le schéma ci-dessous, on suppose qu'il y a convergence en $y_2 - 0$, divergence en $y_1 \pm 0$ et $y_2 + 0$) :



c) Cas général des équations à variables séparées :

(E) $\qquad y' = f(x)g(y)$ avec f, g continues.

• Si $g(y_j) = 0$, la fonction constante $y(x) = y_j$ est solution singulière.

• Sur l'ouvert $U = \{(x, y)\,;\ g(y) \neq 0\}$ on a

$$(\mathrm{E}) \Leftrightarrow \frac{dy}{g(y)} = f(x)dx$$

d'où $G(y) = F(x) + \lambda, \quad \lambda \in \mathbb{R}$, où F est une primitive de f et G une primitive de $1/g$. Comme G est continue strictement monotone sur chaque intervalle $[y_j, y_{j+1}[$, l'application G admet une application réciproque G^{-1} et on obtient

$$y = G^{-1}(F(x) + \lambda).$$

Exemple – Soit l'équation $y' = \sqrt{\dfrac{1-y^2}{1-x^2}}$. Le domaine de définition est la réunion

$$\{|x| < 1 \text{ et } |y| \leq 1\} \quad \cup \quad \{|x| > 1 \text{ et } |y| \geq 1\}.$$

On va donc se placer dans l'ouvert

$$U = \{|x| < 1 \text{ et } |y| < 1\} \quad \cup \quad \{|x| > 1 \text{ et } |y| > 1\}.$$

• Dans le carré $\{|x|<1 \text{ et } |y|<1\}$ l'équation s'écrit :

$$\frac{dy}{\sqrt{1-y^2}} = \frac{dx}{\sqrt{1-x^2}},$$

d'où $\operatorname{Arc\,sin} y = \operatorname{Arc\,sin} x + \lambda$, $\lambda \in \mathbb{R}$. Comme Arcsin est une bijection de $]-1,1[$ sur $\left]-\frac{\pi}{2},\frac{\pi}{2}\right[$, on a nécessairement $\lambda \in]-\pi,\pi[$. On doit avoir de plus

$$\operatorname{Arc\,sin} x \in \left]-\frac{\pi}{2},\frac{\pi}{2}\right[\cap \left]-\frac{\pi}{2}-\lambda,\frac{\pi}{2}-\lambda\right[= \begin{cases} \left]-\frac{\pi}{2},\frac{\pi}{2}-\lambda\right[& \text{si} \quad \lambda \geq 0, \\ \left]-\frac{\pi}{2}-\lambda,\frac{\pi}{2}\right[& \text{si} \quad \lambda \leq 0. \end{cases}$$

De même $\operatorname{Arc\,sin} y$ est dans $\left]-\frac{\pi}{2}+\lambda,\frac{\pi}{2}\right[$ si $\lambda \geq 0$, et dans $\left]-\frac{\pi}{2},\frac{\pi}{2}+\lambda\right[$ si $\lambda \leq 0$. Les courbes intégrales admettent pour équation

$$y = \sin(\operatorname{Arc\,sin} x + \lambda) = x\,\cos\,\lambda + \sqrt{1-x^2}\,\sin\,\lambda$$

avec

$$\begin{aligned} x &\in]-1,\cos\,\lambda[, \quad y \in]-\cos\,\lambda,1[\quad \text{si} \quad \lambda \geq 0, \\ x &\in]-\cos\,\lambda,1[, \quad y \in]-1,\cos\,\lambda[\quad \text{si} \quad \lambda \leq 0. \end{aligned}$$

L'équation ci-dessus implique $(y - x\,\cos\,\lambda)^2 + x^2\,\sin^2\lambda = \sin^2\lambda$, donc les courbes intégrales sont des arcs d'ellipse.

• L'ouvert $\{|x|>1 \text{ et } |y|>1\}$ est formé de 4 composantes connexes. Plaçons-nous par exemple dans $\{x>1 \text{ et } y>1\}$. On a

$$(E) \Leftrightarrow \frac{dy}{\sqrt{y^2-1}} = \frac{dx}{\sqrt{x^2-1}},$$

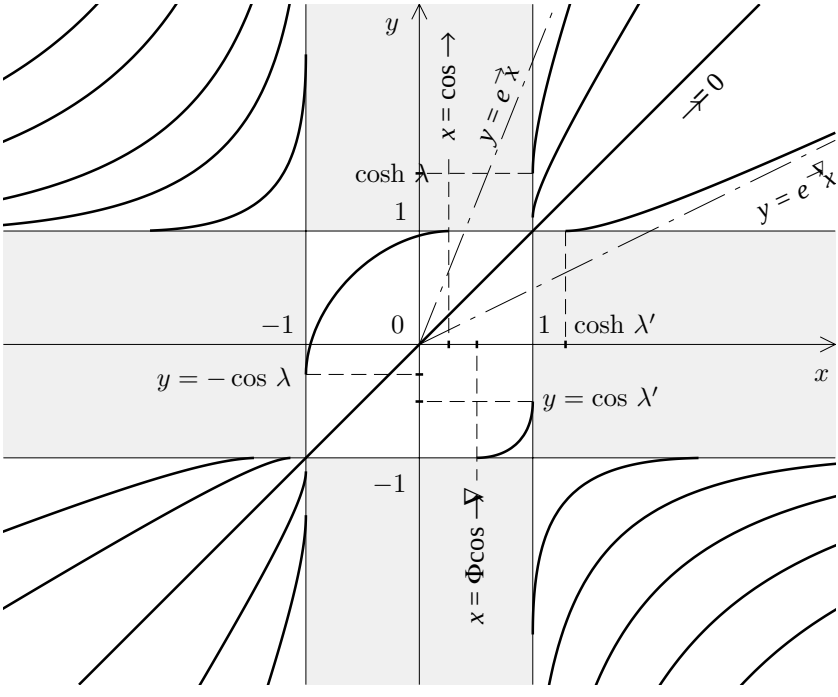
d'où $\operatorname{Arg\,cosh} y = \operatorname{Arg\,cosh} x + \lambda$, $\lambda \in \mathbb{R}$. Arg cosh est une bijection de $]1,+\infty[$ sur $]0,+\infty[$; en raisonnant comme ci-dessus, on obtient

$$y = x\,\cosh\,\lambda + \sqrt{x^2-1}\,\sinh\,\lambda$$

avec

$$\begin{aligned} x &\in]1,+\infty[, \quad y \in]\cosh\,\lambda,+\infty[\quad \text{si} \quad \lambda \geq 0, \\ x &\in]\cosh\,\lambda,+\infty[, \quad y \in]1,+\infty[\quad \text{si} \quad \lambda \leq 0, \end{aligned}$$

par suite $(y - x\,\cosh\,\lambda)^2 - x^2\,\sinh^2\lambda + \sinh^2\lambda = 0$, ce qui est l'équation d'une conique. Comme $\sqrt{x^2-1} = |x|\sqrt{1-\frac{1}{x^2}} = |x| - \frac{1}{2|x|} + O\left(\frac{1}{x^3}\right)$, on voit que la conique admet des asymptotes $y = (\cosh\,\lambda \pm \sinh\,\lambda)x = e^{\pm\lambda}x$ (pour la branche $x>1$ qui nous intéresse, c'est $y = e^{\lambda}x$). On a donc affaire à des arcs d'hyperbole.



On a figuré ici $\lambda > 0$, $\lambda' < 0$.

1.3. Cas où l'on connaît une « intégrale première »

Supposons qu'on cherche à résoudre une équation

(E) $$y' = f(x, y)$$

ou un système différentiel

(S) $$\begin{cases} \dfrac{dx}{dt} = a(x, y) \\ \dfrac{dy}{dt} = b(x, y) \end{cases}$$

dans un ouvert $U \subset \mathbb{R}^2$. Dans les deux cas on a une écriture sous forme différentielle :

$$\text{(E)} \Leftrightarrow f(x, y)dx - dy = 0,$$

$$\text{(S)} \Rightarrow b(x, y)dx - a(x, y)dy = 0.$$

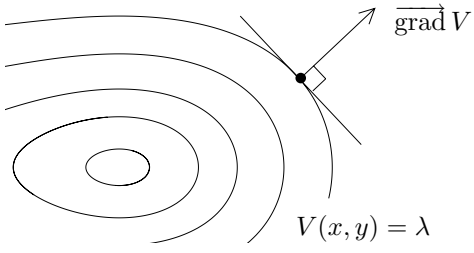
Définition – *On dit qu'une fonction $V : U \to \mathbb{R}$ de classe C^1 est une intégrale première si (E) (respectivement (S)) implique*

$$dV = V'_x(x, y)dx + V'_y(x, y)dy = 0.$$

Dans ce cas, les courbes intégrales $y = \varphi(x)$ vérifient

$$V'_x(x,\varphi(x)) + V'_y(x,\varphi(x))\varphi'(x) = \frac{d}{dx}\,[V(x,\varphi(x)] = 0.$$

Les courbes intégrales sont donc contenues dans les lignes de niveau $V(x,y) = \lambda$, où $\lambda \in \mathbb{R}$ est une constante.



En tout point où $\overrightarrow{\text{grad}}\, V \neq \vec{0}$, la ligne de niveau correspondante possède une tangente perpendiculaire à $\overrightarrow{\text{grad}}\, V$. Le champ des tangentes est dirigé par le vecteur

$$\vec{k}\begin{pmatrix} 1 \\ f(x,y)\end{pmatrix}, \quad \text{resp.} \quad \vec{k}\begin{pmatrix} a(x,y) \\ b(x,y)\end{pmatrix} \quad \text{dans le cas de (E) (resp. (S)).}$$

La condition d'orthogonalité $\overrightarrow{\text{grad}}\, V \perp \vec{k}$ équivaut à la proportionnalité de l'équation $V'_x dx + V'_y dy = 0$ à l'équation différentielle (E) (ou (S)). On peut donc énoncer :

Propriété caractéristique – *V est une intégrale première si et seulement si $\overrightarrow{\text{grad}}\, V$ est orthogonal au champ des tangentes de l'équation différentielle considérée.*

Exemple – Soit $y' = \frac{y}{x+y^2}$ sur $U = \{x + y^2 \neq 0\}$. L'équation se récrit

$$\text{(E)} \qquad y dx - (x+y^2)dy = 0.$$

Cette différentielle n'est pas une différentielle exacte $dV = Pdx + Qdy$ (on devrait avoir $\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$, ce qui n'est pas le cas). On observe néanmoins que

$$d\Big(\frac{x}{y}\Big) = \frac{ydx - xdy}{y^2}.$$

Multiplions alors l'équation (E) par $\frac{1}{y^2}$, en se plaçant dans l'ouvert $y \neq 0$:

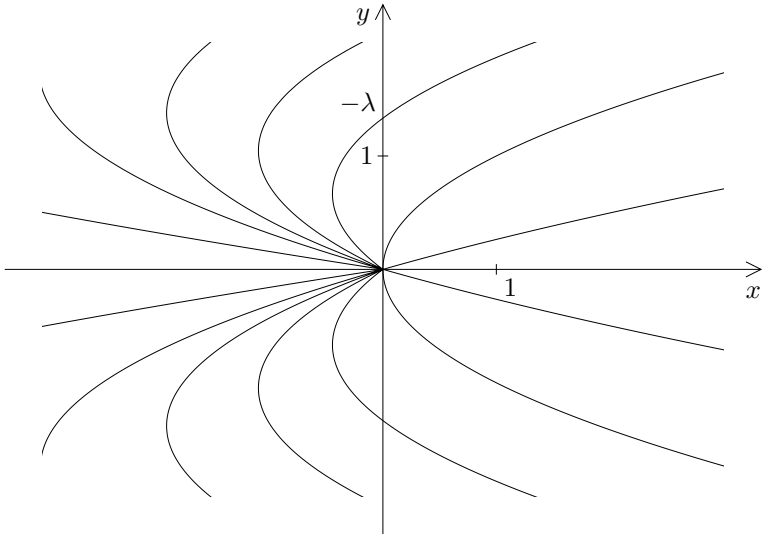
$$\text{(E)} \Leftrightarrow \frac{ydx - xdy}{y^2} - dy = 0 \Leftrightarrow d\Big(\frac{x}{y} - y\Big) = 0.$$

Les courbes intégrales y sont donc données par

$$\frac{x}{y} - y = \lambda \quad \Leftrightarrow \quad x = y^2 + \lambda y.$$

Ce sont des arcs de la parabole d'axe $y = -\frac{\lambda}{2}$ et de sommet $\begin{pmatrix} -\lambda^2/4 \\ -\lambda/2 \end{pmatrix}$, délimités par les points tels que $x + y^2 = 2y^2 + \lambda y = 0$, c'est-à-dire $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ et le sommet, qui doivent être exclus. Par ailleurs, $y = 0$ est une solution singulière, fournissant deux solutions maximales pour $x \in]-\infty, 0[$ et $x \in]0, +\infty[$ respectivement.

Remarque – On dit que $\frac{1}{y^2}$ est un « facteur intégrant » de la forme différentielle $ydx - (x + y^2)dy = 0$.



1.4. Équations linéaires du premier ordre

Ce sont les équations de la forme

$$y' = a(x)y + b(x) \tag{E}$$

où $a, b : I \to \mathbb{R}$ (ou $\mathbb{C}$) sont des fonctions continues.

Supposons qu'on connaisse une solution particulière $y_{(1)}$ de l'équation (E). Alors on obtient par soustraction $y' - y'_{(1)} = a(x)(y - y_{(1)})$, c'est-à-dire que $z = y - y_{(1)}$ vérifie l'équation linéaire « sans second membre »

$$z' = a(x)z. \tag{E_0}$$

Inversement, si z est solution de (E_0), alors $y = y_{(1)} + z$ est solution de (E).

Théorème 1 – *La solution générale de* (E) *s'écrit*

$$y = y_{(1)} + z$$

où $y_{(1)}$ *est une solution particulière de* (E) *et où* z *est la solution générale de* (E_0).

a) *Solutions de* (E_0)

Comme $f(x,z) = a(x)z$ est continue et de dérivée partielle $\frac{\partial f}{\partial z}(x,z) = a(x)$ continue, on sait que le problème de Cauchy admet une solution unique en tout point $(x_0, z_0) \in I \times \mathbb{R}$. Or $z(x) \equiv 0$ est clairement solution de (E_0). D'après l'unicité, aucune autre solution ne peut s'annuler en un quelconque point $x_0 \in I$. Si $z \neq 0$, on peut donc écrire

$$\begin{aligned} \frac{z'}{z} &= a(x), \\ \ln|z| &= A(x) + C, \quad C \in \mathbb{R}, \end{aligned}$$

où A est une primitive de a sur I. On en déduit

$$\begin{aligned} |z(x)| &= e^C e^{A(x)}, \\ z(x) &= \varepsilon(x) e^C e^{A(x)} \quad \text{avec} \quad \varepsilon(x) = \pm 1. \end{aligned}$$

Comme z est continue et ne s'annule pas, le signe de z ne change pas, d'où

$$z(x) = \lambda e^{A(x)}$$

avec $\lambda = \pm e^C$. Inversement, toute fonction

$$z(x) = \lambda e^{A(x)}, \quad \lambda \in \mathbb{R}$$

et visiblement solution de (E_0). On peut donc énoncer :

Théorème 2 – *Les solutions maximales de* (E_0) *:* $z' = a(x)z$ *forment un espace vectoriel de dimension 1, ayant pour base* $x \mapsto e^{A(x)}$*.*

b) *Recherche d'une solution particulière* $y_{(1)}$ *de* (E)*.*

Si aucune solution évidente n'apparaît, on peut utiliser la méthode dite de *variation des constantes*, c'est-à-dire que l'on cherche $y_{(1)}$ sous la forme

$$y_{(1)}(x) = \lambda(x) e^{A(x)},$$

où λ est différentiable. Il vient

$$\begin{aligned} y'_{(1)}(x) &= \lambda(x)a(x)e^{A(x)} + \lambda'(x)e^{A(x)} \\ &= a(x)y_{(1)}(x) + \lambda'(x)e^{A(x)}. \end{aligned}$$

$y_{(1)}$ est donc solution de (E) si on prend

$$\begin{aligned} &\lambda'(x)e^{A(x)} = b(x), \\ &\lambda'(x) = b(x)e^{-A(x)}, \\ &\lambda(x) = \int_{x_0}^{x} b(t)e^{-A(t)}, \quad x_0 \in I. \end{aligned}$$

On obtient ainsi la solution particulière

$$y_{(1)}(x) = e^{A(x)} \int_{x_0}^{x} b(t) e^{-A(t)} dt$$

telle que $y_{(1)}(x_0) = 0$. La solution générale est donnée d'après le théorème 1 par

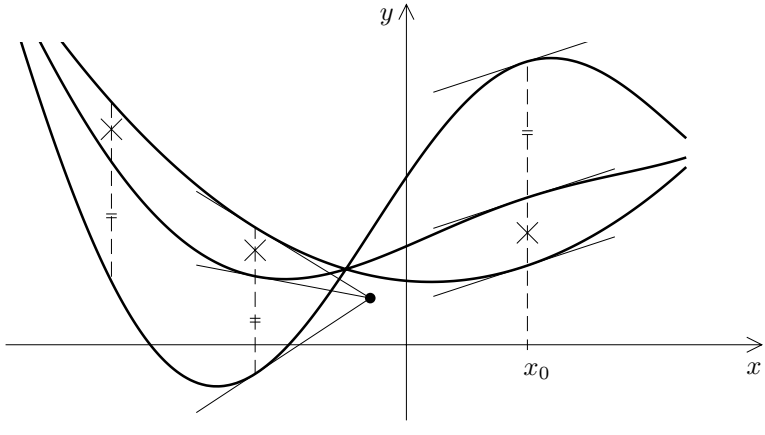
$$y(x) = e^{A(x)} \Big(\lambda + \int_{x_0}^{x} b(t) e^{-A(t)} dt \Big).$$

La solution du problème de Cauchy $y(x_0) = y_0$ est obtenue pour $\lambda = e^{-A(x_0)} y_0$.

Exercice – *Propriétés géométriques liées aux équations linéaires* (*cf. schéma*).

(a) *Si $y_{(1)}$, $y_{(2)}$, $y_{(3)}$ sont trois solutions d'une équation linéaire, montrer que la fonction $y_{(3)} - y_{(2)}$ est proportionnelle à $y_{(2)} - y_{(1)}$.*

(b) *Montrer qu'une équation $y' = f(x, y)$ est linéaire si et seulement si le champ des tangentes a la propriété suivante : pour tout x_0 fixé, les tangentes aux différents points (x_0, y) sont concourantes ou toutes parallèles.*



1.5. Équations se ramenant à des équations linéaires

a) Équations de Bernoulli

Ce sont les équations de la forme

$$\frac{dy}{dx} = p(x)y + q(x)y^{\alpha}, \quad \alpha \in \mathbb{R} \setminus \{1\}, \tag{E}$$

avec $p, q : I \to \mathbb{R}$ continues (pour $\alpha = 1$, (E) est linéaire).
On se place dans le demi-plan supérieur $U = \mathbb{R} \times]0, +\infty[= \{(x, y)\,;\ y > 0\}$. En multipliant par $y^{-\alpha}$, on obtient

$$(\mathrm{E}) \quad \Leftrightarrow \quad y^{-\alpha}\frac{dy}{dx} = p(x)y^{1-\alpha} + q(x)$$

Posons $z = y^{1-\alpha}$; alors $\frac{dz}{dx} = (1-\alpha)y^{-\alpha}\frac{dy}{dx}$, d'où

$$(\mathrm{E}) \quad \Leftrightarrow \quad \frac{1}{1-\alpha}\,\frac{dz}{dx} = p(x)z + q(x)$$

On est donc ramené à une équation linéaire en z.

b) Équations de Riccati

Ce sont les équations de la forme

$$(\mathrm{E}) \qquad y' = a(x)y^2 + b(x)y + c(x)$$

avec $a, b, c : I \to \mathbb{R}$ continues, c'est-à-dire que $f(x, y)$ est un polynôme de degré ≤ 2 en y. Montrons que l'on sait résoudre (E) dès que l'on connaît une solution particulière $y_{(1)}$. Posons $y = y_{(1)} + z$. Il vient

$$\begin{aligned} y'_{(1)} + z' &= a(x)(y_{(1)}^2 + 2y_{(1)}z + z^2) + b(x)(y_{(1)} + z) + c(x) \\ &= a(x)y_{(1)}^2 + b(x)y_{(1)} + c(x) + (2a(x)y_{(1)} + b(x))z + a(x)z^2. \end{aligned}$$

Comme $y'_{(1)}$ se simplifie, on en déduit

$$z' = (2a(x)y_{(1)}(x) + b(x)) + a(x)z^2.$$

C'est une équation linéaire de Bernoulli avec $\alpha = 2$. On la ramène à une équation linéaire en posant $w = z^{1-\alpha} = \frac{1}{z}$.

Exemple – Soit l'équation $(1 - x^3)y' + x^2y + y^2 - 2x = 0$.

On remarque que $y_{(1)}(x) = x^2$ est solution particulière. En posant $y = x^2 + z$ on se ramène à

$$(1 - x^3)z' + 3x^2z + z^2 = 0$$

puis, après division par z^2, à

$$-(1 - x^3)w' + 3x^2w + 1 = 0 \quad \text{avec} \quad w = \frac{1}{z},$$

soit

$$w' = \frac{3x^2}{1 - x^3}\,w + \frac{1}{1 - x^3}, \quad \text{si} \quad x \neq 1.$$

L'équation linéaire sans second membre $\frac{w'}{w} = \frac{3x^2}{1-x^3}$ donne

$$\ln|w| = -\ln|1-x^3| + C, \quad \text{d'où} \quad w = \frac{\lambda}{1-x^3}.$$

La méthode de variation des constantes conduit à

$$\frac{\lambda'}{1-x^3} = \frac{1}{1-x^3} \quad \text{soit} \quad \lambda' = 1, \quad \lambda(x) = x.$$

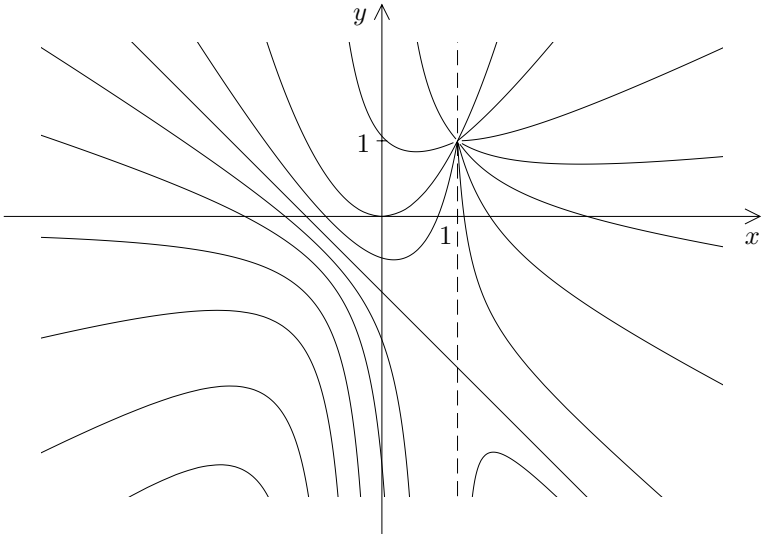
La solution générale de l'équation linéaire complète est donc

$$w(x) = \frac{x+\lambda}{1-x^3},$$

d'où $y = x^2 + z = x^2 + \frac{1}{w} = x^2 + \frac{1-x^3}{x+\lambda}$, soit encore

$$y(x) = \frac{\lambda x^2 + 1}{x+\lambda} = \lambda x - \lambda^2 + \frac{1+\lambda^3}{x+\lambda}.$$

Pour $\lambda = -1$, on obtient la droite $y = -x - 1$. Pour $\lambda \neq -1$, il s'agit d'une hyperbole $(y - \lambda x + \lambda^2)(x+\lambda) = 1 + \lambda^3$, admettant pour asymptotes les droites $x = -\lambda$ et $y = \lambda x - \lambda^2$. La solution singulière $y_{(1)}(x) = x^2$ est la solution limite obtenue quand $|\lambda|$ tend vers $+\infty$.



1.6. Équations homogènes

Une équation homogène est une équation qui peut se mettre sous la forme

(E) $$y' = f\left(\frac{y}{x}\right) \quad \text{où} \quad f : I \to \mathbb{R} \quad \text{est continue.}$$

C'est le cas par exemple des équations $y' = \frac{P(x,y)}{Q(x,y)}$ où P, Q sont des polynômes homogènes de même degré d : une division par x^d au numérateur et au dénominateur nous ramène à $y' = \frac{P(1,y/x)}{Q(1,y/x)}$.

Méthode – On pose $z = \frac{y}{x}$, c'est-à-dire $y = xz$. Il vient

$$y' = z + xz' = f(z),$$

donc z satisfait l'équation à variables séparées

$$z' = \frac{f(z) - z}{x}.$$

- On a d'une part les solutions singulières

$$z(x) = z_j, \qquad y(x) = z_j x \qquad \text{(droites passant par 0)},$$

où $\{z_j\}$ est l'ensemble des racines de $f(z) = z$.

- Pour $f(z) \neq z$ on peut écrire

$$\frac{dz}{f(z) - z} = \frac{dx}{x},$$
$$F(z) = \ln|x| + C = \ln(\lambda x), \qquad \lambda \in \mathbb{R}^*,$$

où F est une primitive de $z \mapsto 1/(f(z) - z)$ sur $]z_j, z_{j+1}[$. On en déduit que $z = F^{-1}(\ln(\lambda x))$, d'où la famille de courbes intégrales

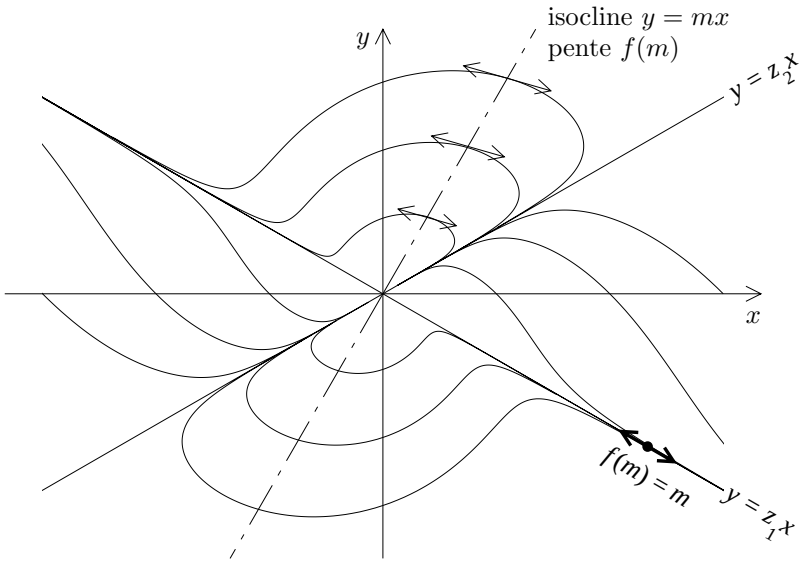
$$C_\lambda : y = xF^{-1}(\ln(\lambda x)),$$

définies dans le secteur angulaire $z_j < \frac{y}{x} < z_{j+1}$, $\lambda x > 0$.

En cas de divergence de F aux points z_j, z_{j+1}, on a $F^{-1} :]-\infty, +\infty[\to]z_j, z_{j+1}[$ monotone bijective et $\frac{y}{x} \to z_j$ ou z_{j+1} quand $x \to 0$ ou ∞. On a donc d'une part une branche infinie de direction asymptotique $y = z_{j+1}x$ (resp. $y = z_j x$) et une tangente $y = z_j x$ (resp. $y = z_{j+1}x$) au point 0 si F est croissante (resp. décroissante). Noter que la droite $y = z_j x$ n'est pas nécessairement asymptote : voir l'exemple ci-dessous.

Observons enfin que les lignes isoclines sont les droites $y = mx$, la pente correspondante étant $f(m)$. Le champ des tangentes est donc invariant par les homothéties de centre O. Ceci permet de voir que *l'homothétique d'une courbe intégrale est encore une courbe intégrale.*

Exercice – *Vérifier que* $C_\lambda = h_{1/\lambda}(C_1)$ *où* $h_\lambda(x, y) = (\lambda x, \lambda y)$.



Exemple – L'équation $xy'(2y - x) = y^2$ peut se récrire

$$y' = \frac{y^2}{x(2y - x)} \quad \text{si} \quad x \neq 0, \quad y \neq \frac{x}{2}.$$

y' est donc une fonction rationnelle en x, y dont le numérateur et le dénominateur sont des polynômes homogènes de degré 2. En divisant le numérateur et dénominateur par x^2 on obtient

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

Posons $z = \frac{y}{x}$, soit $y = xz$. Il vient

$$y' = xz' + z = \frac{z^2}{2z - 1},$$
$$xz' = \frac{z^2}{2z - 1} - z = \frac{z - z^2}{2z - 1} = \frac{z(1 - z)}{2z - 1}.$$

• Solutions singulières :

$$z = 0, \qquad z = 1,$$
$$y = 0, \qquad y = x.$$

• Pour $z \neq 0$, $z \neq 1$ l'équation se récrit

$$\frac{2z - 1}{z(1 - z)}\, dz = \frac{dx}{x}.$$

La fonction

$$\frac{2z-1}{z(1-z)} = \frac{z-(1-z)}{z(1-z)} = \frac{1}{1-z} - \frac{1}{z}$$

admet pour primitive

$$-\ln|1-z| - \ln|z| = -\ln|z(1-z)|,$$

d'où le calcul des courbes intégrales :

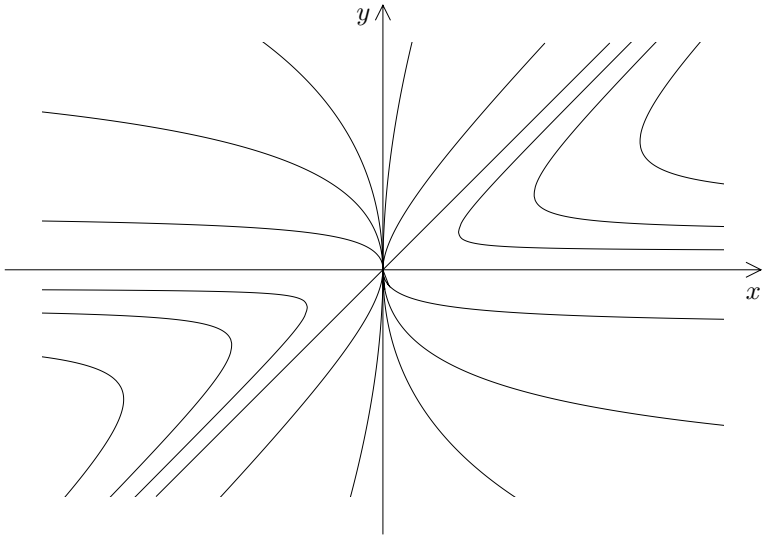
$$\begin{aligned} &\ln|z(1-z)| = -\ln|x| + C, \\ &z(1-z) = \frac{\lambda}{x}, \quad \frac{y}{x}\left(1-\frac{y}{x}\right) = \frac{\lambda}{x}, \\ &y(x-y) = \lambda x. \end{aligned}$$

Les courbes intégrales sont donc des coniques. On peut mettre l'équation sous la forme

$$(y-\lambda)(x-y-\lambda) = \lambda^2$$

c'est-à-dire $XY = \lambda$ avec $X = x-y-\lambda$ et $Y = y-\lambda$. Il s'agit d'une hyperbole d'asymptotes $y = \lambda$, $y = x-\lambda$ (parallèles aux directions asymptotiques $y = 0$, $y = x$ données par les droites intégrales singulières).

Exercice – *Montrer que chaque hyperbole passe par* $(0,0)$ *avec tangente* $x = 0$.



Autre Méthode de résolution – Utilisation des coordonnées polaires.

Pour $r > 0$ et $\theta \in \mathbb{R}$ on pose

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta \end{cases}.$$

Il vient

$$\frac{dy}{dx} = \frac{dr\,\sin\theta + r\,\cos\theta\,d\theta}{dr\,\cos\theta - r\,\sin\theta\,d\theta} = \frac{dr\,\tan\theta + r\,d\theta}{dr - r\,\tan\theta\,d\theta}.$$

L'équation (E) $y' = f\left(\frac{y}{x}\right)$ se transforme alors en

$$\begin{aligned} &dr\,\tan\theta + r\,d\theta = (dr - r\,\tan\theta\,d\theta) f(\tan\theta),\\ &dr(f(\tan\theta) - \tan\theta) = r d\theta(1 + \tan\theta\, f(\tan\theta)),\\ &\frac{dr}{r} = \frac{1 + \tan\theta\, f(\tan\theta)}{f(\tan\theta) - \tan\theta}\, d\theta. \end{aligned}$$

On aboutit donc à une équation à variables séparées r, θ. Les intégrales singulières correspondent aux droites $\theta = \theta_j$ telles que $f(\tan\theta_j) = \tan\theta_j$.

Exercice – *Résoudre $y' = \frac{x+y}{x-y}$ à l'aide des deux méthodes proposées. Quelle est la nature des courbes intégrales ?*

2. ÉQUATIONS DU PREMIER ORDRE NON RÉSOLUES EN y'

2.1. DÉFINITIONS ET PREMIÈRES PROPRIÉTÉS

On appelle équation du premier ordre non résolue en y' une équation de la forme

$$f(x, y, y') = 0 \qquad \text{(E)}$$

où $(x, y, p) \mapsto f(x, y, p)$ est une fonction de classe C^1 dans un ouvert $U \subset \mathbb{R}^3$. Plaçons-nous au voisinage d'un point $(x_0, y_0) \in \mathbb{R}^2$. On suppose que l'équation $f(x_0, y_0, p) = 0$ admet des racines $p_1, p_2, \ldots, p_N$ et que ces racines sont simples, c'est-à-dire

$$\frac{\partial f}{\partial p}\,(x_0, y_0, p_j) \neq 0.$$

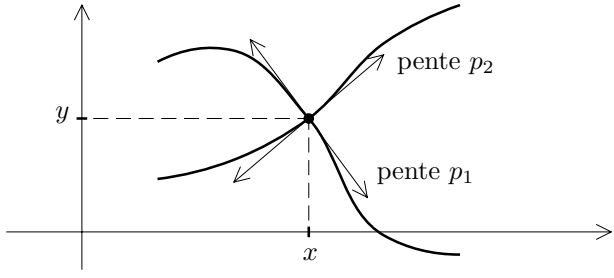
D'après le théorème des fonctions implicites, on sait alors qu'il existe un voisinage V de (x_0, y_0), un réel $h > 0$ et une fonction $g_j : V \to]p_j - h, p_j + h[$ de classe C^1, $1 \leq j \leq N$, tels que pour tout $(x, y, p) \in V \times]p_j - h, p_j + h[$ on ait

$$f(x, y, p) = 0 \quad \Leftrightarrow \quad p = g_j(x, y).$$

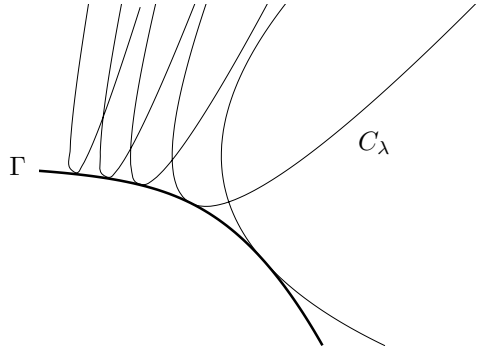
L'équation différentielle $f(x, y, y') = 0$ nous amène alors à résoudre dans V les N équations différentielles

$$y' = g_j(x, y). \qquad (\mathrm{E}_j)$$

Comme g_j est de classe C^1, on voit que par tout point $(x, y) \in V$ il passe exactement N courbes intégrales dont les pentes sont les racines p de $f(x, y, p) = 0$.



Remarque – Dans cette situation, il arrive fréquemment qu'on ait une famille de courbes intégrales C_λ admettant une *enveloppe* Γ, c'est-à-dire une courbe Γ qui est tangente en chacun de ses points à l'une des courbes C_λ.



La courbe Γ est alors elle-même une courbe intégrale, car en chaque point sa tangente appartient au champ des tangentes de l'équation (E) (elle coïncide avec la tangente de l'une des courbes C_λ). Γ est donc une solution singulière. On notera qu'une telle courbe Γ doit satisfaire simultanément les deux équations $f(x, y, y') = 0$ et $\partial f/\partial p(x, y, y') = 0$: chaque point $(x, y) \in \Gamma$ est en effet limite d'une suite de points en lesquels deux tangentes du champ viennent se confondre, de sorte que $p = y'$ est racine double de $f(x, y, p) = 0$. En particulier les hypothèses faites ci-dessus pour appliquer le théorème des fonctions implicites ne sont pas satisfaites si $(x_0, y_0) \in \Gamma$.

Méthode de Résolution – Pour résoudre les équations différentielles non résolues en y', le principe général est de chercher une paramétrisation de x, y, y' en fonction d'un paramètre t qui sera alors choisi comme nouvelle variable.

2.2. Cas des équations non résolues incomplètes

a) Équations du type $(E) : f(x, y') = 0$

Supposons que l'équation $f(x, p) = 0$ admette une paramétrisation de classe C^1

$$\begin{cases} x = \varphi(t) \\ p = \psi(t). \end{cases}$$

On a alors

$$dx = \varphi'(t)\,dt$$
$$dy = y'\,dx = \psi(t)\,dx = \psi(t)\varphi'(t)\,dt$$

On en déduit

$$y = \int_{t_0}^{t} \psi(u)\varphi'(u)du + \lambda = \rho(t) + \lambda,$$

ce qui donne une paramétrisation des courbes intégrales :

$$\begin{cases} x = \varphi(t) \\ y = \rho(t) + \lambda. \end{cases}$$

b) Équations du type $(E) : f(y, y') = 0$,

connaissant une paramétrisation $\begin{cases} y = \varphi(t) \\ y' = \psi(t). \end{cases}$

On obtient $dy = \varphi'(t)dt = \psi(t)dx$, d'où $dx = \frac{\varphi'(t)}{\psi(t)}\,dt$. Les courbes intégrales sont paramétrées par

$$\begin{cases} x = \rho(t) + \lambda \\ y = \varphi(t) \end{cases}$$

avec $\rho(t) = \displaystyle\int_{t_0}^{t} \frac{\varphi'(u)}{\psi(u)}\,du.$

2.3. Équations homogènes non résolues

Ce sont les équations pouvant être mises sous la forme

$$f\Big(\frac{y}{x}, y'\Big) = 0. \qquad \text{(E)}$$

Supposons qu'on connaisse une paramétrisation

$$\begin{cases} \frac{y}{x} = \varphi(t) \\ y' = \psi(t). \end{cases}$$

On a alors

$$y = x\varphi(t),$$
$$\begin{cases} dy = \varphi(t)dx + x\varphi'(t)dt \\ dy = \psi(t)\,dx, \end{cases}$$

d'où $(\psi(t) - \varphi(t))\,dx = x\varphi'(t)\,dt.$

- On a d'une part des solutions singulières correspondant aux racines t_j de $\psi(t) = \varphi(t)$, donnant des droites

$$y = x\varphi(t_j).$$

- D'autre part, pour $t \neq t_j$ on obtient

$$\frac{dx}{x} = \frac{\varphi'(t)}{\psi(t) - \varphi(t)}\, dt,$$

ce qui donne par intégration de $\varphi'/(\psi - \varphi)$:

$$\begin{aligned} &\ln|x| = \rho(t) + C, \\ &\begin{cases} x = \lambda e^{\rho(t)} \\ y = x\varphi(t) = \lambda\varphi(t)e^{\rho(t)} \end{cases}, \qquad \lambda \in \mathbb{R}. \end{aligned}$$

Il est clair sur ces dernières formules que les courbes intégrales se déduisent les unes des autres par les homothéties de centre O.

Exemple – Soit l'équation $x^2(y + 3xy') = (y + xy')^3$.
En divisant par x^3 on trouve

$$\frac{y}{x} + 3y' = \Big(\frac{y}{x} + y'\Big)^3,$$

c'est donc une équation homogène non résolue en y'. On obtient une paramétrisation en posant

$$\begin{cases} \frac{y}{x} + y' = t \\ \frac{y}{x} + 3y' = t^3, \end{cases}$$

d'où

$$\begin{cases} \frac{y}{x} = \frac{1}{2}\,(3t - t^3) \\ y' = \frac{1}{2}\,(t^3 - t). \end{cases} \qquad (*)$$

En différentiant $y = \frac{1}{2}\,(3t - t^3)x$ on obtient

$$\begin{aligned} dy &= \frac{1}{2}\,(3 - 3t^2)dt \cdot x + \frac{1}{2}\,(3t - t^3)\,dx \\ &= y'\,dx = \frac{1}{2}\,(t^3 - t)\,dx, \end{aligned}$$

d'où l'équation

$$\begin{aligned} &(t^3 - 2t)dx = \frac{1}{2}\,(3 - 3t^2)dt \cdot x, \\ &\frac{dx}{x} = \frac{3(1 - t^2)}{2t(t^2 - 2)}\, dt. \end{aligned}$$

• Solutions singulières : $t(t^2 - 2) = 0 \Leftrightarrow t = 0, \sqrt{2}, -\sqrt{2}$. En remplaçant dans $(*)$ on obtient les droites

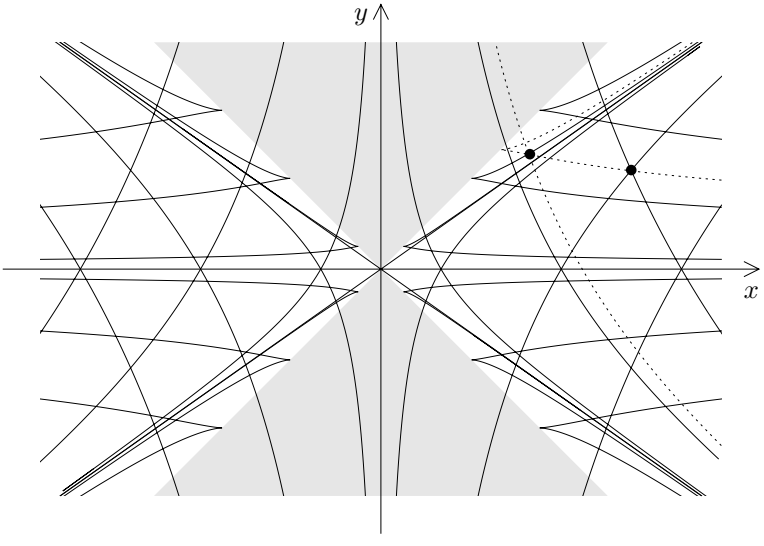
$$y = 0, \qquad y = \frac{\sqrt{2}}{2}\,x, \qquad y = -\frac{\sqrt{2}}{2}\,x.$$

• Solution générale :

$$\frac{1-t^2}{t(t^2-2)} = \frac{1-\frac{t^2}{2}-\frac{t^2}{2}}{t(t^2-2)} = -\frac{1}{2t} - \frac{t}{2(t^2-2)},$$
$$\frac{3(1-t^2)}{2t(t^2-2)}\,dt = -\frac{3}{4}\,\frac{dt}{t} - \frac{3}{4}\,\frac{tdt}{t^2-2}.$$

On en déduit

$$\ln|x| = -\frac{3}{4}\ln|t| - \frac{3}{8}\ln|t^2-2| + C,$$
$$\begin{cases} x = \lambda|t|^{-3/4}|t^2-2|^{-3/8} \\ y = \frac{y}{x}\cdot x = \frac{\lambda}{2}\,(3t-t^3)|t|^{-3/4}|t^2-2|^{-3/8}. \end{cases}$$



Exercice – *Montrer que par tout point (x,y) tel que $|y|<|x|$ il passe exactement trois courbes intégrales, alors qu'il n'en passe qu'une si $|y|>|x|$. Combien en passe-t-il si $|y|=|x|$?* [Indication : *étudier le nombre de valeurs de t et y' associées à une valeur donnée de y/x*].

2.4. Équations de Lagrange (ou équations à isoclines rectilignes)

Cherchons à déterminer les équations différentielles dont les courbes isoclines sont des droites. La courbe isocline $y' = p$ sera une droite $y = a(p)x + b(p)$ (pour simplifier, on écarte le cas des droites parallèles à $y'Oy$). L'équation différentielle correspondante est donc

$$y = a(y')x + b(y'). \tag{E}$$

On supposera que a, b sont au moins de classe C^1.

Méthode de Résolution – On choisit $p = y'$ comme nouvelle variable paramétrant chaque courbe intégrale ; ceci est légitime à condition que y' ne soit pas une constante sur un morceau de la courbe intégrale considérée. Dans le cas contraire, si $y' = p_0 =$ constante, la courbe intégrale est contenue dans la droite $y = a(p_0)x + b(p_0)$, ce qui n'est compatible avec la condition $y' = p_0$ que si $a(p_0) = p_0$.

• On a donc des solutions singulières $y = p_j x + b(p_j)$ où les p_j sont les racines de $a(p) = p$.

• Solution générale :

$$\begin{cases} y = a(p)x + b(p), \\ dy = a(p)dx + (a'(p)x + b'(p))dp \\ dy = y'dx = pdx. \end{cases}$$

Il vient

$$(p - a(p))dx = (a'(p)x + b'(p))dp,$$

et pour $p \neq a(p)$ on aboutit à

$$\frac{dx}{dp} = \frac{1}{p - a(p)} \left(a'(p)x + b'(p)\right);$$

c'est une équation linéaire en la fonction $x(p)$. La solution générale sera de la forme

$$\begin{cases} x(p) = x_{(1)}(p) + \lambda z(p), \quad \lambda \in \mathbb{R}, \\ y(p) = a(p)(x_{(1)}(p) + \lambda z(p)) + b(p). \end{cases}$$

Exercice – *Résoudre l'équation* $2y - x(y' + y'^3) + y'^2 = 0$.

2.5. Équations de Clairaut

C'est le cas particulier des équations de Lagrange dans lequel $a(p) = p$ pour toute valeur de p, soit

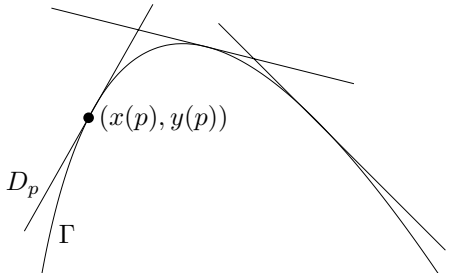
$$y = y'x + b(y'). \tag{E}$$

Les droites

$$D_p : y = px + b(p)$$

qui étaient précédemment des solutions singulières forment maintenant une famille générale de solutions.

Montrons que les droites D_p possèdent toujours une enveloppe Γ. Une telle courbe Γ admet par définition une paramétrisation $(x(p), y(p))$ telle que Γ soit tangente à D_p au point $(x(p), y(p))$.



Le vecteur tangent $(x'(p), y'(p))$ à Γ doit avoir même pente p que D_p, d'où $y'(p) = px'(p)$. Par ailleurs $(x(p), y(p)) \in D_p$, donc

$$y(p) = px(p) + b(p).$$

En différentiant, il vient

$$y'(p) = px'(p) + x(p) + b'(p).$$

Ceci implique $x(p) + b'(p) = 0$, d'où la paramétrisation cherchée de l'enveloppe :

$$\Gamma \begin{cases} x(p) = -b'(p) \\ y(p) = -pb'(p) + b(p). \end{cases}$$

Si b est de classe C^2, on a $y'(p) = -pb''(p) = px'(p)$ de sorte que Γ est bien l'enveloppe des droites D_p. La courbe Γ est une solution singulière de (E).

Exercice – *Résoudre l'équation* $(xy' - y)(1 + y'^2) + 1 = 0$.

3. Problèmes géométriques conduisant à des équations différentielles du premier ordre

3.1. Équation différentielle associée à une famille de courbes

On considère le problème suivant :

Problème – Etant donné une famille de courbes

$$C_\lambda : h(x, y, \lambda) = 0, \quad \lambda \in \mathbb{R},$$

existe-t-il une équation différentielle du premier ordre dont les courbes C_λ soient les courbes intégrales ?

• **Cas particulier.** On suppose que les courbes C_λ sont les lignes de niveau d'une fonction V de classe C^1 :

$$C_\lambda : V(x,y) = \lambda, \qquad \lambda \in \mathbb{R}.$$

Alors les courbes C_λ sont solutions de l'équation différentielle

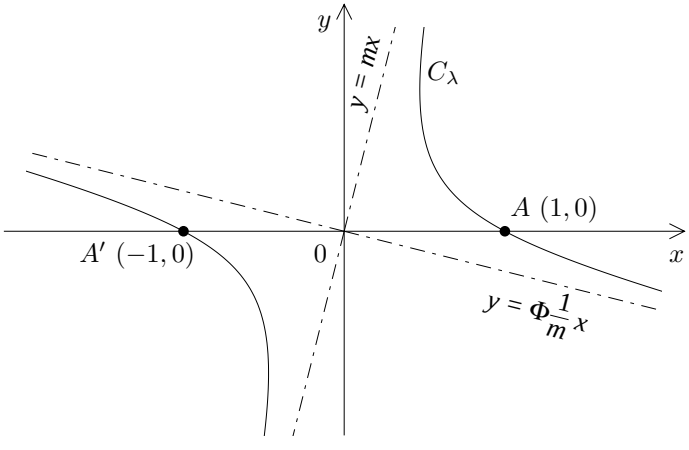
$$V'_x(x,y)dx + V'_y(x,y)dy = 0. \tag{E}$$

• **Cas général.** Si l'équation $h(x,y,\lambda) = 0$ peut se mettre sous la forme $\lambda = V(x,y)$, on est ramené au cas précédent. Sinon on écrit que sur chaque C_λ on a

$$\begin{cases} h(x,y,\lambda) = 0 \\ h'_x(x,y,\lambda)dx + h'_y(x,y,\lambda)dy = 0, \end{cases}$$

et on essaie d'éliminer λ entre les 2 équations pour obtenir une équation ne faisant plus intervenir que x, y, dx, dy.

Exemple – Soit C_λ la famille des hyperboles équilatères de centre O passant par le point $A(1,0)$.



Les asymptotes de C_λ sont alors des droites orthogonales passant par O, soit

$$y = mx, \qquad y = -\frac{1}{m}\,x, \qquad m \in \mathbb{R}^*.$$

Posons $X = y - mx$, $Y = y + \frac{1}{m}\,x$. L'équation de l'hyperbole cherchée s'écrit

$$\begin{aligned} &XY = C \quad \text{(constante)}, \\ &(y - mx)\Big(y + \frac{1}{m}\,x\Big) = C, \\ &y^2 - x^2 + \Big(\frac{1}{m} - m\Big)xy = C. \end{aligned}$$

En faisant $x = 1$, $y = 0$ on trouve $C = -1$, d'où l'équation

$$C_\lambda : y^2 - x^2 + \lambda xy + 1 = 0, \qquad \lambda \in \mathbb{R},$$

avec $\lambda = \frac{1}{m} - m$ (noter que $m \mapsto \frac{1}{m} - m$ est surjective de $\mathbb{R}^*$ sur $\mathbb{R}$). Sur C_λ on a :

$$\lambda = \frac{x^2 - y^2 - 1}{xy},$$
$$d\lambda = 0 = \frac{(2xdx - 2ydy)xy - (x^2 - y^2 - 1)(xdy + ydx)}{x^2y^2}.$$

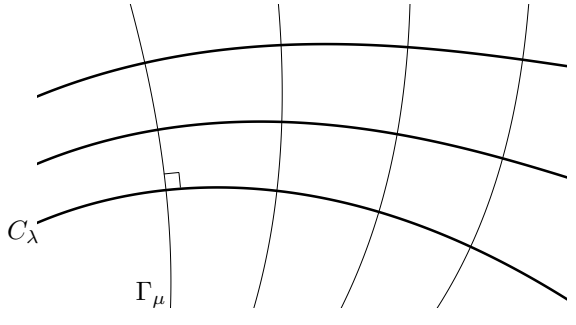
L'équation différentielle des courbes C_λ est donc

$$\text{(E)} \ : \quad (2x^2y - x^2y + y^3 + y)dx + (-2xy^2 - x^3 + xy^2 + x)dy = 0,$$
$$\text{(E)} \ : \quad (x^2 + y^2 + 1)ydx - (x^2 + y^2 - 1)xdy = 0.$$

3.2. Recherche des trajectoires orthogonales à une famille de courbes

Soient (C_λ), (Γ_μ) deux familles de courbes.

Définition – *On dit que C_λ et Γ_λ sont orthogonales si les tangentes à C_λ et Γ_μ sont orthogonales en tout point de $C_\lambda \cap \Gamma_\mu$, quels que soient λ et μ.*



Problème – *Etant donné une famille de courbes C_λ, trouver la famille (Γ_μ) des courbes qui sont orthogonales aux C_λ.*

Pour cela, on suppose que l'on connaît une équation différentielle (E) satisfaite par les courbes C_λ, et on cherche l'équation différentielle ($\text{E}^\perp$) des courbes orthogonales Γ_μ. Distinguons quelques cas.

• (C_λ) satisfait (E) : $y' = f(x,y)$.
En un point (x,y) donné, la pente de la tangente à C_λ est $y' = f(x,y)$. La pente de la tangente à Γ_μ est donc $-1/f(x,y)$. Les courbes (Γ_μ) sont donc solutions de

$$(\mathrm{E}^\perp) \ : \qquad y' = -\frac{1}{f(x,y)}.$$

• (C_λ) satisfait (E) : $\dfrac{d\overrightarrow{M}}{dt} = \overrightarrow{V}(M) \Leftrightarrow \begin{cases} \dfrac{dx}{dt} = a(x,y) \\ \dfrac{dy}{dt} = b(x,y) \end{cases}$.

La tangente à C_λ est portée par $\overrightarrow{V}(M)$, celle de (Γ_μ) est donc portée par le vecteur orthogonal $\overrightarrow{V}(M)^\perp \begin{pmatrix} -b(x,y) \\ a(x,y) \end{pmatrix}$. Par suite (Γ_μ) est solution de

$$(\mathrm{E}^\perp) \quad \begin{cases} \dfrac{dx}{dt} = -b(x,y) \\ \dfrac{dy}{dt} = a(x,y) \end{cases}$$

• (C_λ) satisfait (E) : $\alpha(x,y)dx + \beta(x,y)dy = 0$.
Alors (Γ_μ) vérifie $(\mathrm{E}^\perp)$: $-\beta(x,y)dx + \alpha(x,y)dy = 0$.

Cas particulier. Supposons que les courbes C_λ sont les lignes de niveau $V(x,y) = \lambda$ de la fonction V. Elles vérifient alors

$$(\mathrm{E}) \qquad V'_x(x,y)dx + V'_y(x,y)dy = 0.$$

Leurs trajectoires orthogonales (Γ_μ) sont les lignes de champ du gradient $\overrightarrow{\mathrm{grad}}\, V$:

$$(\mathrm{E}^\perp) \quad \begin{cases} \dfrac{dx}{dt} = V'_x(x,y) \\ \dfrac{dy}{dt} = V'_y(x,y) \end{cases}$$

Exemple – Soit $C_\lambda : y^2 - x^2 + \lambda xy + 1 = 0$ (cf. § 3.1).
Nous avons vu que C_λ vérifie

$$(\mathrm{E}) \ : \qquad (x^2+y^2+1)ydx - (x^2+y^2-1)xdy = 0.$$

Donc Γ_μ vérifie

$$\begin{aligned} (\mathrm{E}^\perp) \ : \quad & (x^2+y^2-1)xdx + (x^2+y^2+1)ydy = 0 \\ \Leftrightarrow\ & (x^2+y^2)(xdx+ydy) - xdx + ydy = 0. \end{aligned}$$

Une intégrale première apparaît immédiatement :

$$d\Big[\frac{1}{4}\,(x^2+y^2)^2 - \frac{x^2}{2} + \frac{y^2}{2}\Big] = 0$$

Les courbes Γ_μ sont donc les lignes de niveau

$$(x^2+y^2)^2 - 2x^2 + 2y^2 = \mu, \qquad \mu \in \mathbb{R},$$

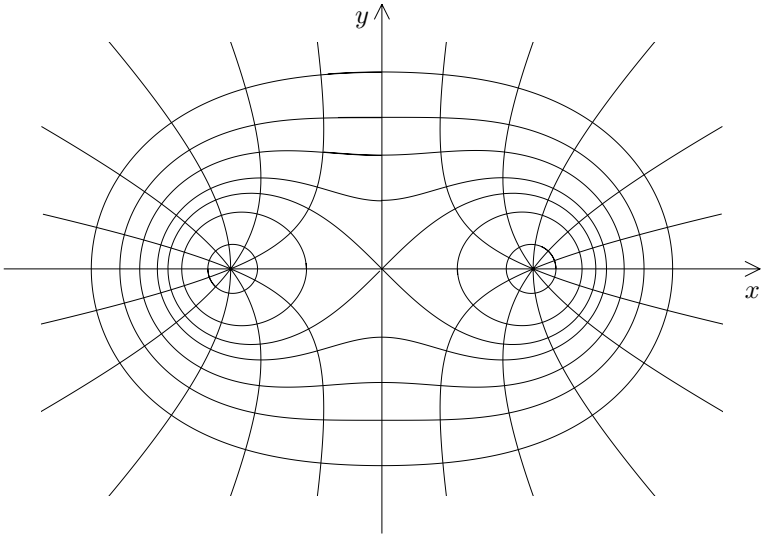
ce qui peut encore s'écrire

$$\begin{aligned}
&(x^2+y^2+1)^2 - 4x^2 = \mu + 1,\\
&(x^2-2x+1+y^2)(x^2+2x+1+y^2) = \mu+1,\\
&((x-1)^2+y^2)((x+1)^2+y^2) = \mu+1,\\
&MA\cdot MA' = C = \sqrt{\mu+1},
\end{aligned}$$

avec

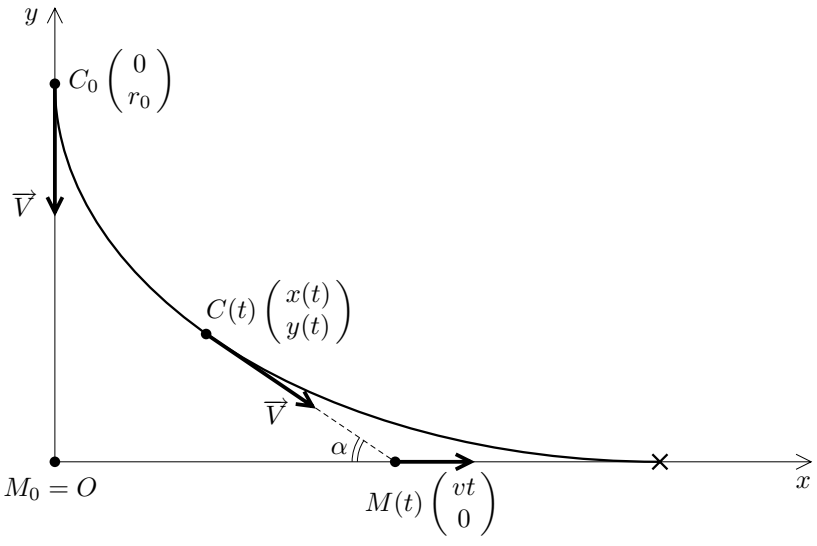
$$M\begin{pmatrix} x\\ y\end{pmatrix}, \qquad A\begin{pmatrix} A\\ 0\end{pmatrix}, \qquad A'\begin{pmatrix} -1\\ 0\end{pmatrix}.$$

Les courbes $MA\cdot MA' = C$ s'appellent des ovales de Cassini. Leur allure est la suivante.



3.3. *Courbe de poursuite du chien*

Nous présentons ici la célèbre « courbe du chien » comme exemple de courbe de poursuite. Voici le problème : un chien et son maître se déplacent l'un et l'autre à des vitesses scalaires constantes V (pour le chien) et v (pour le maître), avec $V > v$. On suppose que le maître se déplace en ligne droite, disons sur l'axe Ox, dans la direction positive, suivant la loi $x = vt$. A l'instant $t = 0$, le chien se trouve au point $x = 0$ $y = r_0$, à distance r_0 du maître. Le chien cherche à rejoindre son maître en pointant son vecteur vitesse $\overrightarrow{V}$ en direction du maître. Le problème est de déterminer la loi du mouvement $C(t)$ du chien.



Notons α l'angle (non orienté) $\alpha = (Ox, \overrightarrow{CM})$ et $r = \|\overrightarrow{CM}\|$; on a bien entendu $\alpha = \alpha(t)$ et $r = r(t)$. Comme d'habitude en Physique, on désignera par des points surlignants les dérivées temporelles $\dot{r}(t) = dr/dt$, $\dot{\alpha}(t) = d\alpha/dt$, ... A l'instant t, la position et la vitesse du chien sont données par

$$\begin{cases} x(t) = vt - r\cos\alpha, & \dot{x}(t) = v - \dot{r}\cos\alpha + r\dot{\alpha}\sin\alpha = V\cos\alpha, \\ y(t) = r\sin\alpha, & \dot{y}(t) = \dot{r}\sin\alpha + r\dot{\alpha}\cos\alpha = -V\sin\alpha, \end{cases}$$

Ces équations fournissent aisément l'expression de $\dot{r}$ et $r\dot{\alpha}$:

$$\begin{cases} \dot{r} = v\cos\alpha - V, \\ r\dot{\alpha} = -v\sin\alpha. \end{cases}$$

En prenant le quotient on élimine dt et on trouve donc

$$\frac{dr}{r\,d\alpha} = -\operatorname{cotan}\alpha + \frac{V}{v}\frac{1}{\sin\alpha}.$$

Notons $\lambda = V/v > 1$ le rapport des vitesses respectives du chien et du maître. Après intégration, et compte tenu de ce que $r = r_0$ et $\alpha = \pi/2$ quand $t = 0$, il vient

$$\ln r = -\ln\sin\alpha + \lambda \ln\tan(\alpha/2) + \text{Cte} \quad \Longrightarrow \quad r = r_0 \frac{\tan(\alpha/2)^\lambda}{\sin\alpha}$$

Nous en déduisons

$$\frac{d\alpha}{dt} = \dot{\alpha} = -\frac{v\sin\alpha}{r} = -\frac{v}{r_0}(\sin\alpha)^2 \tan(\alpha/2)^{-\lambda}.$$

En posant $\theta = \tan(\alpha/2)$ et $\sin\alpha = 2\sin\dfrac{\alpha}{2}\cos\dfrac{\alpha}{2} = \dfrac{2\theta}{1+\theta^2}$, on trouve

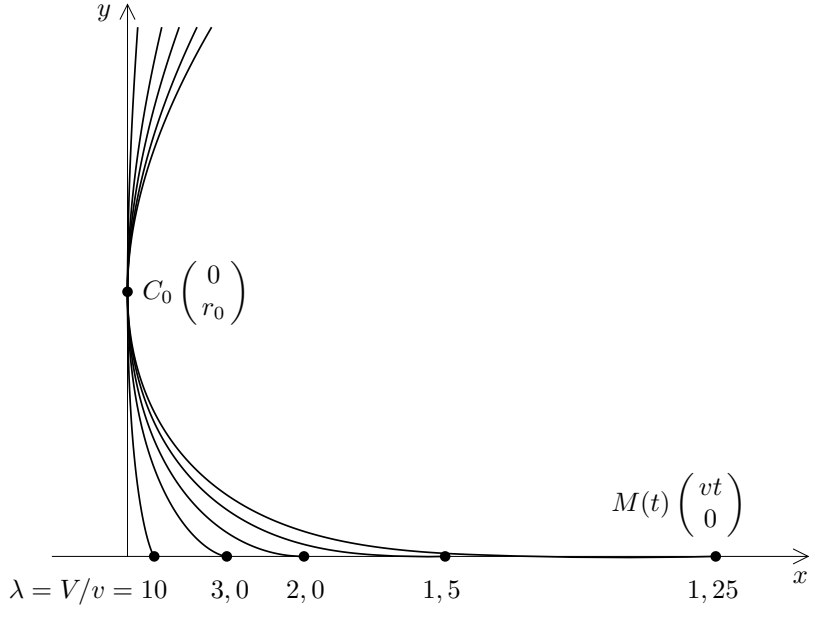
$$dt = -\frac{r_0}{v}\,\frac{\tan(\alpha/2)^\lambda}{(\sin\alpha)^2}\,d\alpha = -\frac{r_0}{2v}(1+\theta^2)\theta^{\lambda-2}d\theta,$$
$$t = \frac{r_0}{2v}\Big(\frac{1-\theta^{\lambda-1}}{\lambda-1} + \frac{1-\theta^{\lambda+1}}{\lambda+1}\Big)$$

compte tenu du fait que $\theta = 1$ en $t = 0$. Par substitution dans les expressions de x et y, et d'après l'égalité $\tan\alpha = 2\theta/(1-\theta^2)$, on obtient les équations paramétriques de la « courbe du chien », à savoir

$$\begin{cases} x = \dfrac{r_0}{2}\Big(\dfrac{1-\theta^{\lambda-1}}{\lambda-1} + \dfrac{1-\theta^{\lambda+1}}{\lambda+1}\Big) - \dfrac{r_0}{2}(1-\theta^2)\theta^{\lambda-1} \\ y = r_0\theta^\lambda \\ t = \dfrac{r_0}{2v}\Big(\dfrac{1-\theta^{\lambda-1}}{\lambda-1} + \dfrac{1-\theta^{\lambda+1}}{\lambda+1}\Big), \qquad \theta \in [0,1]. \end{cases}$$

Au terme de la poursuite ($y = \theta = 0$), le maître a parcouru la distance

$$x = \frac{\lambda}{\lambda^2-1}\,r_0 \qquad \text{pendant le temps} \quad t = \frac{\lambda}{\lambda^2-1}\,\frac{r_0}{v}.$$



Remarque – Les équations ont encore un sens lorsque $t < 0$. On a dans ce cas $\alpha \in]\pi/2, \pi[$, $\theta \in]1, +\infty[$, le chien se trouve dans le quadrant $x > 0$, $y > r_0$ et se dirige vers le maître qui parcourt de son côté la demi-droite $x < 0$.

4. Équations différentielles du second ordre

4.1. Remarques générales

On considère une équation différentielle

$$y'' = f(x, y, y') \tag{E}$$

où $f : U \to \mathbb{R}$, $U \subset \mathbb{R}^3$, est une application continue localement lipschitzienne en ses deuxième et troisième variables.

La solution générale y définie au voisinage d'un point x_0 dépend alors de deux paramètres $\lambda, \mu \in \mathbb{R}$ qui apparaissent le plus souvent comme des constantes d'intégration :

$$y(x) = \varphi(x, \lambda, \mu).$$

Le théorème de Cauchy-Lipschitz montre qu'on peut choisir $y_0 = y(x_0)$, $y_1 = y'(x_0)$ comme paramètres.

Il existe très peu de cas où on sait résoudre explicitement une équation du second ordre : même les équations linéaires du second ordre sans second membre ne se résolvent pas explicitement en général.

4.2. Équations incomplètes du second ordre

a) Équations du type (E) : $y'' = f(x, y')$

Si on considère la nouvelle fonction inconnue $v = y'$, (E) se ramène à l'équation du premier ordre

$$v' = f(x, v).$$

La solution générale de cette dernière sera de la forme $v(x, \lambda)$, $\lambda \in \mathbb{R}$, et on obtient donc

$$y(x) = \int_{x_0}^{x} v(t, \lambda)dt + \mu, \quad \mu \in \mathbb{R}.$$

b) Équations du type (E) : $y'' = f(y, y')$

La méthode consiste à prendre y comme nouvelle variable et $v = y'$ comme variable fonction inconnue (en la variable y).

• Il peut y avoir des solutions constantes $y(x) = y_0$, auquel cas y ne peut être choisi comme variable. On a donc des solutions singulières

$$y(x) = y_j, \quad \text{avec} \quad f(y_j, 0) = 0.$$

• Cas général

$$y'' = \frac{dy'}{dx} = \frac{dy}{dx} \cdot \frac{dy'}{dy} = v\,\frac{dv}{dy}$$

L'équation se ramène alors à l'équation du premier ordre

$$v\,\frac{dv}{dy} = f(y, v).$$

La résolution de cette dernière donne une solution générale $v(y, \lambda)$, $\lambda \in \mathbb{R}$. On doit ensuite résoudre

$$y' = v(y, \lambda) \Leftrightarrow \frac{dy}{v(y, \lambda)} = dx,$$

d'où la solution générale

$$\int \frac{dy}{v(y, \lambda)} = x + \mu, \qquad \mu \in \mathbb{R}.$$

c) Équations du type (E) : $y'' = f(y)$

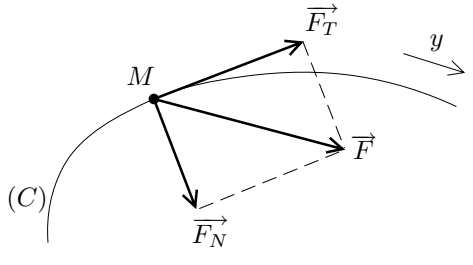
C'est un cas particulier du cas b) précédent, mais on peut ici préciser davantage la méthode de résolution. On a en effet

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

et en intégrant il vient $\frac{1}{2}\,y'^2 = \varphi(y) + \lambda$, $\lambda \in \mathbb{R}$, où φ est une primitive de f. On obtient donc

$$\begin{aligned} y' &= \pm\sqrt{2(\varphi(y) + \lambda)}, \\ &\pm \frac{dy}{\sqrt{2(\varphi(y) + \lambda)}} = dx, \\ &\pm \int_{y_0}^{y} \frac{du}{\sqrt{2(\varphi(u) + \lambda)}} = x + \mu, \quad \mu \in \mathbb{R}. \end{aligned}$$

Interprétation physique – On étudie la loi du mouvement d'un point matériel M de masse m astreint à se déplacer sur une courbe (C). On suppose que la composante tangentielle $\overrightarrow{F_T}$ de la force $\overrightarrow{F}$ qui s'exerce sur M ne dépend que de la position de M, repérée par son abscisse curviligne y sur (C).



Par hypothèse, il existe une fonction f telle que $F_T = f(y)$. Le principe fondamental de la dynamique donne

$$F_T = m\gamma_T = m\,\frac{d^2y}{dt^2},$$

d'où

$$my'' = f(y) \tag{E}$$

avec $y'' = d^2y/dt^2$. On en déduit $my'y'' - f(y)y' = 0$, donc $\frac{1}{2}\,my'^2 - \varphi(y) = \lambda$, où φ est une primitive de f. La quantité $\frac{1}{2}\,my'^2 = E_{\rm c}$ est "l'énergie cinétique" de la particule tandis que

$$-\varphi(y) = -\int f(y)dy = -\int \overrightarrow{F_T}(M)\cdot d\overrightarrow{M} = -\int \overrightarrow{F}(M)\cdot d\overrightarrow{M}$$

est "l'énergie potentielle" $E_{\rm p}$. L'énergie totale

$$E_{\rm t} = E_{\rm c} + E_{\rm p} = \frac{1}{2}\,my'^2 - \varphi(y)$$

est constante quel que soit le mouvement du point M. On dit que $U(y,y') = \frac{1}{2}\,my'^2 - \varphi(y)$ est une "intégrale première" de (E) (dans le sens que c'est une relation différentielle obtenue à l'aide d'une première intégration de l'équation du second ordre, une deuxième intégration restant nécessaire pour établir la loi du mouvement). Si $E_{\rm t}$ désigne l'énergie totale, la loi du mouvement est donnée par

$$t - t_0 = \pm\sqrt{m/2}\int_{y_0}^{y} \frac{du}{\sqrt{E_{\rm t} + \varphi(u)}}$$

au voisinage de tout donnée initiale (t_0, y_0, y_0') telle que $\frac{1}{2}my_0'^2 = E_{\rm t} + \varphi(y_0) > 0$.

Complément – Supposons que la fonction $f : \mathbb{R} \to \mathbb{R}$ soit de classe C^1, strictement décroissante et telle que $f(0) = 0$; on a donc en particulier $f(y) > 0$ pour $y < 0$ et $f(y) < 0$ pour $y < 0$ (physiquement, ceci signifie que la force est une "force de rappel" vers la position neutre $y = 0$, dont l'intensité s'accroît avec la distance à la position neutre). Alors les solutions maximales $t \mapsto y(t)$ sont *périodiques*. Pour le voir, posons par exemple $\varphi(u) = \int_0^u f(y)dy$ et observons que φ est une fonction concave négative ou nulle, passant par un maximum en $\varphi(0) = 0$. Comme $\frac{1}{2}my'^2 - \varphi(y) = E_{\rm t}$ avec $y'^2 \geq 0$ et $-\varphi(y) > 0$ si $y \neq 0$, les solutions non triviales n'existent que pour une valeur $E_{\rm t} > 0$ de l'énergie totale, et elles vérifient $|y'| \leq \sqrt{2E_{\rm t}/m}$ et $-\varphi(y) \leq E_{\rm t}$. De plus, si $a < 0$, $b > 0$ sont les uniques réels négatif et positif tels que $\varphi(a) = \varphi(b) = -E_{\rm t}$, on a $a \leq y(t) \leq b$ pour tout t. Ceci implique déjà que les solutions maximales sont définies sur $\mathbb{R}$ tout entier [si par exemple une solution maximale n'était définie que sur un intervalle ouvert $]t_1, t_2[$, le critère de Cauchy uniforme montrerait que y se prolonge par continuité à droite en t_1 et à gauche en t_2, puisque y est lipschitzienne de rapport $\leq \sqrt{2E_{\rm t}/m}$, et de même y' et y'' se prolongeraient grâce à la relation $y'' = \frac{1}{m}f(y)$; l'existence de solutions locales au voisinage de t_1 et t_2 contredirait alors la maximalité de y].

La relation $\frac{1}{2}my'^2 - \varphi(y) = E_\mathrm{t}$ montre que $y' \neq 0$ lorsque $a < y < b$ (car on a alors $-\varphi(y) < E_\mathrm{t}$). Par continuité, la fonction y' est donc de signe constant sur tout intervalle de temps où $a < y < b$. La solution explicite donnée plus haut montre que y est alternativement croissant de a à b puis décroissant de b à a, avec demi-période

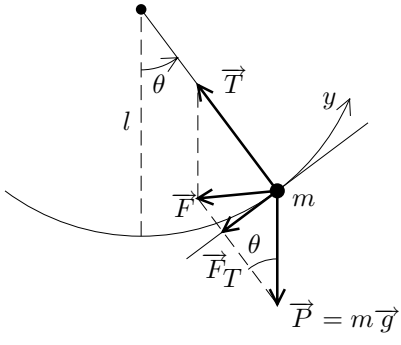
$$\frac{T}{2} = \sqrt{m/2}\int_a^b \frac{du}{\sqrt{E_\mathrm{t} + \varphi(u)}},$$

et, en choisissant t_0 tel que $y(t_0) = \min y = a$, on a les relations

$$t = t_0 + \sqrt{m/2}\int_a^y \frac{du}{\sqrt{E_\mathrm{t} + \varphi(u)}} + nT, \qquad t \in [t_0 + nT, t_0 + nT + T/2],$$
$$t = t_0 - \sqrt{m/2}\int_a^y \frac{du}{\sqrt{E_\mathrm{t} + \varphi(u)}} + (n+1)T, \quad t \in [t_0 + nT + T/2, t_0 + (n+1)T].$$

On observera que l'intégrale donnant la période T est convergente, car on a $\varphi'(a) = f(a) > 0$, $\varphi'(b) = f(b) < 0$, de sorte que $E_\mathrm{t} + \varphi(u) \sim f(a)(u-a)$ au voisinage de a, et de même au voisinage de b.

Exemple – Mouvement d'un pendule simple de masse m suspendu à un fil de longueur l.



On a ici $y = l\theta$ et $F_T = P \sin\theta = -mg \sin\theta$, d'où

$$ml\theta'' = -mg \sin\theta,$$
$$\theta'' = -\frac{g}{l} \sin\theta.$$

L'énergie totale est

$$E_\mathrm{t} = E_\mathrm{c} + E_\mathrm{p} = \frac{1}{2}my'^2 - mgl\cos\theta = \frac{1}{2}ml^2\theta'^2 - mgl\cos\theta.$$

Les solutions $t \mapsto \theta(t)$ vérifient

$$\theta'^2 - \frac{2g}{l}\cos\theta = \lambda = \frac{2E_\mathrm{t}}{ml^2}, \qquad \lambda \in \mathbb{R},$$
$$\pm\int_0^\theta \frac{d\varphi}{\sqrt{\lambda + \frac{2g}{l}\cos\varphi}} = t - t_0, \qquad t_0 \in \mathbb{R}.$$

L'intégrale ne se calcule pas explicitement, sauf si $\lambda = \frac{2g}{l}$, auquel cas

$$t - t_0 = \pm\sqrt{\frac{l}{g}} \int_0^\theta \frac{d\varphi}{2\cos\varphi/2} = \pm\sqrt{\frac{l}{g}}\, \ln\tan\Big(\frac{\theta}{4} + \frac{\pi}{4}\Big),$$

et le pendule atteint la position verticale haute $\theta = \pm\pi$ en un temps infini. Dans les autres cas, les solutions maximales sont périodiques. Si $\lambda < 2g/l$, l'équation différentielle implique $\cos\theta \geq -\lambda l/2g > -1$ et l'amplitude angulaire est donc majorée en valeur absolue par une amplitude maximale $\theta_m \in]0, \pi[$ telle que $\cos\theta_m = -\lambda l/2g$; dans ce cas le mouvement est oscillatoire autour de la position d'équilibre $\theta = 0$ (ceci correspond à la situation étudiée dans la remarque, avec une fonction $f(\theta) = -\sin\theta$ strictement décroissante sur $[-\theta_m, \theta_m]$). La demi-période est donnée par

$$\frac{T}{2} = \int_{-\theta_m}^{\theta_m} \frac{d\varphi}{\sqrt{\lambda + \frac{2g}{l}\,\cos\,\varphi}} = 2\int_0^{\theta_m} \frac{d\varphi}{\sqrt{\lambda + \frac{2g}{l}\,\cos\,\varphi}}.$$

Si $\lambda > 2g/l$, on n'est plus dans la situation de la remarque, mais on a cependant encore un mouvement périodique de demi-période

$$\frac{T}{2} = \int_0^{\pi} \frac{d\varphi}{\sqrt{\lambda + \frac{2g}{l}\,\cos\,\varphi}},$$

le pendule effectuant des rotations complètes sans jamais changer de sens de rotation.

En physique, on s'intéresse généralement aux oscillations de faible amplitude du pendule. Ceci permet de faire l'approximation usuelle $\sin\theta \simeq \theta$ et on obtient alors les solutions approchées classiques $\theta = \theta_m \cos\omega(t - t_0)$ avec $\omega = \sqrt{g/l}$. Nous reviendrons sur cette question au paragraphe 2.4 du chapitre XI, et nous indiquerons en particulier une méthode permettant d'évaluer l'erreur commise.

4.3. Équations linéaires homogènes du second ordre

La théorie générale des équations et systèmes différentiels linéaires sera faite au chapitre suivant. Indiquons un cas où l'on peut se ramener à un calcul de primitives. Soit

$$\text{(E)} \qquad a(x)y'' + b(x)y' + c(x)y = 0.$$

Supposons qu'on connaisse une solution particulière $y_{(1)}$ de (E). On peut alors chercher la solution générale par la méthode de variation des constantes :

$$y(x) = \lambda(x)y_{(1)}(x).$$

Il vient

$$a(x)\Big(\lambda'' y_{(1)} + 2\lambda' y'_{(1)} + \lambda y''_{(1)}\Big) + b(x)\Big(\lambda' y_{(1)} + \lambda y'_{(1)}\Big) + c(x)\lambda y_{(1)} = 0,$$

$$\lambda\Big(a(x)y''_{(1)} + b(x)y'_{(1)} + c(x)y_{(1)}\Big) + \lambda'\Big(2a(x)y'_{(1)} + b(x)y_{(1)}\Big) + \lambda'' a(x)y_{(1)} = 0,$$

$$\lambda'\Big(2a(x)y'_{(1)} + b(x)y_{(1)}\Big) + \lambda'' a(x)y_{(1)} = 0.$$

La fonction $\mu = \lambda'$ est donc solution d'une équation différentielle linéaire du premier ordre, qui se peut se récrire

$$\frac{\mu'}{\mu} = \frac{\lambda''}{\lambda'} = -2\,\frac{y'_{(1)}}{y_{(1)}} - \frac{b(x)}{a(x)}.$$

La solution générale est donnée par $\mu = \alpha\mu_{(1)}$, $\alpha \in \mathbb{R}$, d'où $\lambda = \alpha\lambda_{(1)} + \beta$, $\beta \in \mathbb{R}$, où $\lambda_{(1)}$ est une primitive de $\mu_{(1)}$. La solution générale de (E) est donc :

$$y(x) = \alpha\lambda_{(1)}(x)y_{(1)}(x) + \beta y_{(1)}(x), \quad (\alpha, \beta) \in \mathbb{R}^2.$$

Les solutions forment un espace vectoriel de dimension 2.

Exercice – *Résoudre* $x^2(1-x^2)y'' + x^3y' - 2y = 0$ *en observant que* $y_{(1)}(x) = x^2$ *est solution.*

4.4. ÉQUATIONS DIFFÉRENTIELLES ISSUES DE PROBLÈMES VARIATIONNELS*

Les problèmes variationnels conduisent très souvent à la résolution d'équations différentielles du second ordre. Avant de donner un exemple, nous allons résoudre un problème variationnel général dans une situation simple. On considère un opérateur fonctionnel (c'est-à-dire une fonction dont la variable est une fonction)

$$\begin{aligned} \varphi : C^2([a,b]) &\to \mathbb{R} \\ u &\mapsto \varphi(u) = \int_a^b F(x, u(x), u'(x))dx, \end{aligned}$$

où $F : [a,b] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $(x,y,z) \mapsto F(x,y,z)$ est une application de classe C^2. Le problème typique du calcul des variations est de rechercher les extrema de $\varphi(u)$ lorsque u décrit $C^2([a,b])$ avec la « contrainte aux bornes » suivante : les valeurs aux bornes de l'intervalle $u(a) = u_1$, $u(b) = u_2$ sont fixées.

Soit $h \in C^2([a,b])$ avec $h(a) = h(b) = 0$. Pour tout $t \in \mathbb{R}$, la fonction $u + th$ vérifie la même contrainte aux bornes que la fonction u. Si u est un extremum de φ sous les conditions précisées plus haut, alors $t = 0$ est un extremum de la fonction d'une variable réelle

$$\psi_h(t) = \varphi(u + th) = \int_a^b F(x, u(x) + th(x), u'(x) + th'(x))dx.$$

On doit donc avoir $\psi'(0) = 0$, et ceci quel que soit la fonction $h \in C^2([a,b])$ vérifiant $h(a) = h(b) = 0$. D'après le théorème de dérivation sous le signe somme il vient

$$\psi'_h(0) = \int_a^b \big(h(x)F'_y(x,u,u') + h'(x)F'_z(x,u,u')\big)dx.$$

En intégrant par parties le terme en $h'(x)$ on obtient

$$\psi'_h(0) = \int_a^b h(x)\Big(F'_y(x,u,u') - \frac{d}{dx}\,F'_z(x,u,u')\Big)dx.$$

Par densité de l'ensemble des fonctions h considérées dans l'espace $L^1([a,b])$ des fonctions intégrables sur $[a,b]$, on aura donc $\psi_h'(0)=0$ pour tout h si et seulement si u satisfait l'équation différentielle

$$\text{(E)} \qquad F_y'(x,u,u') - \frac{d}{dx}\,\Big(F_z'(x,u,u')\Big) = 0,$$

ou encore :

$$F_y'(x,u,u') - F_{xz}''(x,u,u') - u'F_{yz}''(x,u,u') - u''F_{zz}''(x,u,u') = 0.$$

Cette équation différentielle du second ordre en u est appelée *équation d'Euler-Lagrange* associée au problème variationnel défini par l'opérateur φ.

Application à la chaînette* – On cherche à déterminer la courbe représentant la position à l'équilibre d'un fil souple inextensible de masse linéique $\mu = dm/ds$ constante, lorsque ce fil est suspendu par ses extrémités en des points situés à la même hauteur (cette courbe est appelée « chaînette »). On admettra comme physiquement évident que la courbe cherchée est symétrique et située dans le plan vertical contenant les extrémités. Soit Oxy un repère orthonormé de ce plan tel que Oy est la verticale orienté vers le haut et passant par le point le plus bas de la courbe, les extrémités ayant pour coordonnées $(\pm a, 0)$. Soit enfin $s \in [-\ell/2, \ell/2]$ l'abscisse curviligne mesurée le long du fil avec le point le plus bas pris comme origine (ℓ désigne la longueur du fil). La position d'équilibre correspond à la position la plus basse possible du centre de gravité G. Par symétrie, on a (x étant choisi comme variable) :

$$y_G = \frac{1}{m/2}\int_0^a y\,dm = \frac{1}{\mu\ell/2}\int_0^a y\mu\,ds = \frac{2}{\ell}\int_0^a y\,ds,$$

où $m = \mu\ell$ est la masse du fil. Une intégration par parties donne

$$\begin{aligned} y_G &= \frac{2}{\ell}\,\big[ys\big]_0^a - \frac{2}{\ell}\int_0^a s\,dy \\ &= -\frac{2}{\ell}\int_0^a s\,dy = -\frac{2}{\ell}\int_0^a s\sqrt{s'^2-1}\,dx \end{aligned}$$

avec $s' = ds/dx = \sqrt{dx^2+dy^2}/dx$. Le problème revient donc à déterminer les fonctions $s = s(x)$ réalisant le maximum de l'opérateur

$$\varphi(s) = \int_0^a s\sqrt{s'^2-1}\,dx,$$

avec les contraintes $s(0)=0$, $s(a)=\ell/2$. L'équation d'Euler-Lagrange appliquée à $F(x,s,t) = s\sqrt{t^2-1}$ donne

$$\sqrt{s'^2-1} - \frac{d}{dx}\,\Big(\frac{ss'}{\sqrt{s'^2-1}}\Big) = \sqrt{s'^2-1} - \frac{s'^2+ss''}{\sqrt{s'^2-1}} + \frac{ss'^2s''}{(s'^2-1)^{3/2}} = 0.$$

Après multiplication par $(s'^2-1)^{3/2}$ on obtient l'équation

$$\text{(E)} \qquad (s'^2-1)^2 - (s'^2-1)(s'^2+ss'') + ss'^2s'' = 1 - s'^2 + ss'' = 0.$$

On résout cette équation grâce à la méthode décrite au paragraphe 4.2.b), consistant à choisir s comme nouvelle variable et $v = s'$ comme nouvelle fonction inconnue. Il vient successivement

$$\begin{aligned}
s'' &= \frac{ds'}{dx} = \frac{ds}{dx}\cdot\frac{ds'}{ds} = v\,\frac{dv}{ds},\\
\text{(E)} &\Rightarrow 1 - v^2 + sv\,\frac{dv}{ds} = 0,\\
\frac{ds}{s} &= \frac{vdv}{v^2-1} \Rightarrow \ln\, s = \frac{1}{2}\,\ln(v^2-1) + C,\\
s &= \lambda\sqrt{v^2-1} = \lambda\sqrt{s'^2-1},\\
s' &= \frac{ds}{dx} = \sqrt{1+\frac{s^2}{\lambda^2}} \Rightarrow dx = \frac{ds}{\sqrt{1+s^2/\lambda^2}},\\
x &= \lambda\ \mathrm{Arg\,sinh}\,\frac{s}{\lambda} \Rightarrow s = \lambda\,\sinh\,\frac{x}{\lambda},\\
\frac{ds}{dx} &= \sqrt{1+\Big(\frac{dy}{dx}\Big)^2} = \mathrm{ch}\,\frac{x}{\lambda} \Rightarrow \frac{dy}{dx} = \sinh\,\frac{x}{\lambda}.
\end{aligned}$$

On en déduit l'équation de la chaînette, en tenant compte du fait que $y(a) = 0$:

$$y = \lambda\Big(\cosh\,\frac{x}{\lambda} - \cosh\,\frac{a}{\lambda}\Big).$$

Le paramètre λ se calcule à partir de la relation $\lambda\sinh a/\lambda = \ell/2$, obtenue en égalant $s(a) = \ell/2$.

Remarque – Notre raisonnement n'est pas parfaitement rigoureux dans la mesure où $F(x,s,t) = s\sqrt{t^2-1}$ est de classe C^2 seulement sur $\mathbb{R}\times\mathbb{R}\times\{|t|>1\}$, alors que $|s'|$ est ≥ 1 mais prend la valeur 1 pour $x = 0$ (on notera que $ds/dx = 1/\cos\theta$ où θ est l'angle de la tangente à la courbe avec l'axe $0x$). Supposons $s'(x) > 1$ pour $x > 0$, comme c'est le cas pour la solution physique observée. Le raisonnement de dérivation sous la signe somme et l'intégration par parties appliqués dans les considérations générale du début fonctionnent encore pour $|t|$ petit si on suppose $h(x) = 0$ sur un voisinage de 0 (et aussi bien sûr $h(a) = 0$).

Ces fonctions h sont encore denses dans $L^1([0,a])$, donc s doit effectivement satisfaire l'équation différentielle (E) sur $]0,a]$.

Calcul de géodésiques** – Nous étudions ici une autre application importante du calcul des variations, à savoir le calcul des géodésiques d'une surface (ou d'une variété de dimension plus grande). Si nous avons une surface $S\subset\mathbb{R}^3$ donnée comme un graphe $z = h(x,y)$ d'une fonction $h:\Omega\to\mathbb{R}$ sur un ouvert $\Omega\subset\mathbb{R}^2$, l'élément de longueur infinitésimal de la surface S est donné pour tout $(x,y)\in\Omega$ par

$$\begin{aligned}
ds^2 &= dx^2 + dy^2 + dz^2 = dx^2 + dy^2 + (h'_x dx + h'_y dy)^2\\
&= (1+h'^2_x)dx^2 + 2h'_x h'_y dx\,dy + (1+h'^2_y)dy^2.
\end{aligned}$$

Plus généralement, une *métrique riemannienne* sur un ouvert $\Omega \subset \mathbb{R}^m$ est une expression de l'élément de longueur infinitésimal par une forme quadratique définie positive, dépendant du point $x \in \Omega$ considéré :

$$ds^2 = q(x, dx) = \sum_{1 \le i,j \le m} a_{ij}(x)\, dx_i dx_j$$

(avec une matrice symétrique $(a_{ij}(x))$ définie positive). On supposera en outre que les coefficients $a_{ij}(x)$ sont suffisamment réguliers, disons de classe C^2. Étant donné une courbe $\gamma : [a, b] \to \Omega$ de classe C^1, sa longueur (riemannienne) est par définition

$$ds = \|\gamma'(t)\, dt\|_q = \sqrt{q\big(\gamma(t), \gamma'(t)dt\big)} = \sqrt{\sum_{1 \le i,j \le m} a_{ij}(\gamma(t))\, \gamma_i'(t)\gamma_j'(t)}\, dt,$$

$$\mathrm{long}(\gamma) = \int_a^b ds = \int_a^b \sqrt{\sum_{1 \le i,j \le m} a_{ij}(\gamma(t))\, \gamma_i'(t)\gamma_j'(t)}\, dt.$$

Pour deux points $x, y \in \Omega$, la distance géodésique $d_q(x, y)$ est par définition $\inf_\gamma \mathrm{long}(\gamma)$ pour tous les chemins $\gamma : [a, b] \to \Omega$ de classe C^1 d'extrêmités $\gamma(a) = x$, $\gamma(b) = y$. Si un chemin réalise l'infimum, on dit qu'il s'agit d'une géodésique de la métrique riemannienne (on notera qu'en général un tel chemin n'existe pas nécessairement, et s'il existe il peut ne pas être unique). Un problème fondamental est de déterminer l'équation des géodésiques afin entre autres de calculer la distance géodésique. Pour cela il est commode d'introduire « l'énergie d'un chemin » qui est par définition

$$E(\gamma) = \int_a^b \|\gamma'(t)\|_q^2\, dt = \int_a^b \sum_{1 \le i,j \le m} a_{ij}(\gamma(t))\, \gamma_i'(t)\gamma_j'(t)\, dt.$$

L'inégalité de Cauchy-Schwarz donne

$$\Big(\int_a^b \|\gamma'(t)\|_q\, dt\Big)^2 = \Big(\int_a^b 1 \cdot \|\gamma'(t)\|_q\, dt\Big)^2 \le (b-a) \int_a^b \|\gamma'(t)\|_q^2\, dt$$

soit $\mathrm{long}(\gamma) \le \big((b-a)E(\gamma)\big)^{1/2}$, avec égalité si et seulement si

$$\frac{ds}{dt} = \|\gamma'(t)\|_q = \mathrm{Cte},$$

condition qui peut toujours être réalisée en reparamétrisant le chemin γ par son abscisse curviligne s. Il en résulte que les chemins qui minimisent l'énergie sont exactement les géodésiques paramétrées par l'abscisse curviligne (à un facteur constant près). Or la fonctionnelle d'énergie $\gamma \mapsto E(\gamma)$ admet pour différentielle

$$\begin{aligned} E'(\gamma) \cdot h &= \int_a^b \Big(\sum_{i,j,k} \frac{\partial a_{ij}}{\partial x_k}(\gamma(t))\, \gamma_i'(t)\gamma_j'(t)h_k(t) + 2\sum_{i,j} a_{ij}(\gamma(t))\, \gamma_i'(t)h_j'(t) \Big)\, dt \\ &= \int_a^b \sum_k h_k(t) \Big(\sum_{i,j} \frac{\partial a_{ij}}{\partial x_k}(\gamma(t))\, \gamma_i'(t)\gamma_j'(t) - 2\frac{d}{dt}\sum_i a_{ik}(\gamma(t))\, \gamma_i'(t) \Big)\, dt \end{aligned}$$

après intégration par parties (on suppose bien sûr $h_j(a) = h_j(b) = 0$). Il en résulte que le coefficient de chaque terme $h_k(t)$ doit être identiquement nul. En multipliant par $-1/2$ et en développant la dérivée d/dt, on obtient le système d'équations d'Euler-Lagrange caractérisant les géodésiques :

$$\sum_i a_{ik}(\gamma(t))\,\gamma_i''(t) + \sum_{i,j}\Big(\frac{\partial a_{ik}}{\partial x_j} - \frac{1}{2}\frac{\partial a_{ij}}{\partial x_k}\Big)(\gamma(t))\,\gamma_i'(t)\gamma_j'(t) = 0, \qquad 1 \le k \le m.$$

5. PROBLÈMES

5.1. On considère l'équation différentielle à variables séparées

$$(F_\alpha) \qquad \frac{dy}{dt} = y^\alpha + 1, \qquad \alpha > 0.$$

(a) Exprimer la solution générale de (F_α) en introduisant la fonction auxiliaire

$$G(y) = \int_0^y \frac{dx}{x^\alpha + 1}$$

(b) Plus précisément :

- Déterminer (en distinguant les deux cas $0 < \alpha \le 1$ et $\alpha > 1$) le comportement de $G(y)$ sur $[0, +\infty[$;
- en déduire dans chaque cas l'allure des solutions maximales de (F_α) ;
- traiter complètement et explicitement les deux cas $\alpha = 1$ et $\alpha = 2$.

5.2. On considère l'équation différentielle

$$xy' - y^2 + (2x+1)y = x^2 + 2x.$$

(a) Possède-t-elle une solution particulière de type polynôme ? En donner la solution générale.

(b) Quelle est l'équation de l'isocline de pente 0 dans le nouveau repère de vecteurs de base $((1,1),(0,1))$? Dessiner cette isocline en précisant les tangentes aux points d'abscisse 0 dans l'ancien repère.

(c) Dessiner l'allure générale des solutions.

(d) Soit (x_0, y_0) un point de $\mathbb{R}^2$. Combien passe-t-il de solutions maximales de classe C^1 par (x_0, y_0) ? On précisera l'intervalle de définition de ces solutions et le cas échéant on indiquera les solutions globales.

5.3. On considère l'équation différentielle

$$\frac{dy}{dt} = y^2 - (2x-1)y + x^2 - x + 1.$$

(a) Déterminer explicitement les solutions de cette équation ; on pourra commencer par chercher s'il existe des solutions polynomiales simples.

(b) Montrer que les courbes intégrales maximales correspondant à des solutions non polynomiales forment deux familles de courbes se déduisant les unes des autres par translations. Tracer celles de ces courbes qui sont asymptotes à l'axe $y'Oy$.

5.4. On considère l'équation différentielle (1) $y'^2 = yy' + x$, où y est une fonction de x à valeurs réelles, de classe C^1 par morceaux.

(a) Par quels points (x, y) de $\mathbb{R}^2$ passe-t-il une solution de (1) ? Faire un graphique.

(b) En paramétrant (1) par $dy = tdx$, montrer que y est solution d'une équation différentielle (2) $f\big(y, t, \frac{dy}{dt}\big) = 0$.

(c) Intégrer (2) puis (1) ; on pourra poser $t = \tan\varphi$ avec $\varphi \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$. On obtient une famille de courbes C_λ dépendant d'un paramètre λ.

(d) Pour $\lambda = 0$ préciser les limites quand $\varphi \to \frac{\pi}{2} - 0$ de $x, y, y/x$, puis préciser $\frac{dx}{d\varphi}$ et $\frac{dy}{d\varphi}$. Quelle est la norme euclidienne du vecteur $\big(\frac{dx}{d\varphi}, \frac{dy}{d\varphi}\big)$? On se rappellera que $\frac{dy}{dx} = \tan\varphi$.

(e) On pose $z(\varphi) = \sqrt{\big(\frac{dx}{d\varphi}\big)^2 + \big(\frac{dy}{d\varphi}\big)^2}$ pour $0 \leq \varphi \leq \frac{\pi}{2}$. Etudier les variations de $z(\varphi)$ puis tracer la courbe C_0.

5.5. On considère la famille de paraboles (P_λ) d'équation

$$P_\lambda : x = y^2 + \lambda y.$$

(a) Montrer que ces courbes sont solutions d'une équation différentielle du premier ordre que l'on précisera.

(b) Déterminer la famille des courbes orthogonales aux courbes (P_λ).

5.6. On considère la famille de courbes (C_λ) dans $\mathbb{R}^2$ définies par l'équation

$$x^2 - y^2 + \lambda y^3 = 0,$$

où λ est un paramètre réel.

(a) Tracer les courbes (C_0), (C_1) dans un repère orthonormé Oxy (unité : 4 cm). Quelle relation existe-t-il entre (C_1) et (C_λ) ?

(b) Montrer que les courbes (C_λ) sont solutions d'une équation différentielle du premier ordre.

(c) Déterminer l'équation des trajectoires orthogonales aux courbes (C_λ). Quelle est la nature de ces courbes ? Représenter la trajectoire orthogonale passant par le point $(1, 0)$ sur le même schéma que (C_0) et (C_1).

5.7. On considère dans le plan euclidien $\mathbb{R}^2$ la famille de courbes

$$(C_\lambda) \qquad x^4 = y^4 + \lambda x.$$

(a) Déterminer l'équation différentielle vérifiée par la famille (C_λ).

(b) Écrire l'équation différentielle des trajectoires orthogonales aux courbes (C_λ).

En observant que cette équation est d'un type classique, déterminer l'équation des trajectoires orthogonales.

5.8. On considère le problème de Cauchy

$$(\mathrm{P}) \qquad y' = t^2 + y^2 + 1\,; \qquad y(0) = 0.$$

(a) Soient T et R deux réels > 0, et soit $\Omega(T, R)$ le rectangle défini par les inégalités

$$0 \le t \le T\,; \quad -R \le y \le R.$$

(α) Montrer que si la condition

$$T^2 + R^2 + 1 \le R/T$$

est vérifiée, alors $\Omega(T, R)$ est un rectangle de sécurité pour (P).

(β) Montrer que pour $T > 0$ suffisamment petit, par exemple pour $T < T_0 = \sqrt{\frac{\sqrt{2}-1}{2}}$, il existe $R > 0$ tel que $\Omega(T, R)$ soit un rectangle de sécurité pour (P).

(γ) Montrer que $\Omega(1/3, 2)$ est un rectangle de sécurité pour (P), et en déduire avec précision que (P) admet une solution y et une seule sur l'intervalle $[0, 1/3]$.

(b) On va montrer que la solution y de (P) mise en évidence en (a) (γ) ne se prolonge pas à $[0, +\infty[$. On introduit à cet effet le problème de Cauchy auxiliaire

$$(\mathrm{P}_1) \qquad z' = z^2 + 1, \qquad z(0) = 0,$$

où z désigne une nouvelle fonction inconnue de t.

(α) Déterminer explicitement l'unique solution z de (P_1), et indiquer son intervalle de définition maximal.

(β) Soit $[0, T]$ un intervalle sur lequel y et z soient simultanément définies. Montrer que $u = y - z$ est solution d'un problème de Cauchy

$$(\mathrm{P}_2) \qquad u' = a(t)u + b(t)\,; \qquad u(0) = 0\,;$$

où b vérifie $b(t) \ge 0$ pour tout t dans $[0, T]$. En déduire que $y(t) \ge z(t)$ pour tout t dans $[0, T]$.

(γ) Déduire de ce qui précède que si $T \geq \frac{\pi}{2}$, alors y ne se prolonge certainement pas jusqu'à $t = T$.

(δ) Tracer avec le maximum de précision possible le graphe de y sur son intervalle de définition maximal $[0, T_1[$ (forme du graphe au voisinage de 0 ; sens de variation, convexité ; asymptote ; etc...).

5.9****.** On appelle métrique de Poincaré du disque unité $D = \{|z| < 1\}$ du plan complexe la métrique riemannienne

$$ds^2 = \frac{|dz|^2}{(1-|z|^2)^2} = \frac{dx^2+dy^2}{(1-(x^2+y^2))^2}, \qquad z = x+iy.$$

(a) Montrer (avec les notations du § 4.4, et en gardant les fonctions complexes dans les calculs) que la différentielle de l'énergie est donnée par

$$E(\gamma)\cdot h = 2\,\mathrm{Re}\int_a^b \Big(\frac{\gamma'(t)}{\big(1-\gamma(t)\overline{\gamma(t)}\big)^2}\overline{h'(t)} + 2\frac{\gamma(t)\gamma'(t)\overline{\gamma'(t)}}{\big(1-\gamma(t)\overline{\gamma(t)}\big)^3}\overline{h(t)}\Big)dt.$$

En déduire que l'équation d'Euler-Lagrange des géodésiques est

$$\gamma''(t) + \frac{2\gamma'(t)^2\overline{\gamma(t)}}{1-\gamma(t)\overline{\gamma(t)}} = 0.$$

(b) Montrer que le chemin $\gamma(t) = \tanh(kt)$ (qui décrit le diamètre $]-1,1[$ du disque) est solution de l'équation pour tout $k \in \mathbb{R}_+^*$.

(c) Montrer que si $t \mapsto \gamma(t)$ est solution, alors $t \mapsto \lambda\gamma(t)$ est encore solution pour tout nombre complexe λ de module 1, et également que $h_a \circ \gamma$ est solution, pour toute homographie complexe h_a de la forme

$$h_a(z) = \frac{z+a}{1+\overline{a}z}, \qquad a \in D.$$

(d) En utilisant un argument d'unicité des solutions du problème de Cauchy, montrer que les géodésiques sont toutes données par $\gamma(t) = h_a(\lambda\tanh(kt))$, $a \in D$, $|\lambda| = 1$, $k \in \mathbb{R}_+^*$ [on pourra calculer $\gamma(0)$ et $\gamma'(0)$].

(e) Montrer que les trajectoires des géodésiques sont les diamètres et les arcs de cercle orthogonaux au cercle unité $|z| = 1$.

CHAPITRE VII

SYSTÈMES DIFFÉRENTIELS LINÉAIRES

Les systèmes différentiels linéaires ont une grande importance pratique, car de nombreux phénomènes naturels peuvent se modéliser par de tels systèmes, au moins en première approximation. On sait d'autre part résoudre complétement les systèmes à coefficients constants, le calcul des solutions se ramenant à des calculs d'algèbre linéaire (diagonalisation ou triangulation de matrices). Dans toute la suite, $\mathbb{K}$ désigne l'un des corps $\mathbb{R}$ ou $\mathbb{C}$.

1. GÉNÉRALITÉS

1.1. DÉFINITION

Un système différentiel linéaire du premier ordre dans $\mathbb{K}^m$ est une équation

$$\frac{dY}{dt} = A(t)Y + B(t) \tag{E}$$

où $Y(t) = \begin{pmatrix} y_1(t) \\ \vdots \\ y_m(t) \end{pmatrix} \in \mathbb{K}^m$ est la fonction inconnue et où

$$A(t) = (a_{ij}(t))_{1 \leq i,j \leq m} \in M_m(\mathbb{K}), \quad B(t) = \begin{pmatrix} b_1(t) \\ \vdots \\ b_m(t) \end{pmatrix} \in \mathbb{K}^m$$

sont des fonctions *continues* données :

$$A : I \to M_m(\mathbb{K}) = \{\text{matrices carrées } m \times m \text{ sur } \mathbb{K}\},$$
$$B : I \to \mathbb{K}^m,$$

définies sur un intervalle $I \subset \mathbb{R}$.

On observe que la fonction $f(t, Y) = A't)Y + B(t)$ est continue sur $I \times \mathbb{K}^m$ et lipschitzienne en Y de rapport

$$k(t) = |||A(t)|||.$$

D'après le critère V 3.4 sur l'existence de solutions globales, on peut énoncer :

Théorème – *Par tout point $(t_0, V_0) \in I \times \mathbb{K}^m$ il passe une solution maximale unique, définie sur I tout entier.*

1.2. Cas d'un système linéaire sans second membre

On entend par là un système linéaire avec $B = 0$ identiquement :

$$\text{(E}_0\text{)} \qquad \frac{dY}{dt} = A(t)Y.$$

Soit $\mathcal{S}$ l'ensemble des solutions maximales. Alors pour tous $Y_{(1)}, Y_{(2)} \in \mathcal{S}$ et tous scalaires $\lambda_1, \lambda_2 \in \mathbb{K}$ on a $\lambda_1 Y_{(1)} + \lambda_2 Y_{(2)} \in \mathcal{S}$, donc $\mathcal{S}$ est un $\mathbb{K}$-*espace vectoriel*. Considérons l'application d'évaluation au temps t_0 :

$$\begin{aligned} \phi_{t_0} : \mathcal{S} &\to \mathbb{K}^m \\ Y &\mapsto Y(t_0). \end{aligned}$$

ϕ_{t_0} est un isomorphisme linéaire, la surjectivité provenant du théorème d'existence, et l'injectivité du théorème d'unicité relatif au problème de Cauchy.

Conséquence – *L'ensemble $\mathcal{S}$ des solutions maximales est un espace vectoriel de dimension m sur $\mathbb{K}$.*

1.3. Cas général

Revenons au système linéaire le plus général :

$$\text{(E)} \qquad \frac{dY}{dt} = A(t)Y + B(t).$$

On sait qu'il existe au moins une solution globale $Y_{(1)}$. Si Y est une solution quelconque, il est clair que $Z = Y - Y_{(1)}$ satisfait l'équation sans second membre $\text{(E}_0\text{)}$: $dZ/dt = A(t)Z$, et réciproquement. Par conséquent, l'ensemble des solutions maximales est donné par

$$Y_{(1)} + \mathcal{S} = \{Y_{(1)} + Z\,;\ Z \in \mathcal{S}\},$$

où $\mathcal{S}$ est l'ensemble des solutions maximales de l'équation sans second membre $\text{(E}_0\text{)}$ associée. L'ensemble $Y_{(1)} + \mathcal{S}$ des solutions est un translaté de $\mathcal{S}$, c'est donc un espace affine de dimension m sur $\mathbb{K}$, admettant $\mathcal{S}$ comme direction vectorielle.

2. Systèmes différentiels linéaires à coefficients constants

Ce sont les systèmes de la forme

$$\text{(E)} \qquad \frac{dY}{dt} = AY + B(t)$$

où la matrice $A = (a_{ij}) \in M_n(\mathbb{K})$ est indépendante de t.

2.1. Solutions exponentielles élémentaires de $\dfrac{dY}{dt} = AY$

On cherche une solution de la forme $Y(t) = e^{\lambda t}V$ où $l \in \mathbb{K}$, $V \in \mathbb{K}^m$ sont des constantes. Cette fonction est solution si et seulement si $\lambda e^{\lambda t}V = e^{\lambda t}AV$, soit

$$AV = \lambda V.$$

On est donc amené à chercher les valeurs propres et les vecteurs propres de A.

Cas simple : A est diagonalisable.

Il existe alors une base $(V_1, \ldots, V_m)$ de $\mathbb{K}^m$ constituée de vecteurs propres de A, de valeurs propres respectives $\lambda_1, \ldots, \lambda_m$. On obtient donc m solutions linéairement indépendantes

$$t \mapsto e^{\lambda_j t}V_j, \quad 1 \leq j \leq m.$$

La solution générale est donnée par

$$Y(t) = \alpha_1 e^{\lambda_1 t}V_1 + \ldots + \alpha_m e^{\lambda_m t}V_m, \quad \alpha_j \in \mathbb{K}.$$

Lorsque A est n'est pas diagonalisable, on a besoin en général de la notion d'exponentielle d'une matrice. Toutefois le cas des systèmes 2×2 à coefficients constants est suffisamment simple pour qu'on puisse faire les calculs « à la main ». Le lecteur pourra se reporter au § X 2.2 pour une étude approfondie de ce cas.

2.2. Exponentielle d'une matrice

La définition est calquée sur celle de la fonction exponentielle complexe usuelle, calculée au moyen du développement en série entière.

Définition – *Si $A \in M(\mathbb{K})$, on pose* $e^A = \displaystyle\sum_{n=0}^{+\infty} \frac{1}{n!}\, A^n$.

Munissons $M_n(\mathbb{K})$ de la norme $|||\quad|||$ des opérateurs linéaires sur $\mathbb{K}^m$ associée à la norme euclidienne (resp. hermitienne) de $\mathbb{R}^m$ (resp. $\mathbb{C}^m$). On a alors

$$\left|\left|\left|\frac{1}{n!}\, A^n\right|\right|\right| \leq \frac{1}{n!}\, |||A|||^n,$$

de sorte que la série $\sum \frac{1}{n!}\, A^n$ est absolument convergente. On voit de plus que

$$|||e^A||| \leq e^{|||A|||}.$$

Propriété fondamentale – *Si $A, B \in M_m(\mathbb{K})$ commutent ($AB = BA$), alors*

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

Vérification. On considère la série produit $\sum \frac{1}{p!}\ A^p \cdot \sum \frac{1}{q!}\ B^q$, dont le terme général est

$$C_n = \sum_{p+q=n} \frac{1}{p!q!}\ A^p B^q = \frac{1}{n!} \sum_{p=0}^{n} \frac{n!}{p!(n-p)!}\ A^p B^{n-p} = \frac{1}{n!}\ (A+B)^n$$

d'après la formule du binôme (noter que cette formule n'est vraie que si A et B commutent). Comme les séries de e^A et e^B sont absolument convergentes, on en déduit

$$e^A \cdot e^B = \sum_{n=0}^{+\infty} C_n = e^{A+B}.$$ ■

On voit en particulier que e^A est une *matrice inversible*, d'inverse e^{-A}.

Remarque – La propriété fondamentale tombe en défaut lorsque A et B ne commutent pas. Le lecteur pourra par exemple calculer $e^A \cdot e^B$ et e^{A+B} avec

$$A = \begin{pmatrix} 0 & 0 \\ \theta & 0 \end{pmatrix} \quad \text{et} \quad B = \begin{pmatrix} 0 & -\theta \\ 0 & 0 \end{pmatrix},$$

avec $\theta \in \mathbb{R}$. Que remarque-t-on ?

Méthode générale de calcul dans $M_n(\mathbb{C})$ – Toute matrice $A \in M_n(\mathbb{C})$ peut être mise sous forme de blocs triangulaires correspondant aux différents sous-espaces caractéristiques de A. Il existe donc une matrice de passage P, dont les colonnes sont constituées par des vecteurs formant des bases des sous-espaces caractéristiques, telle que

$$T = P^{-1}AP$$

soit une matrice triangulaire de la forme

$$T = \begin{pmatrix} \boxed{T_1} & & 0 \\ & \boxed{T_2} & \\ & \ddots & \\ 0 & & \boxed{T_s} \end{pmatrix}, \qquad T_j = \begin{pmatrix} \lambda_j & * & \dots & * \\ 0 & \lambda_j & \ddots & \vdots \\ \vdots & & \ddots & * \\ 0 & \dots & 0 & \lambda_j \end{pmatrix},$$

où $\lambda_1, \dots, \lambda_s$ sont les valeurs propres distinctes de A. On a alors de façon évidente

$$T^n = \begin{pmatrix} T_1^n & & 0 \\ & T_2^n & \\ & \ddots & \\ 0 & & T_s^n \end{pmatrix}, \qquad e^T = \begin{pmatrix} e^{T_1} & & 0 \\ & e^{T_2} & \\ & \ddots & \\ 0 & & e^{T_s} \end{pmatrix}$$

Comme $A = PTP^{-1}$, il vient $A^n = PT^nP^{-1}$, d'où

$$e^A = Pe^TP^{-1} = P \begin{pmatrix} e^{T_1} & & 0 \\ & e^{T_2} & \\ & \ddots & \\ 0 & & e^{T_s} \end{pmatrix} P^{-1}$$

On est donc ramené à calculer l'exponentielle e^B lorsque B est un bloc triangulaire de la forme

$$B = \begin{pmatrix} \lambda & * & \dots & * \\ 0 & \lambda & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & \lambda \end{pmatrix} = \lambda I + N \in M_p(\mathbb{K}),$$

où I est la matrice unité et N une matrice nilpotente $N = \begin{pmatrix} 0 & & * \\ \vdots & \ddots & \\ 0 & \dots & 0 \end{pmatrix}$ triangulaire supérieure. La puissance N^n comporte n diagonales nulles à partir de la diagonale principale (celle-ci incluse), en particulier $N^n = 0$ pour $n \geq p$. On obtient donc

$$e^N = I + \frac{1}{1!}\, N + \ldots + \frac{1}{(p-1)!}\, N^{p-1} = \begin{pmatrix} 1 & * & \dots & * \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & 1 \end{pmatrix}.$$

Comme I et N commutent, il vient finalement

$$e^B = e^{\lambda I} e^N = e^\lambda e^N \qquad (\text{car } e^{\lambda I} = e^\lambda I).$$

Formule – $\det(e^A) = \exp(\mathrm{tr}(A))$.

Vérification. Dans le cas d'un bloc triangulaire $B \in M_p(\mathbb{K})$, on trouve

$$\det(e^B) = (e^\lambda)^p \det(e^N) = e^{p\lambda} = \exp(\mathrm{tr}(B)).$$

On en déduit donc

$$\det(e^T) = \det(e^{T_1}) \ldots \det(e^{T_s}) = \exp(\mathrm{tr}(T_1) + \ldots + \mathrm{tr}(T_s)) = \exp(\mathrm{tr}(T))$$

Comme $A = PTP^{-1}$ et $e^A = Pe^T P^{-1}$, on a finalement

$$\det(e^A) = \det(e^T), \quad \mathrm{tr}(A) = \mathrm{tr}(T) = \sum \text{ valeurs propres.}$$ ■

2.3. Solution générale du système sans second membre $\dfrac{dY}{dt} = AY$

L'une des propriétés fondamentales de l'exponentiation des matrices réside dans le fait qu'elle est intimement liée à la résolution des équations linéaires à coefficients constants $dY/dt = AY$.

Théorème – *La solution Y telle que $Y(t_0) = V_0$ est donnée par*

$$Y(t) = e^{(t-t_0)A} \cdot V_0, \qquad \forall t \in \mathbb{R}.$$

Démonstration. On a $Y(t_0) = e^0 \cdot V_0 = IV_0 = V_0$.

D'autre part, la série entière

$$e^{tA} = \sum_{n=0}^{+\infty} \frac{1}{n!}\, t^n A^n$$

est de rayon de convergence $+\infty$. On peut donc dériver terme à terme pour tout $t \in \mathbb{R}$:

$$\frac{d}{dt}\,(e^{tA}) = \sum_{n=1}^{+\infty} \frac{1}{(n-1)!}\, t^{n-1} A^n = \sum_{p=0}^{+\infty} \frac{1}{p!}\, t^p A^{p+1},$$

$$\frac{d}{dt}\,(e^{tA}) = A \cdot e^{tA} = e^{tA} \cdot A.$$

Par conséquent, on a bien

$$\frac{dY}{dt} = \frac{d}{dt}\left(e^{(t-t_0)A} \cdot V_0\right) = Ae^{(t-t_0)A} \cdot V_0 = AY(t). \qquad \blacksquare$$

En prenant $t_0 = 0$, on voit que la solution générale est donnée par $Y(t) = e^{tA} \cdot V$ avec $V \in \mathbb{K}^m$.

Le calcul de e^{tA} se ramène au cas d'un bloc triangulaire $B = \lambda I + N \in M_p(\mathbb{C})$. Dans ce cas on a $e^{tB} = e^{\lambda t I} e^{tN} = e^{\lambda t} e^{tN}$, avec

$$e^{tN} = \sum_{n=0}^{p-1} \frac{t^n}{n!}\, N^n = \begin{pmatrix} 1 & Q_{12}(t) & \dots & Q_{1p}(t) \\ & 1 & Q_{23}(t) & \vdots \\ & & \ddots & \ddots \\ & & & 1 & Q_{p-1\,p}(t) \\ 0 & & & & 1 \end{pmatrix}$$

où $Q_{ij}(t)$ est un polynôme de degré $\leq j - i$, avec $Q_{ij}(0) = 0$. Les composantes de $Y(t)$ sont donc toujours des fonctions exponentielles-polynômes $\sum_{1 \leq j \leq s} P_j(t) e^{\lambda_j t}$ où $\lambda_1, \dots, \lambda_s$ sont les valeurs propres *complexes* de A (même si $\mathbb{K} = \mathbb{R}$).

2.4. Solution générale de $\dfrac{dY}{dt} = AY + B(t)$

Si aucune solution évidente n'apparaît, on peut utiliser la *méthode de variation des constantes*, c'est-à-dire qu'on cherche une solution particulière sous la forme

$$Y(t) = e^{tA} \cdot V(t)$$

où V est supposée différentiable. Il vient

$$\begin{aligned} Y'(t) &= Ae^{tA} \cdot V(t) + e^{tA} \cdot V'(t) \\ &= AY(t) + e^{tA} \cdot V'(t). \end{aligned}$$

Il suffit donc de choisir V telle que $e^{tA} \cdot V'(t) = B(t)$, soit par exemple

$$V(t) = \int_{t_0}^{t} e^{-uA} B(u)du, \quad t_0 \in I.$$

On obtient ainsi la solution particulière

$$Y(t) = e^{tA} \int_{t_0}^{t} e^{-uA} B(u)du = \int_{t_0}^{t} e^{(t-u)A} B(u)du,$$

qui est la solution telle que $Y(t_0) = 0$. La solution générale du problème de Cauchy telle que $Y(t_0) = V_0$ est donc

$$Y(t) = e^{(t-t_0)A} \cdot V_0 + \int_{t_0}^{t} e^{(t-u)A} B(u)du.$$

Exemple – Une particule de masse m et de charge électrique q se déplace dans $\mathbb{R}^3$ sous l'action d'un champ magnétique $\overrightarrow{B}$ et d'un champ électrique $\overrightarrow{E}$ uniformes et indépendants du temps. Quelle est la trajectoire de la particule ?
Si $\overrightarrow{V}$ et $\overrightarrow{\gamma}$ désignent respectivement la vitesse et l'accélération, la loi de Lorentz et le principe fondamental de la dynamique donnent l'équation

$$\overrightarrow{F} = m\overrightarrow{\gamma} = q\overrightarrow{V} \wedge \overrightarrow{B} + q\overrightarrow{E},$$

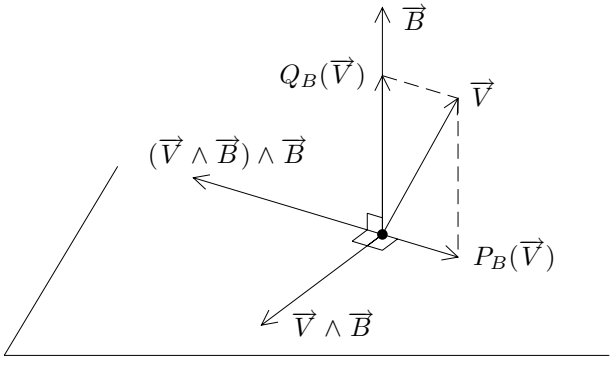
d'où

$$\overrightarrow{\gamma} = \frac{d\overrightarrow{V}}{dt} = \frac{q}{m}\ \overrightarrow{V} \wedge \overrightarrow{B} + \frac{q}{m}\ \overrightarrow{E}.$$

Il s'agit d'un système linéaire où la matrice A (à coefficients constants) est la matrice de l'application linéaire $\overrightarrow{V} \mapsto \frac{q}{m}\overrightarrow{V} \wedge \overrightarrow{B}$. On confondra dans la suite A avec cette application linéaire. Un calcul simple montre que

$$A^2(\overrightarrow{V}) = \Big(\frac{q}{m}\Big)^2 (\overrightarrow{V} \wedge \overrightarrow{B}) \wedge \overrightarrow{B} = -\Big(\frac{q}{m}\Big)^2 B^2 P_B(\overrightarrow{V})$$

où $B = \|\overrightarrow{B}\|$ et où P_B désigne le projection orthogonale sur la plan vectoriel de vecteur normal $\overrightarrow{B}$. Le schéma est le suivant :



On observe pour le calcul que $\overrightarrow{V} \wedge \overrightarrow{B} = P_B(\overrightarrow{V}) \wedge \overrightarrow{B}$. On en déduit alors facilement

$$A^{2p}(\overrightarrow{V}) = (-1)^p \Big(\frac{q}{m}\Big)^{2p} B^{2p} P_B(\overrightarrow{V}), \quad p \geq 1$$
$$A^{2p+1}(\overrightarrow{V}) = (-1)^p \Big(\frac{q}{m}\Big)^{2p+1} B^{2p} \overrightarrow{V} \wedge \overrightarrow{B},$$

cette dernière relation étant encore valable pour $p = 0$. En notant $\omega = \frac{q}{m} B$, on en déduit

$$\begin{aligned} e^{tA}(\overrightarrow{V}) &= \overrightarrow{V} + \sum_{p=1}^{+\infty} (-1)^p \frac{\omega^{2p} t^{2p}}{(2p)!}\ P_B(\overrightarrow{V}) + \sum_{p=0}^{+\infty} (-1)^p \frac{\omega^{2p+1} t^{2p+1}}{(2p+1)!}\ \frac{1}{B}\ \overrightarrow{V} \wedge \overrightarrow{B} \\ &= \overrightarrow{V} - P_B(\overrightarrow{V}) + \cos\,\omega t\, P_B(\overrightarrow{V}) + \sin\,\omega t\ \frac{1}{B}\ \overrightarrow{V} \wedge \overrightarrow{B}. \end{aligned}$$

En l'absence de champ électrique, les équations du mouvement sont données par

$$\overrightarrow{V} = Q_B(\overrightarrow{V_0}) + \cos\,\omega t\, P_B(\overrightarrow{V}_0) + \sin\,\omega t\, \frac{1}{B}\ \overrightarrow{V_0} \wedge \overrightarrow{B}$$
$$\overrightarrow{M_0 M} = t Q_B(\overrightarrow{V_0}) + \frac{1}{w}\ \sin\,\omega t\, P_B(\overrightarrow{V_0}) + \frac{1 - \cos\,\omega t}{\omega B}\ \overrightarrow{V}_0 \wedge \overrightarrow{B}$$

où M_0, $\overrightarrow{V_0}$ désignent la position et la vitesse en $t = 0$ et Q_B la projection orthogonale sur la droite $\mathbb{R} \cdot \overrightarrow{B}$. Il est facile de voir qu'il s'agit d'un mouvement hélicoïdal uniforme de pulsation ω, tracé sur un cylindre d'axe parallèle à $\overrightarrow{B}$ et de rayon $R = \|P_B(\overrightarrow{V_0})\|/\omega$.

En présence d'un champ électrique $\overrightarrow{E}$, le calcul est aisé si $\overrightarrow{E}$ est parallèle à $\overrightarrow{B}$. On a donc dans ce cas une solution particulière évidente $\overrightarrow{V} = t\ \frac{q}{m}\ \overrightarrow{E}$, d'où les lois générales des vitesses et du mouvement :

$$\overrightarrow{V} = Q_B(\overrightarrow{V_0}) + t\ \frac{q}{m}\ \overrightarrow{E} + \cos\,\omega t\, P_B(\overrightarrow{V_0}) + \sin\,\omega t\, \frac{1}{B}\, \overrightarrow{V}_0 \wedge \overrightarrow{B},$$
$$\overrightarrow{M_0 M} = \Big(t Q_B(\overrightarrow{V_0}) + t^2\, \frac{q}{2m}\ \overrightarrow{E}\Big) + \frac{1}{\omega}\ \sin\,\omega t\, P_B(\overrightarrow{V_0}) + \frac{1 - \cos\,\omega t}{\omega B}\ \overrightarrow{V_0} \wedge \overrightarrow{B}.$$

Le mouvement est encore tracé sur un cylindre à base circulaire et sa pulsation est constante, mais le mouvement est accéléré dans la direction de l'axe.

Dans le cas général, on décompose $\overrightarrow{E} = \overrightarrow{E}_{//} + \overrightarrow{E}_{\perp}$ en ses composantes parallèles et orthogonales à $\overrightarrow{B}$, et on observe qu'il existe un vecteur $\overrightarrow{U}$ orthogonal à $\overrightarrow{B}$ et $\overrightarrow{E}_{\perp}$, tel que $\overrightarrow{U} \wedge \overrightarrow{B} = \overrightarrow{E}_{\perp}$. L'équation différentielle devient

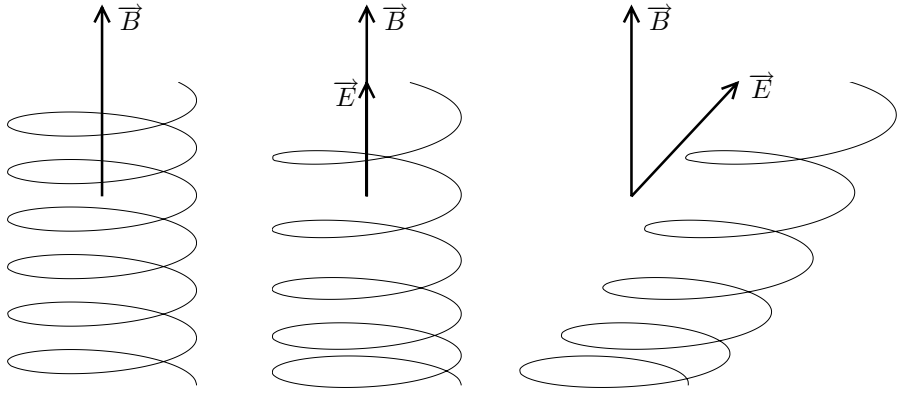
$$\frac{d\overrightarrow{V}}{dt} = \frac{q}{m}\ (\overrightarrow{V} + \overrightarrow{V}) \wedge \overrightarrow{B} + \frac{q}{m}\ \overrightarrow{E}_{//}$$

de sorte que $\overrightarrow{V} + \overrightarrow{U}$ satisfait l'équation différentielle correspondant à un champ électrique parallèle à $\overrightarrow{B}$. En substituant $\overrightarrow{V} + \overrightarrow{U}$ à $\overrightarrow{V}$ et $\overrightarrow{V_0} + \overrightarrow{U}$ à $\overrightarrow{V_0}$ dans les

formules, on obtient

$$\begin{aligned}\overrightarrow{V}+\overrightarrow{U} &= Q_B(\overrightarrow{V}_0)+t\,\frac{q}{m}\,\overrightarrow{E}_{//}+\cos\omega t\,\big(P_B(\overrightarrow{V}_0)+\overrightarrow{U}\big)\\ &\quad+\sin\omega t\,\frac{1}{B}\,(\overrightarrow{V}_0\wedge\overrightarrow{B}+\overrightarrow{E}_{\perp}),\\ \overrightarrow{M_0M} &= t\big(Q_B(\overrightarrow{V}_0)-\overrightarrow{U}\big)+t^2\,\frac{q}{2m}\,\overrightarrow{E}_{//}+\frac{\sin\omega t}{\omega}\,\big(P_B(\overrightarrow{V}_0)+\overrightarrow{U}\big)\\ &\quad+\frac{1-\cos\omega t}{\omega B}\,(\overrightarrow{V}_0\wedge\overrightarrow{B}+\overrightarrow{E}_{\perp}).\end{aligned}$$

Il s'agit encore d'un mouvement de type hélicoïdal accéléré, mais cette fois le mouvement n'est plus tracé sur un cylindre.



3. Équations différentielles linéaires d'ordre p à coefficients constants

On considère ici une équation différentielle sans second membre

$$a_p y^{(p)}+\ldots+a_1y'+a_0y=0 \qquad \text{(E)}$$

où $y:\mathbb{R}\to\mathbb{K}$, $t\mapsto y(t)$ est la fonction inconnue, et où les $a_j\in\mathbb{K}$ sont des constantes, $a_p\neq 0$.

D'après le paragraphe V 4.2, on sait que l'équation (E) est équivalente à un système différentiel (S) d'ordre 1 dans $\mathbb{K}^p$, qui est le système linéaire sans second membre $Y'=AY$ avec

$$A=\begin{pmatrix}0&1&0&\ldots&0\\0&0&1&\ldots&0\\\vdots&&&&\\0&0&0&\ldots&1\\c_0&c_1&c_2&\ldots&c_{p-1}\end{pmatrix},\quad c_j=-\frac{a_j}{a_p}.$$

Grâce au § 1.1, on peut donc énoncer :

Théorème – *L'ensemble* $\mathcal{S}$ *des solutions globales de* (E) *est un* $\mathbb{K}$*-espace vectoriel de dimension* p.

Plaçons-nous maintenant sur le corps $\mathbb{C}$ (si $\mathbb{K} = \mathbb{R}$, les solutions réelles s'obtiennent simplement en prenant la partie réelle et la partie imaginaire des solutions complexes). Cherchons les solutions exponentielles de la forme

$$y(t) = e^{\lambda t}, \quad \lambda \in \mathbb{C}.$$

Comme $y^{(j)}(t) = \lambda^j e^{\lambda t}$, on voit que y est solution de (E) si et seulement si λ est racine du *polynôme caractéristique*

$$P(\lambda) = a_p \lambda^p + \ldots + a_1 \lambda + a_0.$$

3.1. Cas où P a toutes ses racines simples

Si P possède p racines distinctes $\lambda_1, \ldots, \lambda_p$, on obtient p solutions distinctes

$$t \mapsto e^{\lambda_j t}, \quad 1 \leq j \leq p.$$

On verra plus loin que ces solutions sont linéairement indépendantes sur $\mathbb{C}$. L'ensemble des solutions est donc l'espace vectoriel de dimension p des fonctions

$$y(t) = \alpha_1 e^{\lambda_1 t} + \ldots + \alpha_p e^{\lambda_p t}, \quad \alpha_j \in \mathbb{C}.$$

3.2. Cas où P a des racines multiples

On peut alors écrire

$$P(\lambda) = a_p \prod_{j=1}^{s} (\lambda - \lambda_j)^{m_j}$$

où m_j est la multiplicité de la racine λ_j, avec

$$m_1 + \ldots + m_s = p.$$

Considérons l'opérateur différentiel

$$P\Big(\frac{d}{dt}\Big) = \sum_{i=0}^{p} a_i \, \frac{d^i}{dt^i}.$$

On voit que l'équation différentielle étudiée peut se récrire

$$P\Big(\frac{d}{dt}\Big) y = 0 \tag{E}$$

et on a d'autre part la formule

$$P\Big(\frac{d}{dt}\Big) e^{\lambda t} = P(\lambda) e^{\lambda t}, \qquad \forall \lambda \in \mathbb{C}.$$

Comme les dérivées partielles $\frac{d}{dt}$ et $\frac{d}{d\lambda}$ commutent d'après le théorème de Schwarz, on obtient

$$P\Big(\frac{d}{dt}\Big)(t^q e^{\lambda t}) = P\Big(\frac{d}{dt}\Big)\Big(\frac{d^q}{d\lambda^q}\ e^{\lambda t}\Big) = \frac{d^q}{d\lambda^q}\Big(P\Big(\frac{d}{dt}\Big)e^{\lambda t}\Big) = \frac{d^q}{d\lambda^q}\Big(P(\lambda)e^{\lambda t}\Big),$$

d'où, grâce à la formule de Leibnitz :

$$P\Big(\frac{d}{dt}\Big)(t^q e^{\lambda t}) = \sum_{i=0}^{q} C_q^i P^{(i)}(\lambda) e^{\lambda t}.$$

Comme λ_j est racine de multiplicité m_j, on a $P^{(i)}(\lambda_j) = 0$ pour $0 \le i \le m_j - 1$, et $P^{(m_j)}(\lambda_j) \neq 0$. On en déduit

$$P\Big(\frac{d}{dt}\Big)\Big(t^q e^{\lambda_j t}\Big) = 0, \qquad 0 \le q \le m_j - 1.$$

L'équation (E) admet donc les solutions

$$y(t) = t^q e^{\lambda_j t}, \qquad 0 \le q \le m_{j-1}, \qquad 1 \le j \le s$$

soit au total $m_1 + \ldots + m_s = p$ solutions.

Lemme – *Si $\lambda_1, \ldots, \lambda_s \in \mathbb{C}$ sont des nombres complexes deux à deux distincts, alors les fonctions*

$$y_{j,q}(t) = t^q e^{\lambda_j t}, \quad 1 \le j \le s, \quad q \in \mathbb{N}$$

sont linéairement indépendantes.

Démonstration. Considérons une combinaison linéaire finie

$$\sum \alpha_{j,q} y_{j,q} = 0, \quad \alpha_{j,q} \in \mathbb{C}.$$

Si les coefficients sont non tous nuls, soit N le maximum des entiers q tels qu'il existe j avec $\alpha_{j,q} \neq 0$. Supposons par exemple $\alpha_{1,N} \neq 0$. On pose alors

$$Q(\lambda) = (\lambda - \lambda_1)^N (\lambda - \lambda_2)^{N+1} \ldots (\lambda - \lambda_s)^{N+1}$$

Il vient $Q^{(i)}(\lambda_j) = 0$ pour $j \ge 2$ et $0 \le i \le N$, tandis que $Q^{(i)}(\lambda_1) = 0$ pour $0 \le i < N$ et $Q^{(N)}(\lambda_1) \neq 0$. On en déduit

$$\begin{aligned} Q\Big(\frac{d}{dt}\Big)(t^q e^{\lambda_j t}) &= \sum_{i=0}^{q} C_q^i Q^{(i)}(\lambda_j) t^{q-i} e^{\lambda_j t} \\ &= 0 \quad \text{pour} \quad 0 \le q \le N, \quad 1 \le j \le s, \end{aligned}$$

sauf si $q = N$, $j = 1$, auquel cas

$$Q\Big(\frac{d}{dt}\Big)(t^N e^{\lambda_1 t}) = Q^{(N)}(\lambda_1) e^{\lambda_1 t}.$$

En appliquant l'opérateur $Q\left(\frac{d}{dt}\right)$ à la relation $\sum \alpha_{j,q} t^q e^{\lambda_j t} = 0$ on obtient alors $\alpha_{1,N} Q^{(N)}(\lambda_1) e^{\lambda_1 t} = 0$, ce qui est absurde puisque $\alpha_{1,N} \neq 0$ et $Q^{(N)}(\lambda_1) \neq 0$. Le lemme est démontré. On peut donc énoncer :

Théorème – *Lorsque le polynôme caractéristique $P(\lambda)$ a des racines complexes $\lambda_1, \ldots, \lambda_s$ de multiplicités respectives $m_1, \ldots, m_s$, l'ensemble $\mathcal{S}$ des solutions est le $\mathbb{C}$-espace vectoriel de dimension p ayant pour base les fonctions*

$$t \mapsto t^q e^{\lambda_j t}, \quad 1 \leq j \leq s, \quad 0 \leq q \leq m_j - 1.$$

3.3. ÉQUATIONS LINÉAIRES D'ORDRE p AVEC SECOND MEMBRE

Soit à résoudre l'équation différentielle

$$\text{(E)} \qquad a_p y^{(p)} + \ldots + a_1 y' + a_0 y = b(t),$$

où $b : I \to \mathbb{C}$ est une fonction continue donnée. On commence par résoudre l'équation sans second membre

$$\text{(E}_0\text{)} \qquad a_p y^{(p)} + \ldots + a_1 y' + a_0 y = 0.$$

Soit $(v_1, \ldots, v_p)$ une base des solutions de (E$_0$). On cherche alors une solution particulière de (E).

Dans un certain nombre de cas, une solution simple peut être trouvée rapidement. Par exemple, si b est un polynôme de degré d et si $a_0 \neq 0$, l'équation (E) admet une solution polynomiale y de degré d, que l'on peut rechercher par identification des coefficients. Si $b(t) = \alpha e^{\lambda t}$ et si λ n'est pas racine du polynôme caractéristique, l'équation admet pour solution $(\alpha / P(\lambda)) e^{\lambda t}$. Si b est une fonction exponentielle-polynôme, (E) admet une solution du même type (noter que les fonctions trigonométriques se ramènent à ce cas).

En général, le principe consiste à appliquer la *méthode de variation des constantes au système différentiel* (S) *d'ordre 1 associé à* (E).

Si on pose $y = y_0$, $y' = y_1, \ldots, y^{(p-1)} = y_{p-1}$, l'équation (E) équivaut au système

$$\text{(S)} \quad \begin{cases} y_0' = y_1 \\ \vdots \\ y_{p-2}' = y_{p-1} \\ y_{p-1}' = -\frac{1}{a_p} \left(a_0 y_0 + a_1 y_1 + \ldots + a_{p-1} y_{p-1}\right) + \frac{1}{a_p} b(t). \end{cases}$$

Ce système linéaire peut se récrire (S): $Y' = AY + B(t)$ avec

$$Y = \begin{pmatrix} y_0 \\ \vdots \\ y_{p-1} \end{pmatrix} \quad \text{et} \quad B(t) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{1}{a_p} b(t). \end{pmatrix}$$

Le système homogène (S_0) $Y' = AY$ admet pour base de solutions les fonctions

$$V_1 = \begin{pmatrix} v_1 \\ v_1' \\ \vdots \\ v_1^{(p-1)} \end{pmatrix} \qquad V_2 = \begin{pmatrix} v_2 \\ v_2' \\ \vdots \\ v_2^{(p-1)} \end{pmatrix} \quad \dots \quad V_p = \begin{pmatrix} v_p \\ v_p' \\ \vdots \\ v_p^{(p-1)} \end{pmatrix}.$$

On cherche alors une solution particulière de (S) sous la forme

$$Y(t) = \alpha_1(t)V_1(t) + \ldots + \alpha_p(t)V_p(t).$$

Comme $V_j' = AV_j$, il vient

$$\begin{aligned} Y'(t) &= \sum \alpha_j(t)V_j'(t) + \sum \alpha_j'(t)V_j(t) \\ &= AY(t) + \sum \alpha_j'(t)V_j(t). \end{aligned}$$

Il suffit donc de choisir les α_j tels que $\sum \alpha_j'(t)V_j(t) = B(t)$, c'est-à-dire

$$\begin{cases} \alpha_1'(t)v_1(t) + \ldots + \alpha_p'(t)v_p(t) = 0 \\ \ldots \\ \alpha_1'(t)v_1^{(p-2)}(t) + \ldots + \alpha_p'(t)v_p^{(p-2)}(t) = 0 \\ \alpha_1'(t)v_1^{(p-1)}(t) + \ldots + \alpha_p'(t)v_p^{(p-1)}(t) = \frac{1}{a_p}\, b(t). \end{cases}$$

On obtient ainsi un système linéaire de p équations par rapport aux p inconnues $\alpha_1'(t), \ldots, \alpha_p'(t)$. Le déterminant de ce système est non nul pour tout $t \in \mathbb{R}$ (les vecteurs $V_1(t), \ldots, V_p(t)$ sont linéairement indépendants, car si une combinaison linéaire $Y = \beta_1 V_1 + \ldots + \beta_p V_p$ est telle que $Y(t) = 0$, alors $Y \equiv 0$ d'après le théorème d'unicité, donc $\beta_1 = \ldots = \beta_p = 0$).

La résolution de ce système permet de calculer $\alpha_1', \ldots, \alpha_p'$, puis $\alpha_1, \ldots, \alpha_p$ par intégration, d'où la solution particulière cherchée :

$$y(t) = \alpha_1(t)v_1(t) + \ldots + \alpha_p(t)v_p(t).$$

Exemple – (E) $y'' + 4y = \tan t$, avec $t \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$.

• On commence par résoudre l'équation sans second membre

$$(E_0) \qquad y'' + 4y = 0.$$

Le polynôme caractéristique est $P(\lambda) = \lambda^2 + 4$, et possède deux racines simples $2i$ et $-2i$. L'équation (E_0) admet pour base de solutions les fonctions $t \mapsto e^{2it}$, $t \mapsto e^{-2it}$, ou encore :

$$t \mapsto \cos 2t, \quad t \mapsto \sin 2t.$$

• On cherche ensuite une solution particulière de (E) en posant

$$y(t) = \alpha_1(t) \cos 2t + \alpha_2(t) \sin 2t.$$

Ceci conduit à résoudre le système

$$\begin{cases} \alpha_1'(t)\cos 2t + \alpha_2'(t)\sin 2t = 0 \\ \alpha_1'(t)\cdot(-2\sin 2t) + \alpha_2'(t)\cdot(2\cos 2t) = \tan t. \end{cases}$$

Le déterminant du système étant égal à 2, on obtient

$$\begin{cases} \alpha_1'(t) = -\dfrac{1}{2}\tan t \sin 2t = -\sin^2 t = -\dfrac{1}{2}(1-\cos 2t) \\ \alpha_2'(t) = \dfrac{1}{2}\tan t \cos 2t = \dfrac{1}{2}\tan t(2\cos^2 t - 1) = \sin 2t - \dfrac{1}{2}\tan t, \end{cases}$$

$$\begin{cases} \alpha_1(t) = -\dfrac{t}{2} + \dfrac{1}{4}\sin 2t \\ \alpha_2(t) = -\dfrac{1}{4}\cos 2t + \dfrac{1}{2}\ln(\cos t), \end{cases}$$

d'où la solution particulière

$$y(t) = -\frac{t}{2}\cos 2t + \frac{1}{2}\sin 2t \ln(\cos t).$$

La solution générale est donc

$$y(t) = -\frac{t}{2}\cos 2t + \frac{1}{2}\sin 2t \ln(\cos t) + \alpha_1 \cos 2t + \alpha_2 \sin 2t.$$

4. *Systèmes différentiels linéaires à coefficients variables*

L'objet de ce paragraphe (avant tout théorique) est de généraliser les résultats du § 2 au cas des systèmes linéaires à coefficients variables.

4.1. *Résolvante d'un système linéaire*

Considérons une équation linéaire sans second membre

$$(\mathrm{E}_0) \qquad Y' = A(t)Y$$

où $A : \mathbb{R} \supset I \to M_m(\mathbb{K})$ est une matrice $m \times m$ sur $\mathbb{K}$ à coefficients continus. Soit $\mathcal{S}$ l'ensemble des solutions maximales de (E_0). Pour tout $t_0 \in I$, on sait que

$$\Phi_{t_0} : \mathcal{S} \longrightarrow \mathbb{K}^m, \qquad Y \longmapsto Y(t_0)$$

est un isomorphisme $\mathbb{K}$-linéaire. Pour tout couple $(t, t_0) \in I^2$, on définit

$$\begin{aligned} R(t,t_0) = \Phi_t \circ \Phi_{t_0}^{-1} : \mathbb{K}^m &\xrightarrow{\Phi_{t_0}^{-1}} \mathcal{S} \xrightarrow{\Phi_t} \mathbb{K}^m \\ V &\longmapsto Y \longmapsto Y(t). \end{aligned}$$

On a donc $R(t,t_0)\cdot V = Y(t)$, où Y est la solution telle que $Y(t_0)=V$. Comme $R(t,t_0)$ est un isomorphisme $\mathbb{K}^m \to \mathbb{K}^m$, il sera identifié à la matrice inversible qui lui correspond canoniquement dans $M_m(\mathbb{K})$.

Définition – *$R(t,t_0)$ s'appelle la résolvante du système linéaire* (E_0).

Pour tout vecteur $V\in\mathbb{K}^m$, on a avec les notations ci-dessus

$$\left(\frac{d}{dt}\,R(t,t_0)\right)\cdot V = \frac{d}{dt}\,\Big(R(t,t_0)\cdot V\Big)$$
$$= \frac{dY}{dt} = A(t)Y(t) = A(t)R(t,t_0)\cdot V.$$

On en déduit donc $\frac{d}{dt}\,R(t,t_0) = A(t)R(t,t_0)$.

Propriétés de la résolvante

(i) $\forall t\in I,\qquad R(t,t)=I_m \qquad$ *(matrice unité $m\times m$).*

(ii) $\forall (t_0,t_1,t_2)\in I^3,\qquad R(t_2,t_1)R(t_1,t_0)=R(t_2,t_0).$

(iii) *$R(t,t_0)$ est la solution dans $M_m(\mathbb{K})$ du système différentiel*

$$\frac{dM}{dt} = A(t)M(t)$$

où $M(t)\in M_m(\mathbb{K})$ vérifie la condition initiale $M(t_0)=I_m$.

(i) et (ii) sont immédiats à partir de la définition de $R(t,t_0)$ et (iii) résulte de ce qui précède. Retenons enfin que la solution du problème de Cauchy

$$Y' = A(t)Y \quad \text{avec} \quad Y(t_0)=V_0$$

est donnée par

$$Y(t) = R(t,t_0)\cdot V_0.$$

Remarque – Le système $dM/dt = A(t)M(t)$ peut paraître plus compliqué que le système initial puisqu'on a m^2 équations scalaires au lieu de m (on passe de $\mathbb{K}^m$ à $M_m(\mathbb{K})$). Il est néanmoins parfois utile de considérer ce système plutôt que l'équation initiale, parce que tous les objets sont dans $M_m(\mathbb{K})$ et qu'on peut exploiter la structure d'algèbre de $M_m(\mathbb{K})$.

Exemple – Supposons que

$$A(t)A(u) = A(u)A(t) \quad \text{pour tous } t,u\in I. \qquad (*)$$

Alors

$$R(t,t_0) = \exp\left(\int_{t_0}^{t} A(u)du\right).$$

Pour le voir, il suffit de montrer que $M(t) = \exp\left(\int_{t_0}^{t} A(u)du\right)$ satisfait la condition (iii) ci-dessus. Il est clair que $M(t_0) = I_m$. Par ailleurs l'hypothèse de commutation $(*)$ entraîne que $\int_a^b A(u)du$ et $\int_c^d A(u)du$ commutent pour tous $a, b, c, d \in I$, le produit étant égal dans les deux cas à

$$\iint_{[a,b]\times[c,d]} A(u)A(v)dudv$$

par le théorème de Fubini. On a donc

$$\begin{aligned} M(t+h) &= \exp\left(\int_{t_0}^{t} A(u)du + \int_{t}^{t+h} A(u)du\right) \\ &= \exp\left(\int_{t}^{t+h} A(u)du\right) M(t). \end{aligned}$$

Or $\int_t^{t+h} A(u)du = hA(t) + o(h)$, donc utilisant le développement en série de l'exponentielle on trouve

$$\begin{aligned} M(t+h) &= (I_m + hA(t) + o(h))M(t) \\ &= M(t) + hA(t)M(t) + o(h), \end{aligned}$$

ce qui montre bien que $dM/dt = A(t)M(t)$.

En particulier, si U et V sont des matrices constantes qui commutent et si $A(t) = f(t)U + g(t)V$ pour des fonctions scalaires f, g, alors l'hypothèse $(*)$ est satisfaite. On a donc

$$\begin{aligned} R(t, t_0) &= \exp\left(\int_{t_0}^{t} f(u)du \cdot U + \int_{t_0}^{t} g(u)du \cdot V\right) \\ &= \exp\left(\int_{t_0}^{t} f(u)du \cdot U\right) \exp\left(\int_{t_0}^{t} g(u)du \cdot V\right). \end{aligned}$$

Exercice 1 – *Utiliser la dernière remarque de l'exemple pour calculer la résolvante associée aux matrices*

$$A(t) = \begin{pmatrix} a(t) & -b(t) \\ b(t) & a(t) \end{pmatrix}, \quad \text{resp.} \quad A(t) = \begin{pmatrix} 1 & 0 & \cos^2 t \\ 0 & 1 & \cos^2 t \\ 0 & 0 & \sin^2 t \end{pmatrix}.$$

Exercice 2 – *Résoudre le système linéaire*

$$\begin{cases} \frac{dx}{dt} = \frac{1}{t}\, x + ty \\ \frac{dy}{dt} = y \end{cases} \qquad \text{où} \quad A(t) = \begin{pmatrix} 1/t & t \\ 0 & 1 \end{pmatrix}$$

et en déduire la formule donnant la résolvante $R(t, t_0)$.
Montrer que dans ce cas on a

$$R(t, t_0) \neq \exp\left(\int_{t_0}^{t} A(u)du\right).$$

L'exercice 2 montre que c'est le plus souvent la résolution du système qui permet de déterminer la résolvante, et non pas l'inverse comme pourrait le laisser croire la terminologie.

4.2. Wronskien d'un système de solutions

On va voir ici qu'on sait toujours calculer le déterminant d'un système de solutions, ou ce qui revient au même, le déterminant de la résolvante, même lorsque la résolvante n'est pas connue.

Définition – *Le Wronskien d'un système de m solutions $Y_1, Y_2, \ldots, Y_m$ de* (E_0) *est*

$$W(t) = \det\left(Y_1(t), \ldots, Y_m(t)\right)$$

Posons $V_j = Y_j(t_0)$. Alors $Y_j(t) = R(t, t_0) \cdot V_j$, d'où

$$W(t) = \det\left(R(t, t_0)\right) \cdot \det\left(V_1, \ldots, V_m\right).$$

On est donc ramené à calculer la quantité

$$\Delta(t) = \det\left(R(t, t_0)\right),$$

et pour cela on va montrer que $\Delta(t)$ vérifie une équation différentielle simple. On a

$$\begin{aligned}\Delta(t+h) &= \det\left(R(t+h, t_0)\right) = \det\left(R(t+h, t)R(t, t_0)\right)\\ &= \det\left(R(t+h, t)\right)\Delta(t).\end{aligned}$$

Comme $R(t,t) = I_m$ et $\frac{d}{du}\, R(u,t)|_{u=t} = A(t)R(t,t) = A(t)$, la formule de Taylor donne

$$\begin{aligned}R(t+h, t) &= I_m + hA(t) + o(h),\\ \det\left(R(t+h, t)\right) &= \det\left(I_m + hA(t)\right) + o(h).\end{aligned}$$

Lemme – *Si $A = (a_{ij}) \in M_m(\mathbb{K})$, alors*

$$\det\left(I_m + hA\right) = 1 + \alpha_1 h + \ldots + \alpha_m h^m$$

avec $\alpha_1 = \operatorname{tr} A = \displaystyle\sum_{1 \le i \le m} a_{ii}$.

En effet dans $\det\left(I_m + hA\right)$ le terme diagonal est

$$(1 + ha_{11}) \ldots (1 + ha_{mm}) = 1 + h \sum a_{ii} + h^2 \ldots$$

et les termes non diagonaux sont multiples de h^2. ∎

Le lemme entraîne alors

$$\det\left(R(t+h,t)\right) = 1 + h\ \mathrm{tr}\left(A(t)\right) + o(h),$$
$$\Delta(t+h) = \Delta(t) + h\ \mathrm{tr}\left(A(t)\right)\Delta(t) + o(h).$$

On en déduit

$$\Delta'(t) = \mathrm{tr}\left(A(t)\right)\Delta(t),$$

et comme $\Delta(t_0) = \det\left(R(t_0,t_0)\right) = \det\ I_m = 1$, il vient :

$$\det\ R(t,t_0) = \Delta(t) = \exp\left(\int_{t_0}^{t} \mathrm{tr}\ A(u)du\right),$$
$$W(t) = \exp\left(\int_{t_0}^{t} \mathrm{tr}\ A(u)du\right)\det\left(V_1,\ldots,V_m\right).$$

4.3. Méthode de variation des constantes

Soit à résoudre le système différentiel linéaire

$$\text{(E)} \qquad Y' = A(t)Y + B(t),$$

et soit $R(t,t_0)$ la résolvante du système linéaire sans second membre

$$\text{(E}_0\text{)} \qquad Y' = A(t)Y.$$

On cherche alors une solution particulière de (E) sous la forme

$$Y(t) = R(t,t_0)\cdot V(t)$$

où V est supposée différentiable. Il vient

$$\begin{aligned}\frac{dY}{dt} &= \left(\frac{d}{dt}\ R(t,t_0)\right)\cdot V(t) + R(t,t_0)\cdot V'(t)\\ &= A(t)R(t,t_0)\cdot V(t) + R(t,t_0)\cdot V'(t)\\ &= A(t)Y(t) + R(t,t_0)\cdot V'(t).\end{aligned}$$

Il suffit donc de prendre $R(t,t_0)\cdot V'(t) = B(t)$, c'est-à-dire

$$V'(t) = R(t_0,t)\cdot B(t),$$
$$V(t) = \int_{t_0}^{t} R(t_0,u)\cdot B(u)du,$$
$$Y(t) = R(t,t_0)\cdot V(t) = \int_{t_0}^{t} R(t,t_0)R(t_0,u)\cdot B(u)du,$$
$$Y(t) = \int_{t_0}^{t} R(t,u)B(u)du.$$

On obtient ainsi la solution particulière telle que $Y(t_0) = 0$. La solution telle que $Y(t_0) = V_0$ est donnée par

$$Y(t) = R(t, t_0) \cdot V_0 + \int_{t_0}^{t} R(t, u)B(u)du.$$

Dans le cas où $A(t) = A$ est à coefficients constants on retrouve la formule du § 2.4, dans laquelle $R(t, t_0) = e^{(t-t_0)A}$, et la formule du Wronskien équivaut à l'identité déjà connue

$$\det\,(e^{(t-t_0)A}) = \exp\,((t - t_0)\,\mathrm{tr}\ A).$$

5. Problèmes

5.1. Soient b et c deux fonctions continues sur un intervalle fixé $T = [0, \tau[$. Soit (S) le système différentiel linéaire à coefficients constants et avec second membre

$$\begin{cases} x' = \quad y + b(t) \\ y' = 2x - y + c(t) \end{cases}$$

et soit (S_0) le système sans second membre associé (pour lequel $b(t) = c(t) = 0$).

(a) Écrire la matrice A de (S_0), et calculer e^{tA}.

(b) Déterminer la solution générale du système (S_0).

(c) Déterminer la solution générale du système (S) pour $b(t) = 0$, $c(t) = e^{-t}$.

5.2. Soit t une variable réelle ≥ 0. On considère le système différentiel linéaire

$$(S) \quad \begin{cases} x' = \quad 2y \\ y' = x - y \end{cases}$$

(a) Écrire la matrice A de (S), montrer qu'elle a deux valeurs propres réelles λ et μ ($\lambda > \mu$) et déterminer les sous-espaces propres correspondants.

(b) On pose $e_x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $e_y = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, et on note respectivement $v_\lambda = \begin{pmatrix} x_\lambda \\ y_\lambda \end{pmatrix}$ et $v_\mu = \begin{pmatrix} x_\mu \\ y_\mu \end{pmatrix}$ les vecteurs propres associés à λ et μ tels que $y_\lambda = y_\mu = 1$.
Calculer x_λ, x_μ. Déterminer la matrice de passage P de l'ancienne base (e_x, e_y) à la nouvelle base (v_λ, v_μ), et calculer sa matrice inverse P^{-1}.

(c) On pose $e^{tA} = \begin{pmatrix} a(t) & b(t) \\ c(t) & d(t) \end{pmatrix}$. Calculer explicitement $a(t)$, $b(t)$, $c(t)$, $d(t)$.
Donner la solution du système (S) vérifiant les conditions initiales $x(0) = x_0$, $y(0) = y_0$.

(d) Soit $T(x_0, y_0)$ la trajectoire $t \mapsto \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = M(t)$ correspondant aux conditions initiales $M(0) = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$.

(α) Pour quelles positions de $M(0)$ cette trajectoire $T(x_0, y_0)$ est-elle une demi-droite ?

(β) Pour quelles positions de $M(0)$ tend-elle vers 0 quand $t \to +\infty$?

(γ) Indiquer sur un même figure :

- la forme des trajectoires $T(x_0, 0)$ partant d'un point $(x_0, 0)$, $x_0 > 0$, de l'axe des x ;
- la forme des trajectoires $T(0, y_0)$ partant d'un point $(0, y_0)$, $y_0 > 0$, de l'axe des y.

5.3. On note t une variable réelle, et on considère les deux matrices

$$B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

(a) Pour tout $n \geq 0$, calculer explicitement B^n et C^n, et en déduire e^{tB} et e^{tC}.

(b) Mêmes questions pour la matrice

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

On pose maintenant $T = [0, +\infty[$; on note $b_i(t)$ $(1 \leq i \leq 4)$ quatre fonctions continues sur T, et on considère le système différentielle linéaire avec second membre

$$\text{(S)} \quad \begin{cases} y_1' = & y_2 & + b_1(t) \\ y_2' = y_1 & & + b_2(t) \\ y_3' = & & y_3 + y_4 + b_3(t) \\ y_4' = & & y_4 + b_4(t) \end{cases}.$$

On note (S_0) le système sans second membre associé à (S).

(c) Écrire la solution de (S_0) correspondant à des conditions initiales $y_i(0) = v_i$, les v_i $(1 \leq i \leq 4)$ étant quatre constantes données.

(d) Indiquer comment on peut alors résoudre (S) par la méthode dite de variation des constantes, et appliquer cette méthode au cas particulier

$$b_1(t) = 1, \quad b_2(t) = b_3(t) = 0, \quad b_4 = e^t.$$

5.4. On considère l'équation linéaire du 3$^{\text{e}}$ ordre

$$\text{(E)} \qquad y''' + y'' + y' + y = \cos t,$$

où y désigne une fonction inconnue de la variable $t \geq 0$.

(a) Déterminer la solution générale de l'équation sans second membre associée à (E).

(b) A l'aide de la méthode de variation des constantes, déterminer la solution générale de l'équation (E).

(c) Montrer que (E) admet une solution et une seule de la forme $At\cos t+Bt\sin t$: la déterminer explicitement, et tracer son graphe.

5.5. On considère dans $\mathbb{R}^2$ le système différentiel

$$\begin{cases} \dfrac{dx}{dt} = tx - y \\ \dfrac{dy}{dt} = x + ty \end{cases}$$

où x, y sont des fonctions réelles de la variable réelle t.

(a) Résoudre le problème de Cauchy de donnée initiale (x_0, y_0) au temps $t_0 = 0$ (on pourra poser $z = x + iy$).

(b) Même question pour le système

$$\begin{cases} \dfrac{dx}{dt} = tx - y + t\cos t - t^3 \sin t \\ \dfrac{dy}{dt} = x + ty + t\sin t + t^3 \cos t. \end{cases}$$

5.6. On considère un système différentiel $X' = A(t)X$ où $A(t)$ est une matrice à 2 lignes et 2 colonnes à coefficients de période 2π, bornés et continus par morceaux.

(a) Montrer que l'application qui à $M \in \mathbb{R}^2$ associe la position à l'instant s de la solution $X(t)$ de $X' = A(t)$ vérifiant $X(0) = M$ est une application linéaire bijective. On désignera par U_s cet endormorphisme et on notera $V = U_{2\pi}$.

(b) Montrer que l'équation $X' = A(t)X$ admet une solution 2π-périodique non identiquement nulle si et seulement si 1 est valeur propre de V ; comment peut-on interpréter le fait que V admette pour valeur propre une racine k-ième de l'unité ?

(c) On considère l'équation différentielle $y'' + f(t)y = 0$ où f est une fonction 2π-périodique à valeurs rélles. Mettre cette équation sous la forme d'un système du premier ordre.

(d) On supposera dorénavant que

$$f(t) = \begin{cases} (w+\varepsilon)^2 & \text{si} \quad t \in [0, \pi[\\ (w-\varepsilon)^2 & \text{si} \quad t \in [\pi, 2\pi[\end{cases}$$

où $0 < \varepsilon < w$ sont des constantes. Déterminer U_π ; montrer que V se met sous la forme $B \circ U_\pi$ où l'on déterminera la matrice B (on pourra utiliser que f est constante sur $[\pi, 2\pi[$ ainsi que sur $[0, \pi[$). Vérifier que $\det V = 1$.

(e) Montrer qu'alors une des valeurs propres de V est inférieure à 1 en module et que l'équation $y'' + f(t)y = 0$ admet une solution bornée (non identiquement nulle) sur $[0, +\infty[$ et une solution bornée (non identiquement nulle) sur $]-\infty, 0[$; à quelle condition admet-elle une solution bornée (non identiquement nulle) sur $\mathbb{R}$?

(f) Montrer que la trace de V s'écrit

$$-\Delta \cos 2\pi\varepsilon + (2 + \Delta) \cos 2\pi w$$

où

$$\frac{w+\varepsilon}{w-\varepsilon} + \frac{w-\varepsilon}{w+\varepsilon} = 2(1 + \Delta).$$

En déduire que si w n'est pas la moitié d'un entier et si ε est assez petit, toutes les solutions de $y'' + f(t)y = 0$ sont bornées. Que passe-t-il si w est la moitié d'un entier ?

CHAPITRE VIII

MÉTHODES NUMÉRIQUES À UN PAS

L'objectif de ce chapitre est de décrire un certain nombre de méthodes permettant de résoudre numériquement le problème de Cauchy de condition initiale $y(t_0) = y_0$ pour une équation différentielle

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

où $f : [t_0, t_0+T] \times \mathbb{R} \to \mathbb{R}$ est une fonction suffisamment régulière. Nous avons choisi ici d'exposer le cas des équations unidimensionnelles dans le seul but de simplifier les notations ; le cas des systèmes dans $\mathbb{R}^m$ est tout à fait identique, à condition de considérer y comme une variable vectorielle et f comme une fonction vectorielle dans les algorithmes qui vont être décrits.

Étant donné une subdivision $t_0 < t_1 < \ldots < t_N = t_0 + T$ de $[t_0, t_0 + T]$, on cherche à déterminer des valeurs approchées $y_0, y_1, \ldots, y_N$ des valeurs $y(t_n)$ prises par la solution exacte y. On notera les pas successifs

$$h_n = t_{n+1} - t_n, \qquad 0 \leq n \leq N - 1,$$

et

$$h_{\max} = \max(h_n)$$

le maximum du pas.

On appelle *méthode à un pas* une méthode permettant de calculer y_{n+1} à partir de la seule valeur antérieure y_n. Une méthode à r pas est au contraire une méthode qui utilise les r valeurs antérieures $y_n, \ldots, y_{n-r+1}$ (valeurs qui doivent donc être mémorisées) afin de faire le calcul de y_{n+1}.

1. DÉFINITION DES MÉTHODES À UN PAS, EXEMPLES

1.1. DÉFINITIONS

Les méthodes à un pas sont les méthodes de résolution numérique qui peuvent s'écrire sous la forme

$$y_{n+1} = y_n + h_n \Phi(t_n, y_n, h_n), \qquad 0 \leq n < N,$$

où $\Phi : [t_0, t_0 + T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ est une fonction que l'on supposera continue. Dans la pratique, la fonction $\Phi(t, y, h)$ peut n'être définie que sur une partie de la forme $[t_0, t_0 + T] \times J \times [0, \delta]$ où J est un intervalle de $\mathbb{R}$ (de sorte en particulier que $[t_0, t_0 + T] \times J$ soit contenu dans le domaine de définition de l'équation différentielle).

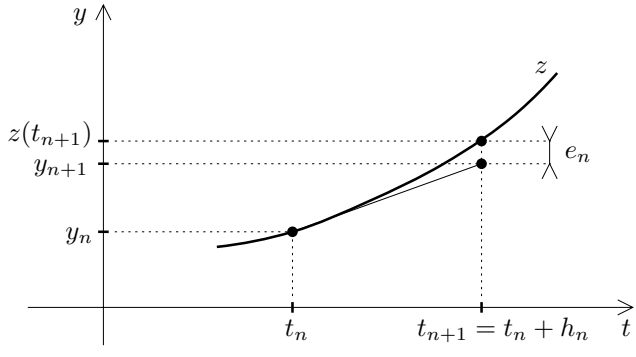
Exemple – La méthode d'Euler est la méthode à un pas associée à la fonction $\Phi(t, y, h) = f(t, y)$, et définie par la formule de récurrence $y_{n+1} = y_n + h_n f(t_n, y_n)$ (voir chapitre V, § 2.3).

Définition – *L'erreur de consistance e_n relative à une solution exacte z est l'erreur*

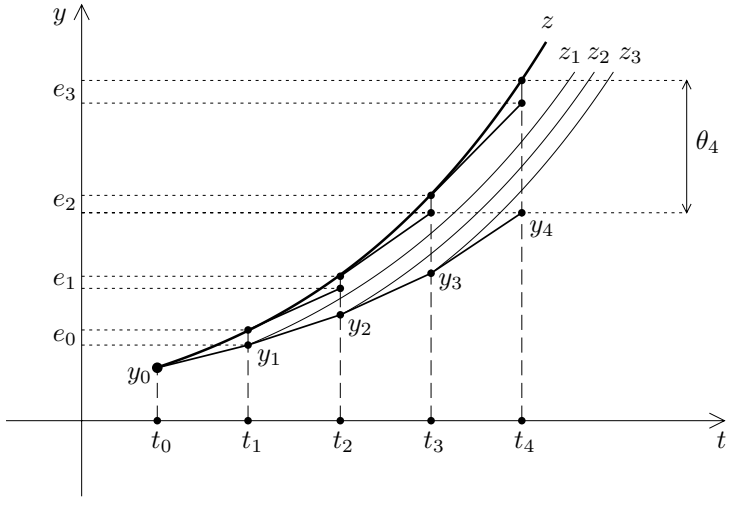
$$e_n = z(t_{n+1}) - y_{n+1}, \quad 0 \leq n < N$$

produite par application de l'algorithme $y_{n+1} = y_n + h_n \Phi(t_n, y_n, h_n)$ à partir de la valeur $y_n = z(t_n)$. Autrement dit, cette erreur mesure l'écart entre la valeur exacte $z(t_{n+1})$ au temps t_{n+1}, et la valeur approchée y_{n+1} issue de la valeur $y_n = z(t_n)$ prise comme valeur initiale au temps t_n (une seule étape de l'algorithme est donc mise en jeu). En termes de la fonction Φ, on a

$$e_n = z(t_{n+1}) - z(t_n) - h_n \Phi(t_n, z(t_n), h_n).$$



Comme le montre le schéma ci-dessous, l'erreur de consistance n'a a priori que peu de rapport avec l'erreur globale $\theta_n = \max\limits_{0 \leq j \leq n} |z(t_j) - y_j|$ résultant d'un calcul de n valeurs successives $y_1, \ldots, y_n$ à partir de la donnée initiale $y_0 = z(t_0)$ (qui est à vrai dire la seule erreur intéressant réellement le numéricien). On imagine cependant, et nous reviendrons là-dessus plus en détail au § 2, que $|\theta_n|$ sera de l'ordre de grandeur de $|e_0| + |e_1| + \cdots + |e_{n-1}|$, sous des hypothèses convenables de régularité pour la fonction f. C'est pourquoi l'évaluation de e_n va gouverner l'évaluation de l'erreur globale.



[Les fonctions z, z_1, z_2, z_3 représentent ici les solutions exactes passant par les points (t_0, y_0) et (t_j, y_j), $j = 1, 2, 3$].

1.2. Retour sur la méthode d'Euler

Soit z une solution exacte de l'équation (E). On a au premier ordre l'approximation

$$z(t_{n+1}) = z(t_n + h_n) \simeq z(t_n) + h_n z'(t_n) = z(t_n) + h_n f(t_n, z(t_n)).$$

Comme on l'a déjà vu au chapitre V, ceci conduit à l'algorithme

$$\begin{cases} y_{n+1} = y_n + h_n f(t_n, y_n) \\ t_{n+1} = t_n + h_n. \end{cases}$$

Par définition de l'erreur de consistance, on a $e_n = z(t_n + h_n) - y_{n+1}$ où

$$y_{n+1} = z(t_n) + h_n f(t_n, z(t_n)) = z(t_n) + h_n z'(t_n).$$

La formule de Taylor-Lagrange donne

$$e_n = z(t_n + h_n) - (z(t_n) + h_n z'(t_n)) = \frac{1}{2}\, h_n^2 z''(t_n) + o(h_n^2),$$

pourvu que z soit de classe C^2. C'est bien le cas si f est de classe C^1, et on sait alors que

$$z''(t) = f^{[1]}(t, z(t)) \qquad \text{où} \qquad f^{[1]} = f'_t + f'_y f.$$

On en déduit par conséquent

$$e_n = \frac{1}{2}\, h_n^2 f^{[1]}(t_n, y_n) + o(h_n^2).$$

Cette erreur en h_n^2 est relativement importante, à moins que le pas h_n ne soit choisi très petit, ce qui augmente considérablement le volume des calculs à effectuer. On va donc essayer de construire des méthodes permettant de réduire l'erreur de consistance e_n.

1.3. Méthode de Taylor d'ordre p

Supposons que f soit de classe C^p. Alors toute solution exacte z est de classe C^{p+1}, et sa dérivée k-ième est $z^{(k)}(t) = f^{[k-1]}(t, z(t))$. La formule de Taylor d'ordre p implique

$$z(t_n + h_n) = z(t_n) + \sum_{k=1}^{p} \frac{1}{k!}\ h_n^k f^{[k-1]}(t_n, z(t_n)) + o(h_n^p).$$

Lorsque h_n est assez petit, l'approximation est d'autant meilleure que p est plus grand. On est donc amené à considérer l'algorithme suivant, appelé méthode de Taylor d'ordre p :

$$\begin{cases} y_{n+1} = y_n + \displaystyle\sum_{k=1}^{p} \frac{1}{k!}\ h_n^k f^{[k-1]}(t_n, y_n) \\ \\ t_{n+1} = t_n + h_n. \end{cases}$$

D'après la définition générale des méthodes à un pas, cet algorithme correspond au choix $\Phi(t, y, h) = \sum_{k=1}^{p} \frac{1}{k!}\ h^{k-1} f^{[k-1]}(t, y)$. Calculons l'erreur de consistance e_n. En supposant $y_n = z(t_n)$, la formule de Taylor d'ordre $p+1$ donne

$$\begin{aligned} e_n = z(t_{n+1}) - y_{n+1} &= z(t_n + h_n) - \sum_{k=0}^{p} \frac{1}{k!}\ h_n^k z^{(k)}(t_n) \\ &= \frac{1}{(p+1)!}\ h_n^{p+1} f^{[p]}(t_n, y_n) + o(h_n^{p+1}). \end{aligned}$$

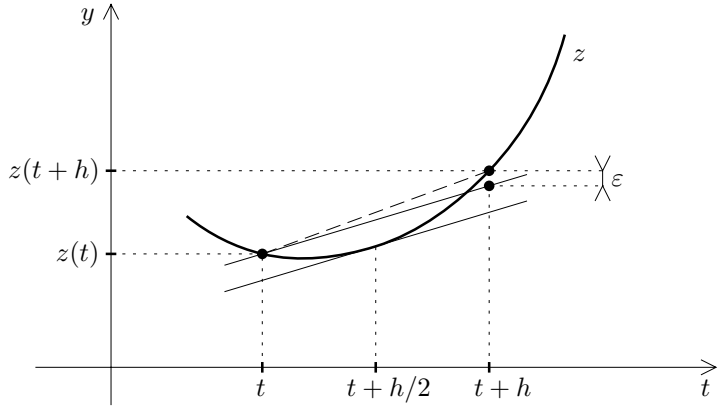
L'erreur est donc maintenant de l'ordre de h_n^{p+1}. On dira d'une manière générale qu'une méthode est *d'ordre p* si l'erreur de consistance est en h_n^{p+1} chaque fois que f est de classe C^p au moins. La méthode d'Euler est le cas particulier $p = 1$ de la méthode de Taylor.

Remarque – Dans la pratique, la méthode de Taylor souffre de deux inconvénients graves qui en font généralement déconseiller l'utilisation pour $p \geq 2$:

- Le calcul des quantités $f^{[k]}$ est souvent complexe et coûteux en temps machine. Il faut aussi pouvoir évaluer explicitement $f^{[k]}$, ce qui n'est pas toujours le cas (par exemple, si f est une donnée expérimentale discrétisée).
- La méthode suppose *a priori* que f soit très régulière ; les erreurs risquent donc de ne pas pouvoir être contrôlées si certaines dérivées de f présentent des discontinuités ou une mauvaise continuité (pentes élevées).

1.4. Méthode du point milieu

Cette méthode est décrite par le schéma suivant.



L'idée est que la corde de la fonction z sur $[t, t+h]$ a une pente voisine de $z'\big(t+\frac{h}{2}\big)$, alors que dans la méthode d'Euler on approxime brutalement cette pente par $z'(t)$. On écrit donc :

$$z(t+h) \simeq z(t) + hz'\Big(t+\frac{h}{2}\Big). \qquad (*)$$

Si z est de classe C^3, il vient

$$\begin{aligned} z(t+h) &= z(t) + hz'(t) + \frac{1}{2}\, h^2 z''(t) + \frac{1}{6}\, h^3 z'''(t) + o(h^3), \\ z'\Big(t+\frac{h}{2}\Big) &= \qquad z'(t) + \frac{1}{2}\, hz''(t) + \frac{1}{8}\, h^2 z'''(t) + o(h^2). \end{aligned}$$

L'erreur commise est donc

$$\varepsilon = z(t+h) - z(t) - hz'\Big(t+\frac{h}{2}\Big) = \frac{1}{24}\, h^3 z'''(t) + o(h^3),$$

soit une erreur en h^3 au lieu de h^2 dans la méthode d'Euler. On par ailleurs

$$z'\Big(t+\frac{h}{2}\Big) = f\Big(t+\frac{h}{2}, z\Big(t+\frac{h}{2}\Big)\Big).$$

Comme la valeur de $z\Big(t+\frac{h}{2}\Big)$ n'est pas connue, on l'approxime par

$$z\Big(t+\frac{h}{2}\Big) \simeq z(t) + \frac{h}{2}\, f(t, z(t)), \qquad (**)$$

d'où en définitive

$$z(t+h) \simeq z(t) + hf\Big(t+\frac{h}{2}, z(t) + \frac{h}{2}\, f(t, z(t))\Big).$$

L'algorithme du point milieu est associé au choix

$$\Phi(t, y, h) = f\Big(t+\frac{h}{2}, y + \frac{h}{2}\, f(t, y)\Big)$$

et donne lieu au schéma numérique

$$\begin{cases} y_{n+\frac{1}{2}} = y_n + \dfrac{h_n}{2}\ f(t_n, y_n) \\ p_n = f\Big(t_n + \dfrac{h_n}{2}, y_{n+\frac{1}{2}}\Big) \\ y_{n+1} = y_n + h_n p_n \\ t_{n+1} = t_n + h_n. \end{cases}$$

Calculons l'erreur de consistance : $e_n = z(t_{n+1}) - y_{n+1}$, avec $y_n = z(t_n)$. On a $e_n = \varepsilon_n + \varepsilon'_n$ où les erreurs

$$\begin{aligned} \varepsilon_n &= z(t_{n+1}) - z(t_n) - h_n z'\Big(t_n + \frac{h_n}{2}\Big), \\ \varepsilon'_n &= h_n z'\Big(t_n + \frac{h_n}{2}\Big) - (y_{n+1} - z(t_n)) \\ &= h_n \left(f\Big(t_n + \frac{h_n}{2}, z\Big(t_n + \frac{h_n}{2}\Big)\Big) - f\Big(t_n + \frac{h_n}{2}, y_{n+\frac{1}{2}}\Big)\right) \end{aligned}$$

proviennent respectivement des approximations $(*)$ et $(**)$. D'après le calcul fait plus haut

$$\varepsilon_n = \frac{1}{24}\ h_n^3 z'''(t_n) + o(h_n^3) = \frac{1}{24}\ h_n^3 f^{[2]}(t_n, y_n) + o(h_n^3).$$

D'autre part

$$\begin{aligned} z\Big(t_n + \frac{h_n}{2}\Big) - y_{n+\frac{1}{2}} &= z\Big(t_n + \frac{h_n}{2}\Big) - \Big(z(t_n) + \frac{h_n}{2}\, z'(t_n)\Big) \\ &= \frac{1}{8}\ h_n^2 z''(t_n) + o(h_n^2) = \frac{1}{8}\ h_n^2 f^{[1]}(t_n, y_n) + o(h_n^2). \end{aligned}$$

D'après le théorème des accroissements finis appliqué en y, on a

$$\begin{aligned} f\Big(t_n + \frac{h_n}{2}, z\Big(t_n + \frac{h_n}{2}\Big)\Big) - \ & f\Big(t_n + \frac{h_n}{2}, y_{n+\frac{1}{2}}\Big) \\ &= f'_y\Big(t_n + \frac{h_n}{2}, c_n\Big)\Big(z\Big(t_n + \frac{h_n}{2}\Big) - y_{n+\frac{1}{2}}\Big) \\ &= \Big(f'_y(t_n, y_n) + o(h_n)\Big)\Big(\frac{1}{8}\ h_n^2 f^{[1]}(t_n, y_n) + o(h_n^2)\Big) \\ &= \frac{1}{8}\ h_n^2 f'_y f^{[1]}(t_n, y_n) + o(h_n^2), \end{aligned}$$

d'où

$$\varepsilon'_n = \frac{1}{8}\ h_n^3 f'_y f^{[1]}(t_n, y_n) + o(h_n^3).$$

On en déduit

$$e_n = \varepsilon_n + \varepsilon'_n = \frac{1}{24}\ h_n^3 \Big(f^{[2]} + 3 f'_y f^{[1]}\Big)(t_n, y_n) + o(h_n^3).$$

La méthode du point milieu est donc d'ordre 2.

1.5.* Méthode du point milieu modifié

Si on observe les algorithmes précédents, on voit que la seule opération éventuellement coûteuse en temps de calcul est l'évaluation de la fonction $f(t, y)$, le reste consistant en un petit nombre d'additions ou de multiplications. On mesure donc le coût d'une méthode d'ordre donné *par le nombre d'évaluations de la fonction f* qu'elle réclame à chaque pas. Pour des méthodes d'ordres différents la comparaison ne tient pas, puisqu'une méthode d'ordre plus élevé exige à précision égale un nombre de pas nettement inférieur.

Dans la méthode du point milieu, on va modifier le calcul successif de $f(t_n, y_n)$ et de la pente intermédiaire $p_n = f\big(t_n + \frac{h_n}{2}, y_{n+\frac{1}{2}}\big)$ en introduisant l'algorithme suivant, qui fait l'économie de l'évaluation de $f(t_n, y_n)$:

$$\begin{cases} \widetilde{y}_{n+\frac{1}{2}} = \widetilde{y}_n + \dfrac{h_n}{2}\ \widetilde{p}_{n-1} \\ \widetilde{p}_n = f\Big(t_n + \dfrac{h_n}{2}, \widetilde{y}_{n+\frac{1}{2}}\Big) \\ \widetilde{y}_{n+1} = \widetilde{y}_n + h_n \widetilde{p}_n \\ t_{n+1} = t_n + h_n. \end{cases}$$

On a donc modifié légèrement le calcul de $y_{n+\frac{1}{2}}$ en remplaçant la pente $f(t_n, y_n)$ par la pente $\widetilde{p}_{n-1}$ calculée à l'étape antérieure. Il s'ensuit naturellement que les valeurs y_n sont elles aussi modifiées en des valeurs $\widetilde{y}_n$.

Remarque – Le démarrage (étape $n = 0$) présente une difficulté car la pente $\widetilde{p}_{-1}$ n'a pas été évaluée. On résout cette difficulté en initialisant $\widetilde{p}_{-1} = f(t_0, y_0)$. On observera que la méthode du point milieu modifié est en fait une méthode à 2 pas (les étapes n et $n-1$ sont utilisées pour calculer $\widetilde{y}_{n+1}$). ■

Évaluons maintenant l'erreur de consistance $\widetilde{e}_n = z(t_{n+1}) - \widetilde{y}_{n+1}$, en supposant $\widetilde{y}_n = z(t_n)$. On peut écrire

$$\widetilde{e}_n = (z(t_{n+1}) - y_{n+1}) + (y_{n+1} - \widetilde{y}_{n+1}) = e_n + \varepsilon''_n,$$

où e_n est l'erreur de consistance de la méthode du point milieu standard (on suppose donc aussi $y_n = z(t_n)$ pour la calculer). Il vient

$$\begin{aligned} \varepsilon''_n = y_{n+1} - \widetilde{y}_{n+1} &= h_n\Big(f\Big(t_n + \frac{h_n}{2}, y_{n+\frac{1}{2}}\Big) - f\Big(t_n + \frac{h_n}{2}, \widetilde{y}_{n+\frac{1}{2}}\Big)\Big). \\ y_{n+\frac{1}{2}} - \widetilde{y}_{n+\frac{1}{2}} &= \frac{h_n}{2}\ \Big(f(t_n, y_n) - \widetilde{p}_{n-1}\Big) \\ &= \frac{h_n}{2}\ \Big(f(t_n, y_n) - f\Big(t_{n-1} + \frac{h_{n-1}}{2}, \widetilde{y}_{n-\frac{1}{2}}\Big)\Big). \end{aligned}$$

Or $t_n - \left(t_{n-1} + \frac{h_{n-1}}{2}\right) = \frac{h_{n-1}}{2}$ et

$$\begin{aligned} y_n - \widetilde{y}_{n-\frac{1}{2}} &= y_n - \Big(\widetilde{y}_{n-1} + \frac{h_{n-1}}{2}\ \widetilde{p}_{n-2}\Big) \\ &= y_n - \Big(\widetilde{y}_n - h_{n-1}\widetilde{p}_{n-1} + \frac{h_{n-1}}{2}\ \widetilde{p}_{n-2}\Big) \\ &= h_{n-1}\Big(\widetilde{p}_{n-1} - \frac{1}{2}\ \widetilde{p}_{n-2}\Big) \\ &= \frac{1}{2}\ h_{n-1} f(t_n, y_n) + o(h_{n-1})\ ; \end{aligned}$$

à la troisième ligne on utilise le fait que $y_n = \widetilde{y}_n = z(t_n)$, et à la quatrième le fait que $\widetilde{p}_{n-i} = f(t_{n-i} + h_{n-i}/2, \widetilde{y}_{n-i} + 1/2)$ converge vers $f(t_n, y_n)$ pour $i = 1, 2$ lorsque $h_{\max}$ tend vers 0. Grâce à la formule de Taylor pour les fonctions de 2 variables, il vient

$$\begin{aligned} f(t_n, y_n) - f\Big(t_{n-1} + \frac{h_{n-2}}{2}, \widetilde{y}_{n-\frac{1}{2}}\Big) & \\ = \frac{h_{n-1}}{2}\ f'_t(t_n, y_n) + \frac{1}{2}\ h_{n-1} f(t_n, y_n) f'_y(t_n, y_n) + o(h_{n-1}) & \\ = \frac{1}{2}\ h_{n-1}(f'_t + f f'_y)(t_n, y_n) + o(h_{n-1}) & \\ = \frac{1}{2}\ h_{n-1} f^{[1]}(t_n, y_n) + o(h_{n-1}), & \end{aligned}$$

d'où

$$y_{n+\frac{1}{2}} - \widetilde{y}_{n+\frac{1}{2}} = \frac{1}{4} h_n h_{n-1} f^{[1]}(t_n, y_n) + o(h_n h_{n-1}).$$

On en déduit finalement

$$\varepsilon''_n = h_n f'_y\Big(t_n + \frac{h_n}{2}, c_n\Big)\big(y_{n+\frac{1}{2}} - \widetilde{y}_{n+\frac{1}{2}}\big) = \frac{1}{4}\ h_n^2 h_{n-1}\big(f'_y f^{[1]}\big)(t_n, y_n) + o(h_n^2 h_{n-1}),$$

d'où l'erreur de consistance

$$\widetilde{e}_n = \frac{1}{24}\ h_n^3\Big(f^{[2]} + 3 f'_y f^{[1]}\Big)(t_n, y_n) + \frac{1}{4}\ h_n^2 h_{n-1}(f'_y f^{[1]})(t_n, y_n) + o(h_n^3 + h_n^2 h_{n-1}).$$

La méthode du point milieu modifié est donc encore une méthode d'ordre 2 (mais ce n'est pas une méthode à un pas !).

2. *Étude générale des méthodes à un pas*

2.1. *Méthodes consistantes, stables et convergentes*

La première notion que nous introduisons a trait au problème de l'accumulation des erreurs de consistance (accumulation purement théorique dans le sens où on tient pas compte du fait que la solution calculée s'écarte de la solution exacte, cf. schémas du § 1.1).

Définition 1 – *On dit que la méthode est* consistante *si pour toute solution exacte z la somme des erreurs de consistance relatives à z, soit $\sum_{0\le n\le N} |e_n|$, tend vers 0 quand $h_{\max}$ tend vers 0.*

Une autre notion fondamentale est la notion de stabilité. Dans la pratique, le calcul récurrent des points y_n est en effet entâché d'erreurs d'arrondi ε_n. Pour que les calculs soient significatifs, il est indispensable que la propagation de ces erreurs reste contrôlable. On est amené à la définition suivante.

Définition 2 – *On dit que la méthode est* stable *s'il existe une constante $S \ge 0$, appelée* constante de stabilité*, telle que pour toutes suites (y_n), $(\widetilde{y}_n)$ définies par*

$$\begin{aligned} y_{n+1} &= y_n + h_n\Phi(t_n, y_n, h_n), && 0 \le n < N \\ \widetilde{y}_{n+1} &= \widetilde{y}_n + h_n\Phi(t_n, \widetilde{y}_n, h_n) + \varepsilon_n, && 0 \le n < N \end{aligned}$$

on ait

$$\max_{0\le n\le N} |\widetilde{y}_n - y_n| \le S\Big(|\widetilde{y}_0 - y_0| + \sum_{0\le n<N} |\varepsilon_n|\Big).$$

Autrement dit, une petite erreur initiale $|\widetilde{y}_0 - y_0|$ et de petites erreurs d'arrondi ε_n dans le calcul récurrent des $\widetilde{y}_n$ provoquent une erreur finale $\max |\widetilde{y}_n - y_n|$ contrôlable. Une dernière notion importante en pratique est la suivante.

Définition 3 – *On dit que la méthode est* convergente *si pour toute solution exacte z, la suite y_n telle que $y_{n+1} = y_n + h_n\Phi(t_n, y_n, h_n)$ vérifie*

$$\max_{0\le n\le N} |y_n - z(t_n)| \to 0$$

quand $y_0 \to z(t_0)$ et quand $h_{\max} \to 0$.

La quantité $\max_{0\le n\le N} |y_n - z(t_n)|$ s'appelle *l'erreur globale* (de la suite y_n calculée par rapport à la solution exacte z). C'est évidemment cette erreur qui importe dans la pratique.

Calcul de l'erreur globale – Posons $\widetilde{y}_n = z(t_n)$. Par définition de l'erreur de consistance (cf. § 2.1) on a

$$\widetilde{y}_{n+1} = \widetilde{y}_n + h_n\Phi(t_n, \widetilde{y}_n, h_n) + e_n.$$

Si la méthode est stable, de constante de stabilité S, l'erreur globale est donc

$$\max_{0\le n\le N} |y_n - z(t_n)| \le S\Big(|y_0 - z(t_0)| + \sum_{0\le n<N} |e_n|\Big).$$

Corollaire – *Si la méthode est stable et consistante, elle est convergente.*

En effet $\sum_{0\leq n<N}|e_n|$ tend vers 0 quand $h_{\max}$ tend vers 0, puisque la méthode est consistante par hypothèse.

2.2. CONDITION NÉCESSAIRE ET SUFFISANTE DE CONSISTANCE

Soit z une solution exacte de l'équation (E) et soient

$$e_n = z(t_{n+1}) - z(t_n) - h_n\Phi(t_n, z(t_n), h_n)$$

les erreurs de consistance correspondantes. D'après le théorème des accroissements finis, il existe $c_n \in]t_n, t_{n+1}[$ tel que

$$z(t_{n+1}) - z(t_n) = h_n z'(c_n) = h_n f(c_n, z(c_n))$$

d'où

$$e_n = h_n(f(c_n, z(c_n)) - \Phi(t_n, z(t_n), h_n)) = h_n(\alpha_n + \beta_n)$$

avec

$$\alpha_n = f(c_n, z(c_n)) - \Phi(c_n, z(c_n), 0),$$
$$\beta_n = \Phi(c_n, z(c_n), 0) - \Phi(t_n, z(t_n), h_n).$$

Comme la fonction $(t,h) \mapsto \Phi(t, z(t), h)$ est continue sur $[t_0, t_0+T] \times [0, \delta]$ qui est compact, elle y est uniformément continue. Par conséquent, pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que $h_{\max} \leq \eta \to |\beta_n| \leq \varepsilon$. Pour $h_{\max} \leq \eta$ on a donc

$$\left| \sum_{0\leq n<N} |e_n| - \sum_{0\leq n<N} h_n|\alpha_n| \right| \leq \sum_{0\leq n<N} h_n|\beta_n| \leq \varepsilon \sum h_n = T\varepsilon.$$

On en déduit

$$\begin{aligned}\lim_{h_{\max}\to 0} \sum_{0\leq n<N} |e_n| &= \lim_{h_{\max}\to 0} \sum_{0\leq n<N} h_n|\alpha_n| \\ &= \int_{t_0}^{t_0+T} |f(t, z(t)) - \Phi(t, z(t), 0)|dt\end{aligned}$$

car $\sum h_n|\alpha_n|$ est une somme de Riemann de l'intégrale précédente. Par définition, la méthode est consistante si et seulement si $\lim \sum |e_n| = 0$ pour toute solution exacte z. On en déduit :

Théorème – *La méthode à* 1 *pas définie par la fonction* Φ *est consistante si et seulement si*

$$\forall (t,y) \in [t_0, t_0+T] \times \mathbb{R}, \quad \Phi(t, y, 0) = f(t, y).$$

Il résulte de ce théorème que les méthodes à un pas déjà mentionnées sont bien consistantes.

2.3. Condition suffisante de stabilité

Pour pouvoir majorer l'erreur globale décrite au § 2.1, il faut savoir estimer d'une part la constante de stabilité S, et d'autre part la somme $\sum_{0\le n<N}|e_n|$. Le résultat suivant permet d'évaluer S.

Théorème – *Pour que la méthode soit stable, il suffit que la fonction Φ soit lipschitzienne en y, c'est-à-dire qu'il existe une constante $\Lambda\ge 0$ telle que*

$\forall t\in[t_0,t_0+T]$, $\forall (y_1,y_2)\in\mathbb{R}^2$, $\forall h\in\mathbb{R}$ *on ait*

$$|\Phi(t,y_1,h)-\Phi(t,y_2,h)|\le\Lambda|y_1-y_2|.$$

Dans ce cas, on peut prendre pour constante de stabilité $S=e^{\Lambda T}$.

Démonstration. Considérons deux suites (y_n), $(\widetilde{y}_n)$ telles que

$$\begin{aligned} y_{n+1}&=y_n+h_n\Phi(t_n,y_n,h_n),\\ \widetilde{y}_{n+1}&=\widetilde{y}_n+h_n\Phi(t_n,\widetilde{y}_n,h_n)+\varepsilon_n.\end{aligned}$$

Par différence, on obtient

$$|\widetilde{y}_{n+1}-y_{n+1}|\le|\widetilde{y}_n-y_n|+h_n\Lambda|\widetilde{y}_n-y_n|+|\varepsilon_n|.$$

En posant $\theta_n=|\widetilde{y}_n-y_n|$, il vient

$$\theta_{n+1}\le(1+\Lambda h_n)\theta_n+|\varepsilon_n|.$$

Lemme de Gronwall (cas discret) – *Soient des suites $h_n,\theta_n\ge 0$ et $\varepsilon_n\in\mathbb{R}$ telles que $\theta_{n+1}\le(1+\Lambda h_n)\theta_n+|\varepsilon_n|$. Alors*

$$\theta_n\le e^{\Lambda(t_n-t_0)}\theta_0+\sum_{0\le i\le n-1}e^{\Lambda(t_n-t_{i+1})}|\varepsilon_i|.$$

Le lemme se vérifie par récurrence sur n. Pour $n=0$, l'inégalité se réduit à $\theta_0\le\theta_0$. Supposons maintenant l'inégalité vraie à l'ordre n. On observe que

$$1+\Lambda h_n\le e^{\Lambda h_n}=e^{\Lambda(t_{n+1}-t_n)}.$$

Par hypothèse on a

$$\begin{aligned}\theta_{n+1}&\le e^{\Lambda(t_{n+1}-t_n)}\theta_n+|\varepsilon_n|\\ &\le e^{\Lambda(t_{n+1}-t_0)}\theta_0+\sum_{0\le i\le n-1}e^{\Lambda(t_{n+1}-t_{i+1})}|\varepsilon_i|+|\varepsilon_n|.\end{aligned}$$

L'inégalité cherchée s'ensuit à l'ordre $n+1$. ∎

Comme $t_n - t_0 \leq T$ et $t_n - t_{i+1} \leq T$, le lemme de Gronwall implique

$$\max_{0\leq n\leq N} \theta_n \leq e^{\Lambda T}\Big(\theta_0 + \sum_{0\leq i\leq N-1} |\varepsilon_i|\Big)$$

Par définition de θ_n, on a donc

$$\max_{0\leq n\leq N} |\widetilde{y}_n - y_n| \leq e^{\Lambda T}\Big(|\widetilde{y}_0 - y_0| + \sum_{0\leq n\leq N} |\varepsilon_n|\Big),$$

et le théorème est démontré. ■

Remarque – Dans la pratique, l'hypothèse lipschitzienne faite sur Φ est très rarement satisfaite globalement pour $y_1, y_2 \in \mathbb{R}$ et $h \in \mathbb{R}$ (ne serait-ce que parce que le domaine de définition de Φ est peut-être plus petit). Par contre cette hypothèse est souvent satisfaite si on se restreint à des $y_1, y_2 \in J$ et $|h| \leq \delta$ où J est un intervalle fermé borné assez petit. Dans ce cas, la constante de stabilité $S = e^{\Lambda T}$ est valable pour des suites $y_n, \widetilde{y}_n \in J$ et pour $h_{\max} \leq \delta$.

Exemples – Supposons que la fonction f soit lipschitzienne de rapport k en y et calculons Λ, S pour les différentes méthodes déjà présentées.

- Dans la méthode d'Euler, $\Phi(t, y, h) = f(t, y)$. On peut prendre $\Lambda = k$, $S = e^{kT}$.
- Dans la méthode du point milieu, on a

$$\Phi(t, y, h) = f\Big(t + \frac{h}{2}, y + \frac{h}{2}\ f(t, y)\Big).$$

On en déduit

$$\begin{aligned}
|\Phi(t, y_1, h) - \Phi(t, y_2, h)| &\leq k\Big|y_1 + \frac{h}{2}\ f(t, y_1) - \Big(y_2 + \frac{h}{2}\ f(t, y_2)\Big)\Big| \\
&\leq k\Big(|y_1 - y_2| + \frac{h}{2}\ |f(t, y_1) - f(t, y_2)|\Big) \\
&\leq k\Big(|y_1 - y_2| + \frac{h}{2}\ k|y_1 - y_2|\Big) = k\Big(1 + \frac{1}{2}\ hk\Big)|y_1 - y_2|.
\end{aligned}$$

On peut prendre ici $\Lambda = k\Big(1 + \frac{1}{2}\ h_{\max}\, k\Big)$, d'où

$$S = \exp\Big(kT\Big(1 + \frac{1}{2}\ h_{\max} k\Big)\Big).$$

Si $h_{\max}$ est petit (par rapport à $1/k^2T$), cette constante est du même ordre de grandeur que dans la méthode d'Euler.

Corollaire – *Lorsque f est lipschitzienne en y, les méthodes d'Euler et du point milieu sont convergentes.*

Le lecteur observera l'analogie des techniques utilisées pour obtenir ce corollaire avec celles utilisées au chapitre V, § 3.1.

2.4. Influence de l'ordre de la méthode sur l'erreur globale

Nous allons voir ici que l'ordre d'une méthode numérique a une influence déterminante sur la précision que cette méthode permet d'atteindre. La définition de l'ordre que nous allons donne prend l'erreur de consistance comme point de départ (le lecteur prendra garde au décalage d'une unité entre l'ordre et l'exposant de h_n dans l'erreur de consistance !)

Définition – *On dit qu'une méthode à* 1 *pas est d'ordre* $\geq p$ *si pour toute solution exacte* z *d'une équation différentielle*

$$\text{(E)} \qquad y' = f(t,y) \quad \text{où } f \text{ est de classe } C^p,$$

il existe une constante $C \geq 0$ *telle que l'erreur de consistance relative à* z *vérifie*

$$|e_n| \leq C h_n^{p+1}, \quad \forall n, \quad 0 \leq n < N.$$

Elle est dite d'ordre p (*exactement*) *si elle est d'ordre* $\geq p$ *mais pas d'ordre* $\geq p+1$.

Rappelons que l'erreur de consistance est donnée par

$$e_n = z(t_{n+1}) - y_n - h_n \Phi(t_n, y_n, h_n) \quad \text{où} \quad y_n = z(t_n).$$

Supposons que Φ soit de classe C^p. La formule de Taylor donne alors

$$\Phi(t_n, y_n, h_n) = \sum_{l=0}^{p} \frac{1}{l!}\, h_n^l \frac{\partial^l \Phi}{\partial h^l}\,(t_n, y_n, 0) + o(h_n^p).$$

Si f est de classe C^p, la solution z est de classe C^{p+1} donc

$$\begin{aligned} z(t_{n+1}) - y_n &= z(t_n + h) - z(t_n) \\ &= \sum_{k=1}^{p+1} \frac{1}{k!}\, h_n^k z^{(k)}(t_n) + o(h_n^{n+1}) \\ &= \sum_{l=0}^{p} \frac{1}{(l+1)!}\, h_n^{l+1} f^{[l]}(t_n, y_n) + o(h_n^{p+1}). \end{aligned}$$

On en déduit aussitôt

$$e_n = \sum_{l=0}^{p} \frac{1}{l!}\, h_n^{l+1} \Big(\frac{1}{l+1}\, f^{[l]}(t_n, y_n) - \frac{\partial^l \Phi}{\partial h^l}\,\Big(t_n, y_n, 0\Big)\Big) + o(h_n^{p+1})$$

Conséquence – *La méthode est d'ordre* $\geq p$ *si et seulement si* Φ *est telle que*

$$\frac{\partial^l \Phi}{\partial h^l}\,(t, y, 0) = \frac{1}{l+1}\, f^{[l]}(t, y), \quad 0 \leq l \leq p-1.$$

Sous cette hypothèse, l'erreur e_n se réduit à

$$e^n = \frac{1}{p!}\ h_n^{p+1}\Big(\frac{1}{p+1}\ f^{[p]}(t_n, y_n) - \frac{\partial^p \Phi}{\partial h^p}\ (t_n, y_n, 0)\Big) + o(h_n^{p+1}).$$

Remarque – *De ce qui précède on déduit les équivalences*

$$\textit{Méthode consistante} \quad \Leftrightarrow \quad \Phi(t,y,0) = f(t,y) \quad \Leftrightarrow \quad \textit{Méthode d'ordre} \quad \geq 1.$$

L'utilisation de la formule de Taylor avec reste de Lagrange implique l'existence de points $\tau_n \in]t_n, t_{n+1}[$ et $\eta_n \in]0, h_{\max}[$ tels que

$$e_n = h_n^{p+1}\left(\frac{1}{(p+1)!}\ f^{[p]}(\tau_n, z(\tau_n)) - \frac{1}{p!}\ \frac{\partial^p \Phi}{\partial h^p}\ (t_n, z(t_n), \eta_n)\right).$$

Ceci permet (au moins théoriquement) de trouver une constante C dans la majoration de l'erreur de consistance : on peut prendre

$$C = \frac{1}{(p+1)!}\ \|f^{[p]}(t, z(t))\|_\infty + \frac{1}{p!}\ \|\frac{\partial^p \Phi}{\partial h^p}\ (t, z(t), h)\|_\infty$$

où les normes $\|\quad\|_\infty$ sont étendues aux $\quad (t,h) \in [t_0, t_0+T] \times [0, h_{\max}]$.

Majoration de l'erreur globale – Compte tenu de la majoration supposée satisfaite pour e_n, on a

$$\sum_{0 \leq n < N} |e_n| \leq \sum C h_n^{p+1} \leq C \sum h_n h_{\max}^p \leq C T h_{\max}^p.$$

Si la méthode est stable avec constante de stabilité S, on obtient donc la majoration

$$\max_{0 \leq n \leq N} |y_n - z(t_n)| \leq S(|y_0 - z(t_0)| + C T h_{\max}^p).$$

L'erreur initiale $|y_0 - z(t_0)|$ est généralement négligeable. L'erreur globale donnée par une méthode stable d'ordre p est donc de l'ordre de grandeur de $h_{\max}^p$ avec une constante de proportionnalité SCT (on retiendra que l'ordre est égal à l'exposant de $h_{\max}$ dans la majoration de l'erreur globale, alors que l'erreur de consistance, elle, est en h_n^{p+1}).

Si la constante SCT n'est pas trop grande (disons $\leq 10^2$), une méthode d'ordre 3 avec pas maximum $h_{\max} = 10^{-2}$ permet d'atteindre une précision globale de l'ordre de 10^{-4}.

2.5. Influence des erreurs d'arrondi

L'erreur globale calculée au § 2.4 est une erreur théorique, c'est-à-dire qu'elle ne tient pas compte des erreurs d'arrondi qui se produisent inévitablement en pratique. Dans la réalité l'ordinateur va calculer non pas la suite récurrente y_n, mais une valeur approchée $\widetilde{y}_n$ de y_n dans laquelle interviendront

- une erreur d'arrondi ρ_n sur $\Phi(t_n, \widetilde{y}_n, h_n)$,
- une erreur d'arrondi σ_n sur le calcul de $\widetilde{y}_{n+1}$.

En définitive, on aura

$$\begin{aligned}\widetilde{y}_{n+1} &= \widetilde{y}_n + h_n(\Phi(t_n, \widetilde{y}_n, h_n) + \rho_n) + \sigma_n \\ &= \widetilde{y}_n + h_n\Phi(t_n, \widetilde{y}_n, h_n) + h_n\rho_n + \sigma_n.\end{aligned}$$

Il se peut également que $\widetilde{y}_0$ diffère légèrement de la valeur théorique y_0 : $\widetilde{y}_0 = y_0 + \varepsilon_0$.

Hypothèse – $\quad \forall n, \quad |\rho_n| \le \rho, \quad |\sigma_n| \le \sigma.$

Les constantes ρ, σ dépendent des caractéristiques de l'ordinateur et de la précision des opérations arithmétiques (si les réels sont codés sur 6 octets, on a typiquement $\rho = 10^{-9}$, $\sigma = 10^{-10}$, $|\varepsilon_0| \le 10^{-10}$).

Si la méthode est stable avec constante de stabilité S, on en déduit

$$\begin{aligned}\max_{0\le n\le N} |\widetilde{y}_n - y_n| &\le S\Big(|\varepsilon_0| + \sum_{0\le n<N} (h_n|\rho_n| + |\sigma_n|)\Big) \\ &\le S(|\varepsilon_0| + T\rho + N\sigma).\end{aligned}$$

A cette erreur due aux arrondis s'ajoute l'erreur globale théorique

$$\max_{0\le n\le N} |y_n - z(t_n)| \le SCTh^p_{\max} \qquad \text{si} \qquad y_0 = z(t_0).$$

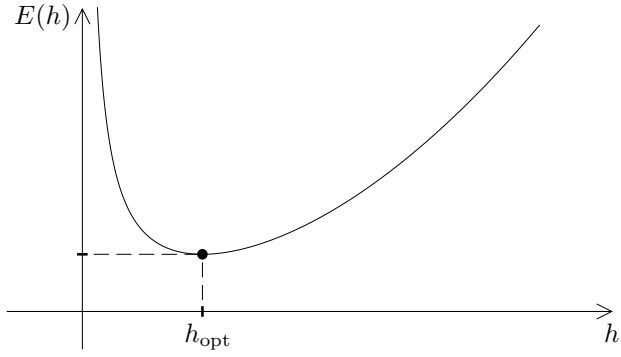
L'erreur totale commise est donc

$$\max_{0\le n\le N} |\widetilde{y}_n - z(t_n)| \le S(|\varepsilon_0| + T\rho + N\sigma + CTh^p_{\max})$$

Supposons le pas $h_n = h$ constant pour simplifier. On a alors $N = \frac{T}{h}$ et l'erreur est majorée par

$$E(h) = S(|\varepsilon_0| + T\rho + \frac{T}{h}\,\sigma + CTh^p) = S(|\varepsilon_0| + T\rho) + ST\Big(\frac{\sigma}{h} + Ch^p\Big)$$

L'étude de $E(h)$ donne la courbe suivante, avec un minimum de $E(h)$ réalisé en $h_{\text{opt}} = \big(\frac{\sigma}{pC}\big)^{\frac{1}{p+1}}$.



Typiquement, pour une méthode d'ordre $p = 2$ où $pC \simeq 10$, on obtient $h_{\text{opt}} \simeq 10^{-3}$. Si on prend un pas plus petit, l'erreur augmente ! Ceci est dû au fait que le nombre de pas $N = \frac{T}{h}$ augmente, et avec lui les erreurs d'arrondi, lorsque le pas h diminue. Les erreurs d'arrondi l'emportent alors sur l'erreur globale théorique $SCTh^p$. L'expérience numérique suivante confirme ces prévisions théoriques.

Exemple – Considérons le problème de Cauchy $y' = y$ avec donnée $y_0 = 1$ en $t_0 = 0$. La solution exacte est $y(t) = e^t$, d'où $y(1) = e \simeq 2,7182818285$. Si l'on utilise la méthode du point milieu avec pas constant h, on obtient l'algorithme

$$y_{n+1} = (1 + h + h^2/2)y_n, \quad y_0 = 1.$$

L'erreur de consistance est donnée par $e_n \sim h^3 y_n/6$, d'où $e_n \leq Ch^3$ avec $C = e/6$ sur $[0, T] = [0, 1]$. Par ailleurs, on peut au mieux espérer $\sigma = 10^{-11}$, ce qui donne

$$h_{\text{opt}} \geq \left(\frac{10^{-11}}{2 \cdot e/6}\right)^{1/3} \simeq 2,224 \cdot 10^{-4}, \qquad N_{\text{opt}} = \frac{T}{h_{\text{opt}}} < 4500.$$

Un calcul sur un ordinateur disposant d'une précision relative maximale des réels de 10^{-11} environ nous a donné en fait les résultats suivants :

Nombre de pas	Valeur y_N associée	Erreur $y(1) - y_N$
$N =$ 10	2,7140808465	$4,2 \cdot 10^{-3}$
$N =$ 100	2,7182368616	$4,5 \cdot 10^{-5}$
$N =$ 500	2,7182800146	$4,8 \cdot 10^{-6}$
$N =$ 1000	2,7182813650	$4,6 \cdot 10^{-7}$
$N =$ 2000	2,7182816975	$1,3 \cdot 10^{-7}$
$N =$ 3000	2,7182817436	$8,5 \cdot 10^{-8}$
$N =$ 4000	2,7182817661	$6,2 \cdot 10^{-8}$
$N =$ 4400	2,7182817882	$4,0 \cdot 10^{-8}$
$N =$ 5000	2,7182817787	$5,0 \cdot 10^{-8}$
$N =$ 6000	2,7182817607	$6,8 \cdot 10^{-8}$
$N =$ 7000	2,7182817507	$7,8 \cdot 10^{-8}$
$N =$ 10000	2,7182817473	$8,1 \cdot 10^{-8}$
$N =$ 20000	2,7182817014	$1,3 \cdot 10^{-7}$

Le nombre de pas optimal observé est $N_{\text{opt}} \simeq 4400$.

2.6. Problèmes bien posés, bien conditionnés, problèmes raides

L'objet de ce paragraphe est de mettre en évidence les difficultés qui peuvent apparaître dans la mise en œuvre des algorithmes de résolution numérique.

Définition 1 – *On dit qu'un problème de Cauchy est mathématiquement bien posé si la solution est unique et dépend continûment de la donnée initiale.*

Exemple – Considérons le problème de Cauchy

$$\begin{cases} y' = 2\sqrt{|y|}, & t \in [0, +\infty[\\ y(0) = 0. \end{cases}$$

Ce problème admet les solutions $y(t) = 0$, $y(t) = t^2$ et plus généralement

$$\begin{cases} y(t) = 0, & t \in [0, a] \\ y(t) = (t-a)^2, & t \in [a, +\infty[. \end{cases}$$

L'utilisation de la méthode d'Euler $y_{n+1} = y_n + 2h_n\sqrt{y_n}$ va conduire aux solutions approchées suivantes :

- si $y_0 = 0$ $\quad y(t) = 0$
- si $y_0 = \varepsilon$ $\quad y(t) \simeq (t + \sqrt{\varepsilon})^2$ quand $h_{\max} \to 0$.

Il n'y a ici ni unicité, ni continuité de la solution. Le problème de Cauchy est donc mathématiquement mal posé.

Les résultats du chapitre V § 3.2 (et ceux à venir du chapitre XI § 1.2) montrent que le problème de Cauchy est *mathématiquement bien posé dès que $f(t, y)$ est localement lipschitzienne* en y.

Définition 2 – *On dit qu'un problème de Cauchy est numériquement bien posé si la continuité de la solution par rapport à la donnée initiale est suffisamment bonne pour que la solution ne soit pas perturbée par une erreur initiale ou des erreurs d'arrondi faibles.*

Par l'expression « continuité suffisamment bonne », on entend en général l'existence d'une constante de Lipschitz petite en regard de la précision des calculs. On notera que la définition 2 ne fait pas référence à la méthode de calcul utilisée.

Exemple 2 – Soit le problème de Cauchy

$$\begin{cases} y' = 3y - 1, & t \in [0, 10] \\ y(0) = \frac{1}{3}. \end{cases}$$

La solution exacte est $y(t) = t + \frac{1}{3}$. La donnée initiale $\widetilde{y}(0) = \frac{1}{3} + \varepsilon$ fournit $\widetilde{y}(t) = t + \frac{1}{3} + \varepsilon e^{3t}$. On a alors

$$\widetilde{y}(10) - y(10) = \varepsilon \cdot e^{30} \simeq 10^{13}\varepsilon.$$

Le problème est ici mathématiquement bien posé, mais numériquement mal posé si la précision des calculs est seulement de 10^{-10}. Le problème redevient numériquement bien posé si la précision des calculs est de 10^{-20}.

Exemple 3 – L'exemple suivant montre que même un problème numériquement bien posé peut soulever des difficultés inattendues :

$$\begin{cases} y' = -150y + 30, \quad t \in [0,1], \\ y(0) = \frac{1}{5}. \end{cases}$$

La solution exacte est $y(t) = \frac{1}{5}$ et la donnée initiale $\widetilde{y}(0) = \frac{1}{5} + \varepsilon$ fournit $\widetilde{y}(t) = \frac{1}{5} + \varepsilon e^{-150t}$. Comme $0 \leq e^{-150t} \leq 1$ sur $[0,1]$, le problème est numériquement bien posé. La méthode d'Euler avec pas constant h donne

$$\begin{aligned} & y_{n+1} = y_n + h(-150y_n + 30) = (1 - 150h)y_n + 30h, \\ & y_{n+1} - \frac{1}{5} = (1 - 150h)(y_n - \frac{1}{5}), \\ \text{d'où} \qquad & y_n - \frac{1}{5} = (1 - 150h)^n \Big(y_0 - \frac{1}{5}\Big). \end{aligned}$$

Supposons $h = \frac{1}{50}$. Une erreur initiale $y_0 = \frac{1}{5} + \varepsilon$ conduit à $y_n = \frac{1}{5} + (-2)^m \varepsilon$, d'où pour $t = 1$

$$y_{50} = \frac{1}{5} + 2^{50}\varepsilon \simeq \frac{1}{50} + 10^{15}\varepsilon!$$

Pour que $|y_n|$ ne diverge pas vers $+\infty$, il est nécessaire de prendre $|1 - 150h| \leq 1$, soit $150h \leq 2$, $h \leq \frac{1}{75}$. Bien que le problème soit tout à fait bien posé, on voit qu'il est nécessaire de prendre un pas assez petit, et donc de faire des calculs plus coûteux que d'ordinaire.

Définition 3 – *On dit qu'un problème est bien conditionné si les méthodes numériques usuelles peuvent en donner la solution en un temps raisonnable.*

Un problème sera bien conditionné si la constante de stabilité S n'est pas trop grande (disons nettement $< 10^{10}$ si la précision des calculs est 10^{-10}). Sinon, on dit qu'on a affaire à un *problème raide*.

On sait qu'en général la constante de stabilité S est majorée par $e^{\Lambda T}$. Dans un problème raide, on peut avoir typiquement $\Lambda T = 10^3$, $e^{\Lambda T} > 10^{400}$. Il existe des algorithmes permettant de traiter certains problèmes raides, mais nous n'aborderons pas cette question. Le lecteur pourra consulter sur ce point le livre de Crouzeix-Mignot.

3. Méthodes de Runge-Kutta

3.1. Principe général

On considère un problème de Cauchy

$$\begin{cases} y' = f(t,y), & t \in [t_0, t_0+T] \\ y(t_0) = y_0 \end{cases}$$

et on cherche à discrétiser ce problème par rapport à une subdivision $t_0 < t_1 < \ldots < t_N = t_0 + T$. L'idée est de calculer par récurrence les points (t_n, y_n) en utilisant des points intermédiaires $(t_{n,i}, y_{n,i})$ avec

$$t_{n,i} = t_n + c_i h_n, \quad 1 \leq i \leq q, \quad c_i \in [0,1].$$

A chacun de ces points on associe la pente correspondante

$$p_{n,i} = f(t_{n,i}, y_{n,i}).$$

Soit z une solution exacte de l'équation. On a

$$\begin{aligned} z(t_{n,i}) &= z(t_n) + \int_{t_n}^{t_{n,i}} f(t, z(t))dt \\ &= z(t_n) + h_n \int_0^{c_i} f(t_n + uh_n, z(t_n + uh_n))du \end{aligned}$$

grâce au changement de variable $t = t_n + uh_n$. De même

$$z(t_{n+1}) = z(t_n) + h_n \int_0^1 f(t_n + uh_n, z(t_n + uh_n)du.$$

On se donne alors pour chaque $i = 1, 2, \ldots, q$ une méthode d'intégration approchée

$$(\mathrm{M}_i) \qquad \int_0^{c_i} g(t)dt \simeq \sum_{1 \leq j < i} a_{ij} g(c_j),$$

ces méthodes pouvant être *a priori* différentes. On se donne également une méthode d'intégration approchée sur $[0,1]$:

$$(\mathrm{M}) \qquad \int_0^1 g(t)dt \simeq \sum_{1 \leq j \leq q} b_j g(c_j).$$

En appliquant ces méthodes d'intégration à $g(u) = f(t_n + uh_n, z(t_n + uh_n))$, il vient

$$\begin{aligned} z(t_{n,i}) &\simeq z(t_n) + h_n \sum_{1 \leq j < i} a_{ij} f(t_{n,j}, z(t_{n,j})), \\ z(t_{n+1}) &\simeq z(t_n) + h_n \sum_{1 \leq j \leq q} b_j f(t_{n,j}, z(t_{n,j})). \end{aligned}$$

La méthode Runge-Kutta correspondante est définie par l'algorithme

$$\begin{cases} \left[\begin{array}{l} t_{n,i} = t_n + c_i h_n \\ y_{n,i} = y_n + h_n \displaystyle\sum_{1\le j<i} a_{ij} p_{n,j} \\ p_{n,i} = f(t_{n,i}, y_{n,i}) \end{array}\right] & 1 \le i \le q \\ t_{n+1} = t_n + h_n \\ y_{n+1} = y_n + h_n \displaystyle\sum_{1\le j\le q} b_j p_{n,j}. \end{cases}$$

On la représente conventionnellement par le tableau

$$\begin{array}{lc|ccccc} (\mathrm{M}_1) & c_1 & 0 & 0 & \dots & 0 & 0 \\ (\mathrm{M}_2) & c_2 & a_{21} & 0 & \dots & 0 & 0 \\ & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & \vdots & \vdots & \vdots & & 0 & 0 \\ (\mathrm{M}_q) & c_q & a_{q1} & a_{q2} & \dots & a_{qq-1} & 0 \\ \hline (\mathrm{M}) & & b_1 & b_2 & \dots & b_{q-1} & b_q \end{array}$$

où les méthodes d'intégration approchées correspondent aux lignes. On pose par convention $a_{ij} = 0$ pour $j \ge i$.

Hypothèse – *On supposera toujours que les méthodes d'intégration (M_i) et (M) sont d'ordre 0 au moins, c'est-à-dire*

$$c_i = \sum_{1\le j<i} a_{ij}, \qquad 1 = \sum_{1\le j\le q} b_j.$$

En particulier, on aura toujours

$$c_1 = 0, \quad t_{n,1} = t_n, \quad y_{n,1} = y_n, \quad p_{n,1} = f(t_n, y_n).$$

3.2. Exemples

Exemple 1 – Pour $q = 1$, le seul choix possible est $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$

On a ici $c_1 = 0$, $a_{11} = 0$, $b_1 = 1$. L'algorithme est donné par

$$\begin{cases} p_{n,1} = f(t_n, y_n) \\ t_{n+1} = t_n + h_n \\ y_{n+1} = y_n + h_n p_{n,1} \end{cases}$$

Il s'agit de la méthode d'Euler.

Exemple 2 – Pour $q = 2$, on considère les tableaux de la forme

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1-\dfrac{1}{2\alpha} & \dfrac{1}{2\alpha} \end{array}, \qquad \text{où} \quad \alpha \in]0,1].$$

L'algorithme s'écrit ici

$$\begin{cases} p_{n,1} = f(t_n, y_n) \\ t_{n,2} = t_n + \alpha h_n \\ y_{n,2} = y_n + \alpha h_n p_{n,1} \\ p_{n,2} = f(t_{n,2}, y_{n,2}) \\ t_{n+1} = t_n + h_n \\ y_{n+1} = y_n + h_n \Big(\Big(1 - \frac{1}{2\alpha}\Big)\ p_{n,1} + \frac{1}{2\alpha}\ p_{n,2} \Big), \end{cases}$$

ou encore, sous forme condensée :

$$y_{n+1} = y_n + h_n \Big(\Big(1 - \frac{1}{2\alpha}\Big)\ f(t_n, y_n) + \frac{1}{2\alpha}\ f(t_n + \alpha h_n, y_n + \alpha h_n f(t, y_n)) \Big).$$

C'est néanmoins la première formulation qui est la plus efficace en pratique, puisqu'elle requiert seulement deux évaluations de la fonction f au lieu de 3 pour la forme condensée.

- Pour $\alpha = \frac{1}{2}$, on retrouve la méthode du point milieu

$$y_{n+1} = y_n + h_n f\Big(t_n + \frac{h_n}{2}, y_n + \frac{h_n}{2}\ f(t_n, y_n)\Big),$$

qui est basée sur la méthode d'intégration du point milieu :

$$\text{(M)} \qquad \int_0^1 g(t)dt \simeq g\Big(\frac{1}{2}\Big).$$

- Pour $\alpha = 1$, on obtient la *méthode de Heun* :

$$y_{n+1} = y_n + h_n \Big(\frac{1}{2}\ f(t_n, y_n) + \frac{1}{2}\ f(t_{n+1}, y_n + h_n f(t_n, y_n))\Big),$$

qui repose sur la méthode d'intégration des trapèzes :

$$\text{(M)} \qquad \int_0^1 g(t)dt \simeq \frac{1}{2}\ (g(0) + g(1)).$$

Exemple 3 – *Méthode de Runge-Kutta « classique » :*
Il s'agit de la méthode définie par le tableau

$$q = 4, \qquad \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$$

L'algorithme correspondant s'écrit

$$\begin{cases} p_{n,1} = f(t_n, y_n) \\ t_{n,2} = t_n + \frac{1}{2}\ h_n \\ y_{n,2} = y_n + \frac{1}{2}\ h_n p_{n,1} \\ p_{n,2} = f(t_{n,2}, y_{n,2}) \\ y_{n,3} = y_n + \frac{1}{2}\ h_n p_{n,2} \\ p_{n,3} = f(t_{n,2}, y_{n,3}) \qquad \text{(noter que } t_{n,3} = t_{n,2}) \\ t_{n+1} = t_n + h_n \qquad \text{(noter que } t_{n,4} = t_{n+1}) \\ y_{n,4} = y_n + h_n p_{n,3} \\ p_{n,4} = f(t_{n+1}, y_{n,4}) \\ y_{n+1} = y_n + h_n \Big(\frac{1}{6}\ p_{n,1} + \frac{2}{6}\ p_{n,2} + \frac{2}{6}\ p_{n,3} + \frac{1}{6}\ p_{n,4} \Big) \end{cases}$$

On verra plus loin que cette méthode est d'ordre 4. Dans ce cas les méthodes d'intégration (M_i) et (M) utilisées sont respectivement :

$$(\mathrm{M}_2) \quad \int_0^{\frac{1}{2}} g(t)dt \simeq \frac{1}{2}\ g(0) : \qquad \text{rectangles à gauche,}$$

$$(\mathrm{M}_3) \quad \int_0^{\frac{1}{2}} g(t)dt \simeq \frac{1}{2}\ g\Big(\frac{1}{2}\Big) : \qquad \text{rectangles à droite,}$$

$$(\mathrm{M}_4) \quad \int_0^{1} g(t)dt \simeq g\Big(\frac{1}{2}\Big) : \qquad \text{point milieu,}$$

$$(\mathrm{M}) \quad \int_0^{1} g(t)dt \simeq \frac{1}{6}\ g(0) + \frac{2}{6}\ g\Big(\frac{1}{2}\Big) + \frac{2}{6}\ g\Big(\frac{1}{2}\Big) + \frac{1}{6}\ g(1) : \quad \text{Simpson.}$$

3.3. Stabilité des méthodes de Runge-Kutta

Les méthodes de Runge-Kutta sont des méthodes à un pas

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

avec $\Phi(t_n, y_n, h_n) = \sum_{1\le j\le q} b_j p_{n,j}$. La fonction Φ est définie de manière explicite par

$$\begin{cases} \Phi(t,y,h) = \sum_{1\le j\le q} b_j f(t+c_j h, y_j) \quad \text{avec} \\ y_i = y + h \sum_{1\le j<i} a_{ij} f(t+c_j h, y_j), \quad 1 \le i \le q. \end{cases} \tag{$*$}$$

Supposons que f soit k-lipschitzienne en y. On va montrer que Φ est alors également lipschitzienne. Soit $z \in \mathbb{R}$ et supposons $\Phi(t,z,h)$ et z_i définis à partir de z comme dans la formule $(*)$.

Lemme – *Soit* $\alpha = \max_i \Big(\sum_{1\le j\le i} |a_{ij}| \Big)$. *Alors*

$$|y_i - z_i| \le (1 + (\alpha k h) + (\alpha k h)^2 + \ldots + (\alpha k h)^{i-1}) |y - z|.$$

On démontre le lemme par récurrence sur i. Pour $i = 1$, on a $y_1 = y$, $z_1 = z$ et le résultat est évident. Supposons l'inégalité vraie pour tout $j < i$. Alors

$$|y_i - z_i| \le |y - z| + h \sum_{j<i} |a_{ij}| \cdot k \cdot \max_{j<i} |y_j - z_j|,$$

$$|y_i - z_i| \le |y - z| + \alpha k h \max_{j<i} |y_j - z_j|.$$

Par hypothèse de récurrence il vient

$$\max_{j<i} |y_j - z_j| \le (1 + \alpha k h + \ldots + (\alpha k h)^{i-2}) |y - z|,$$

et l'inégalité s'ensuit à l'ordre i. ∎

La formule $(*)$ entraîne maintenant

$$|\Phi(t,y,h) - \Phi(t,z,h)| \le \sum_{1\le j\le q} |b_j|\ k\ |y_j - z_j| \le \Lambda |y - z| \qquad \text{avec}$$

$$\Lambda = k \sum_{1\le j\le q} |b_j| (1 + (\alpha k h_{\max}) + \ldots + (\alpha k h_{\max})^{j-1}).$$

Corollaire – *Les méthodes de Runge-Kutta sont stables, avec constante de stabilité* $S = e^{\Lambda T}$.

Remarque – Dans le cas fréquent où les coefficients b_j sont ≥ 0, on a la relation

$$\Lambda \le k(1 + (\alpha k h_{\max}) + \ldots + (\alpha k h_{\max})^{q-1}).$$

Si les coefficients a_{ij} sont eux-mêmes ≥ 0, on a $\alpha = \max_i c_i$.

Lorsque $h_{\max}$ est assez petit devant $1/\alpha k$, la constante de stabilité est donc de l'ordre de grandeur de e^{kT}. Ces observations montrent que les méthodes de Runge-Kutta décrites dans les exemples 1, 2, 3 du § 3.3 possèdent une excellente stabilité (il est facile de voir que e^{kT} est la borne inférieure possible pour S, quelle que soit la méthode utilisée : considérer pour cela l'équation $y' = ky$).

3.4. ORDRE DES MÉTHODES DE RUNGE-KUTTA

Pour déterminer l'ordre, on peut appliquer le critère du § 2.4 consistant à évaluer les dérivées $\frac{\partial^l \Phi}{\partial h^l}\ (t,y,0)$: l'ordre est au moins égal à p si et seulement si cette dérivée est égale à $\frac{1}{l+1}\ f^{[l]}(t,y)$ pour $l \leq p-1$. Grâce à la formule $(*)$ du § 3.3, on obtient facilement les dérivées successives de Φ :

- $\Phi(t,y,0) = \sum_{1\leq j\leq q} b_j f(t,y) = f(t,y).$

Les méthodes de Runge-Kutta sont donc toujours d'ordre ≥ 1 (c'est-à-dire consistantes).

- $\dfrac{\partial \Phi}{\partial h}\ (t,y,h) = \sum_j b_j \Big(c_j f'_t(t+c_jh, y_j) + f'_y(t+c_jh, y_j)\ \dfrac{\partial y_j}{\partial h} \Big),$

$$\frac{\partial y_i}{\partial h} = \sum_{j<i} a_{ij} f(t+c_jh, y_j) + h \sum_{j<i} a_{ij} \Big(c_j f'_t + f'_y\ \frac{\partial y_j}{\partial h} \Big).$$

Pour $h=0$, on obtient donc

$$\frac{\partial y_i}{\partial h}\Big|_{h=0} = \Big(\sum_{j<i} a_{ij} \Big) f(t,y) = c_i f(t,y)$$
$$\frac{\partial \Phi}{\partial h}\ (t,y,0) = \sum_j b_j c_j (f'_t + f'_y f)(t,y) = \Big(\sum b_j c_j \Big) f^{[1]}(t,y).$$

D'après le § 2.4, la méthode est d'ordre ≥ 2 si et seulement si $\sum b_j c_j = \frac{1}{2}$.

- $\dfrac{\partial^2 \Phi}{\partial h^2}\ (t,y,h) = \sum_j b_j \Big(c_j^2 f''_{tt} + 2c_j f''_{ty}\ \dfrac{\partial y_j}{\partial h} + f''_{yy}\ \Big(\dfrac{\partial y_j}{\partial h}\Big)^2 + f'_y\ \dfrac{\partial^2 y_j}{\partial h^2} \Big),$

$$\frac{\partial^2 y_i}{\partial h^2} = 2 \sum_{j<i} a_{ij} \Big(c_j f'_t + f'_y\ \frac{\partial y_j}{\partial h} \Big) + h \sum_{j<i} a_{ij} \Big(c_j^2 f''_{tt} + \dots \Big).$$

Pour $h=0$, il vient

$$\frac{\partial^2 y_i}{\partial h^2}\Big|_{h=0} = 2 \sum a_{ij} c_j (f'_t + f'_y f)(t,y),$$
$$\frac{\partial^2 \Phi}{\partial h^2}\ (t,y,0) = \sum_j b_j c_j^2 (f''_{tt} + 2f''_{ty} f + f''_{yy} f^2)(t,y) + 2 \sum_{i,j} b_i a_{ij} c_j f'_y (f'_t + f'_y f)(t,y).$$

Or $f^{[2]}$ est donné par

$$\begin{aligned} f^{[2]}(t,y) &= (f^{[1]})'_t + (f^{[1]})'_y f \\ &= (f'_t + f'_y f)'_t + (f'_t + f'_y f)'_y f \\ &= f''_{tt} + f''_{ty} f + f'_y f'_t + f''_{ty} f + f''_{yy} f^2 + f'^2_y f. \\ &= (f''_{tt} + 2f''_{ty} f + f''_{yy} f^2) + f'_y (f'_t + f'_y f). \end{aligned}$$

La condition $\frac{\partial^2 \Phi}{\partial h^2}\ (t,y,0) = \frac{1}{3}\ f^{[2]}(t,y)$ se traduit en général par les conditions

$$\sum_j b_j c_j^2 = \frac{1}{3}, \qquad \sum_{i,j} b_i a_{ij} c_j = \frac{1}{6}$$

(prendre respectivement $f(t,y) = t^2$, puis $f(t,y) = t + y$ pour obtenir ces deux conditions). Un calcul analogue (pénible !) de $\frac{\partial^3 \Phi}{\partial h^3}$ conduirait au résultat suivant.

Théorème – *La méthode de Runge-Kutta définie par le tableau des coefficients c_i, a_{ij}, b_j est*

- *d'ordre ≥ 2 ssi* $\displaystyle\sum_j b_j c_j = \frac{1}{2}$.
- *d'ordre ≥ 3 ssi* $\displaystyle\sum_j b_j c_j = \frac{1}{2}\ ;\quad \sum_j b_j c_j^2 = \frac{1}{3}\ ;\quad \sum_{i,j} b_i a_{ij} c_j = \frac{1}{6}$.
- *d'ordre ≥ 4 ssi*

$$\sum_j b_j c_j = \frac{1}{2}\ ;\quad \sum_j b_j c_j^2 = \frac{1}{3}\ ;\quad \sum_j b_j c_j^3 = \frac{1}{4}$$
$$\sum_{i,j} b_i a_{ij} c_j = \frac{1}{6}\ ;\quad \sum_{i,j} b_i a_{ij} c_j^2 = \frac{1}{12}\ ;\quad \sum_{i,j} b_i c_i a_{ij} c_j = \frac{1}{8}\ ;$$
$$\sum_{i,j,k} b_i a_{ij} a_{jk} c_k = \frac{1}{12}.$$

Pour la vérification pratique, on notera que certaines des expressions précédentes sont des produits des matrices $C = \begin{pmatrix} c_1 \\ \vdots \\ c_q \end{pmatrix}$, $A = (a_{ij})$, $B = (b_1 b_2 \dots b_q)$. Ainsi

$$\sum_{i,j} b_i a_{ij} c_j = BAC, \qquad \sum_{i,j,k} b_i a_{ij} a_{jk} c_k = BA^2C.$$

Pour les exemples du § 3.2, on voit ainsi que la méthode d'Euler est d'ordre 1, et que les méthodes de l'exemple 2 sont d'ordre 2. De plus, dans une méthode avec $q = 2$, il y a *a priori* un seul coefficient a_{ij} non nul, à savoir $\alpha = a_{21}$. On a alors $c_2 = \sum_{j<2} a_{2j} = \alpha$ et la méthode est d'ordre 2 au moins ssi $\sum b_j c_j = b_2 \alpha = 1/2$, soit $b_2 = 1/2\alpha$ et $b_1 = 1 - b_2 = 1 - 1/2\alpha$. On voit donc qu'il n'y avait pas d'autres choix possibles pour une méthode d'ordre 2 avec $q = 2$.

Enfin, la méthode Runge-Kutta « classique » présentée dans l'exemple 3 est d'ordre 4 (l'ordre n'est pas ≥ 5 car $\sum b_j c_j^4 \neq 1/5$). C'est si l'on peut dire la « méthode reine » des méthodes à un pas : ordre élevé, grande stabilité (grâce à la positivité des coefficients, voir remarque finale du § 3.3). Il existe des méthodes d'ordre encore plus élevé (voir exercice 5.4), mais leur plus grande complexité les rend peut-être un peu moins praticables.

4. CONTRÔLE DU PAS

La manière la plus simple pour appliquer une méthode de résolution numérique consiste à utiliser un pas constant $h_n = h$.

La principale difficulté est alors de déterminer $h_{\max}$ de façon que l'erreur globale ne dépasse pas une certaine tolérance ε fixée à l'avance ; on ne sait pas en effet quelle sera l'évolution de la solution étudiée, de sorte qu'il est difficile de prévoir *a priori* les erreurs de consistance.

L'utilisation d'algorithmes à pas variables présente de ce point de vue deux avantages majeurs :

- l'adaptation du pas à chaque étape permet d'optimiser l'erreur commise en fonction de la tolérance prescrite ε, sous réserve qu'on dispose d'une estimation « instantanée » de l'erreur de consistance e_n.
- l'approche d'une discontinuité ou d'une singularité de l'équation différentielle ne peut se faire généralement qu'avec une réduction importante du pas. Dans cette circonstance, il convient d'arrêter l'algorithme avant de traverser la discontinuité, faute de quoi les erreurs deviennent imprévisibles. Le calcul du pas sert alors de test d'arrêt.

4.1. PRINCIPE GÉNÉRAL DU CONTRÔLE

Soit $[t_0, t_0 + T]$ l'intervalle de temps considéré. On suppose fixée une tolérance ε pour l'erreur globale

$$\max_{0\le n\le N} |y_n - z(t_n)|.$$

Supposons également qu'on dispose d'un estimation S de la constante de stabilité. En négligeant l'erreur initiale $|y_0 - z(t_0)|$ et les erreurs d'arrondi, on a alors

$$\begin{aligned}\max_{0\le n\le N} |y_n - z(t_n)| &\le S \sum_{0\le n<N} |e_n| = S \sum h_n \frac{|e_n|}{h_n} \\ &\le S\Big(\sum h_n\Big) \max\Big(\frac{|e_n|}{h_n}\Big) \le ST \ \max\Big(\frac{|e_n|}{h_n}\Big).\end{aligned}$$

Il suffit donc de choisir les pas h_n en sorte que

$$\max\Big(\frac{|e_n|}{h_n}\Big) \le \delta = \frac{\varepsilon}{ST} \ ;$$

$\frac{|e_n|}{h_n}$ représente intuitivement l'erreur de consistance par unité de temps, et c'est ce rapport qu'il s'agit de contrôler.

Il est bien entendu impossible de déterminer exactement e_n, sinon on connaîtrait du même coup la solution exacte par la formule $z(t_{n+1}) = y_{n+1} + e_n$! On va supposer néanmoins qu'on dispose d'une estimation e_n^* de e_n.

Dans la pratique, on se fixe un encadrement $[h_{\min}, h_{\max}]$ du pas ($h_{\min}$ est imposé par les limitations du temps de calcul et par l'accroissement des erreurs d'arrondi quand le pas diminue, cf. § 2.5). On essaie alors de choisir $h_n \in [h_{\min}, h_{\max}]$ de façon

que $|e_n^*|/h_n \leq \delta$, et si l'erreur est notablement inférieure on se permet d'augmenter prudemment h_n. Par exemple :

- si $\frac{1}{3}\,\delta \leq \frac{|e_n^*|}{h_n} \leq \delta$, alors $h_{n+1} := h_n$;
- si $\frac{|e_n^*|}{h_n} < \frac{1}{3}\,\delta$, alors $h_{n+1} := \min(1.25\,h_n, h_{\max})$;
- si $\frac{|e_n^*|}{h_n} > \delta$, alors $h_{n+1} := 0.8\,h_n$, avec arrêt de l'algorithme si $h_{n+1} < h_{\min}$.

Ce dernier cas peut correspondre à l'approche d'une discontinuité, l'erreur augmentant continuellement malgré la diminution du pas.

Pour l'initialisation du pas, on prend $h_0 = h_{\min}$, à moins que l'on connaisse une valeur initiale plus appropriée.

4.2. Estimation du rapport $|e_n|/h_n$

Pour estimer e_n, il n'est pas question d'utiliser les expressions analytiques des dérivées $\frac{\partial^l \Phi}{\partial h^l}$ et $f^{[l]}$, beaucoup trop coûteuses en temps de calcul. On est donc amené à rechercher des estimations *ad hoc*, qui n'utilisent si possible que des quantités déjà calculées par l'algorithme, et qui ne réclament pas de nouvelle évaluation de f.

• Méthode d'Euler

Si $p_n = f(t_n, y_n)$, alors

$$\begin{aligned} p_{n+1} - p_n &= f(t_n + h_n, y_n + h_n f(t_n, y_n)) - f(t_n, y_n) \\ &= h_n f'_t(t_n, y_n) + h_n f(t_n, y_n) f'_y(t_n, y_n) + o(h_n) \\ &= h_n f^{[1]}(t_n, y_n) + o(h_n). \end{aligned}$$

Comme $e_n = \frac{1}{2}\ h_n^2 f^{[1]}(t_n, y_n) + o(h_n^2)$ on a une approximation de e_n donnée par

$$\frac{e_n^*}{h_n} = \frac{1}{2}\ (p_{n+1} - p_n).$$

Ceci ne nécessite aucun calcul supplémentaire puisque p_{n+1} est de toute façon nécessaire pour l'étape suivante.

• Méthode de Runge-Kutta d'ordre 2

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1 - \dfrac{1}{2\alpha} & \dfrac{1}{2\alpha} \end{array}$$

D'après le § 2.4 on a ici

$$e_n = h_n^3 \Big(\frac{1}{3!}\, f^{[2]}(t_n, y_n) - \frac{1}{2!}\ \frac{\partial^2 \Phi}{\partial h^2}\,(t_n, y_n, 0) \Big) + o(h_n^3),$$

et les calculs du § 3.4 donnent

$$\begin{aligned} e_n &= h_n^3\Big(\frac{1}{6}\ f^{[2]} - \frac{\alpha}{4}\ (f''_{tt} + 2f''_{ty}f + f''_{yy}f^2)\Big)(t_n, y_n) + o(h_n^3) \\ &= h_n^3\Big(\Big(\frac{1}{6} - \frac{\alpha}{4}\Big)(f''_{tt} + 2f''_{ty}f + f''_{yy}f^2) + \frac{1}{6}\ f'_y f^{[1]}\Big) + o(h_n^3). \end{aligned}$$

On est par ailleurs amené à calculer les quantités

$$\begin{aligned} &p_{n,1} = f(t_n, y_n), \\ &p_{n,2} = f(t_n + \alpha h_n, y_n + \alpha h_n p_{n,1}), \\ &p_{n+1,1} = f\Big(t_n + h_n, y_n + h_n\Big(\Big(1 - \frac{1}{2\alpha}\Big)p_{n,1} + \frac{1}{2\alpha}\, p_{n,2}\Big)\Big). \end{aligned}$$

Des développements limités d'ordre 2 donnent après calcul :

$$\begin{aligned} p_{n,2} - p_{n,1} &= \alpha h_n f^{[1]} + \alpha^2\, \frac{h_n^2}{2}\, (f''_{tt} + 2f''_{ty}f + f''_{yy}f^2) + o(h_n^2), \\ p_{n+1,1} - p_{n,1} &= h_n f^{[1]} + h_n\, \frac{1}{2\alpha}\ (p_{n,2} - p_{n,1}) f'_y \\ &\qquad + \frac{h_n^2}{2}\ (f''_{tt} + 2f''_{ty}f + f''_{yy}f^2) + o(h_n^2), \\ p_{n+1,1} - p_{n,1} - \frac{1}{\alpha}\ (p_{n,2} - p_{n,1}) &= (1-\alpha)\ \frac{h_n^2}{2}\ (f''_{tt} + 2f''_{ty}f + f''_{yy}f^2) \\ &\qquad + \frac{h_n^2}{2}\, f'_y f^{[1]} + o(h_n^2). \end{aligned}$$

On peut approximer très grossièrement e_n/h_n par

$$\frac{e_n^*}{h_n} = \frac{1}{3}\ \Big(p_{n+1,1} - p_{n,1} - \frac{1}{\alpha}\,(p_{n,2} - p_{n,1})\Big).$$

Il n'y a pas de justification théorique sérieuse pour cela, l'idée est simplement que les développements limités se ressemblent formellement.

• Méthode de Runge-Kutta classique

Il n'y a pas ici de méthode simple permettant d'évaluer e_n, même grossièrement. On peut cependant observer que

$$e_n = h_n^5 \times (\text{dérivées d'ordre} \leq 4 \text{ de } f)\ ;$$

D'autre part, des calculs analogues à ceux ci-dessus montrent que les quantités

$$\lambda_n = p_{n,4} - 2p_{n,2} + p_{n,1} \qquad \text{et} \qquad \mu_n = p_{n,3} - p_{n,2}$$

sont de la forme

$$h_n^2 \times (\text{dérivées d'ordre} \leq 2 \text{ de } f).$$

Au lieu de comparer $\frac{|e_n|}{h_n}$ à δ, on peut essayer de comparer

$$\lambda_n^2 + \mu_n^2 \qquad \text{à} \qquad \delta$$

ou (plus rapide)

$$|\lambda_n| + |\mu_n| \qquad \text{à} \qquad \delta' = \sqrt{\delta} = \sqrt{\frac{\varepsilon}{ST}},$$

quitte à ajuster éventuellement les valeurs de δ et δ' par tatonnements.

Si l'on désire une évaluation plus précise de e_n, il est nécessaire d'utiliser des techniques plus élaborées, telles que les méthodes de Runge-Kutta emboîtées (voir par exemple le livre de Crouzeix-Mignot, chapitre 5, § 6).

5. Problèmes

5.1. On étudie la méthode numérique (M) de résolution de l'équation différentielle $y' = f(x, y)$ définie par

$$y_{n+1} = y_n + h_n \Phi(t_n, y_n, h_n),$$
$$\Phi(t, y, h) = \alpha f(t, y) + \beta f(t + \frac{h}{2}, y + \frac{h}{2}\ f(t, y)) + \gamma f(t + h, y + h f(t, h))$$

où α, β, γ sont des réels compris entre 0 et 1.

(a) Pour quelles valeurs du triplet (α, β, γ) retrouve-t-on

- la méthode d'Euler ?
- la méthode du point milieu ?
- la méthode de Heun ?

(b) Dans cette question et la suivante, on supposera que la fonction $f(t, y)$ est de classe C^∞ sur $[t_0, t_0 + \tau] \times \mathbb{R}$, et k-lipschitzienne en y. Pour quelles valeurs de (α, β, γ) la méthode proposée est-elle stable ?

(c) Quelles relations doivent satisfaire (α, β, γ) pour que la méthode soit

consistante ? convergente ? d'ordre ≥ 1 ? d'ordre ≥ 2 ?

La méthode (M) peut-elle être d'ordre supérieur ?

5.2. On considère la méthode de Runge-Kutta définie par le tableau

0	0	0	0
1/4	1	0	0
3/4	−9/20	6/5	0
1	1/9	1/3	5/9

On considère l'équation $y' = f(t,y) = t+y+1$. Résoudre cette équation et calculer $f^{[n]}(0,0)$ pour tout n. Que peut-on dire de $\partial^n\Phi/\partial h^n(0,0;0)$? Déterminer l'ordre de cette méthode.

5.3. On se propose d'obtenir une borne explicite pour l'ordre p d'une méthode de Runge-Kutta dont le nombre de points intermédiaires q est fixé.

(a) Dans l'espace vectoriel $\mathcal{P}_n$ des polynômes à coefficients réels de degré inférieur ou égal à n, on considère la forme bilinéaire

$$\langle P, Q\rangle = \int_0^1 P(x)Q(x)dx.$$

Déterminer la matrice de cette forme bilinéaire dans la base canonique $(1, x, \ldots, x^n)$. En déduire que la matrice symétrique

$$M = \begin{pmatrix} 1 & 1/2 & \ldots & 1/(n+1) \\ 1/2 & 1/3 & & 1/(n+2) \\ & & \vdots & \\ 1/(n+1) & 1/(n+2) & & 1/(2n+1) \end{pmatrix}$$

est définie positive et que $\det M > 0$.

(b) On considère pour $q \in \mathbb{N}$ le système d'équations

$$(\mathrm{S}) \quad \begin{cases} b_1 + b_2 + \ldots b_q = 1 \\ b_1c_1 + b_2c_2 + \ldots + b_qc_q = 1/2 \\ \quad \vdots \\ b_1c_1^{2q} + \ldots + b_qc_q^{2q} = 1/(2q+1) \end{cases}$$

où les b_i et c_i appartiennent à $\mathbb{R}_+^*$. Dans $\mathbb{R}^{q+1}$ on note F l'espace vectoriel engendré par les vecteurs $(1, c_j, c_j^2, \ldots, c_j^q)$, $1 \le j \le q$.

(α) Quelle est la dimension maximum de F ?

(β) Montrer que si (S) avait une solution, les vecteurs

$$V_1 = (1 \quad 1/2 \ldots 1/(q+1))$$
$$V_2 = (1/2 \quad 1/3 \ldots 1/(q+2))$$
$$V_{q+1} = (1/(q+1) \ldots 1/(2q+1))$$

appartiendraient à F. En déduire que $\det(V_1, V_2, \ldots, V_{q+1}) = 0$ puis que le système (S) est impossible.

(c) On considère une méthode de Runge-Kutta

$$\begin{array}{c|c} c_1 & \\ \vdots & A = (a_{ij}) \\ \vdots & \qquad 1 \le i,\, j \le q \\ c_q & \\ \hline & b_1 \ldots \ \ldots \ b_q \end{array}$$

associée à la méthode d'intégration (INT) $\displaystyle\int_0^1 f(x)dx \simeq \sum_{j=1}^{q} b_j f(c_j)$.

On suppose que cette méthode est d'ordre p.

(α) Montrer que (INT) est d'ordre $p-1$.

(β) En déduire que $p \leq 2q$.

5.4. On considère une méthode à un pas de la forme

$$\text{(M)} \qquad y_{n+1} = y_n + h_n \Phi(t_n, y_n, h_n)$$

qu'on suppose être d'ordre p. On se donne par ailleurs une méthode d'intégration approchée

$$\text{(I)} \qquad \int_0^1 g(u)du \simeq \sum_{1\leq j\leq q} b_j g(c_j)$$

d'ordre p au moins (il en existe pour p quelconque).

(a) On considère la méthode à un pas avec points intermédiaires $t_{n,i} = t_n + c_i h_n$ définie comme suit

$$\text{(M}') \qquad \begin{cases} \left[\begin{array}{l} t_{n,i} = t_n + c_i h_n \\ y_{n,i} = y_n + c_i h_n \Phi(t_n, y_n, c_i h_n) \\ p_{n,i} = f(t_{n,i}, y_{n,i}) \end{array}\right] & 1 \leq i \leq q \\ t_{n+1} = t_n + h_n \\ y_{n+1} = y_n + h_n \displaystyle\sum_{1\leq j\leq q} b_j p_{n,j}. \end{cases}$$

Montrer que (M$'$) est d'ordre $\geq p+1$.

(b) Grâce à un raisonnement par récurrence, montrer qu'il existe des méthodes de Runge-Kutta d'ordre p arbitrairement élevé.

CHAPITRE IX

MÉTHODES À PAS MULTIPLES

Comme dans le chapitre précédent, on s'intéresse à la résolution numérique du problème de Cauchy relatif à une équation différentielle

$$\text{(E)} \qquad y' = f(t,y), \qquad (t,y) \in [t_0, t_0+T] \times \mathbb{R}.$$

Si $(t_n)_{0\le n\le N}$ est une subdivision de $[t_0, t_0+T]$ de pas successifs $h_n = t_{n+1} - t_n$, on appelle *méthode numérique à $r+1$ pas* toute méthode numérique de la forme

$$y_{n+1} = \Psi(t_n, y_n, h_n\,;\ldots;\, t_{n-r}, y_{n-r}, h_{n-r}).$$

L'intérêt de ces méthodes vient du fait qu'on peut obtenir un ordre élevé pour une complexité de calcul nettement inférieure à celle des méthodes de Runge-Kutta. L'un des problèmes essentiels, néanmoins, est de s'assurer que la stabilité numérique reste suffisamment bonne.

1. UNE CLASSE DE MÉTHODES AVEC PAS CONSTANT

On suppose ici que le pas $h_n = h$ est constant. On s'intéresse aux méthodes à $r+1$ pas permettant un calcul récurrent des points (t_n, y_n) et des pentes $f_n = f(t_n, y_n)$ sous la forme

$$\text{(M)} \qquad \begin{cases} y_{n+1} = \displaystyle\sum_{0\le i\le r} \alpha_i y_{n-i} + h \sum_{0\le i\le r} \beta_i f_{n-i} \\ t_{n+1} = t_n + h \\ f_{n+1} = f(t_{n+1}, y_{n+1}) \end{cases}$$

où les α_i, β_i, $0 \le i \le r$ sont des constantes réelles.

Démarrage de l'algorithme – Le point initial (t_0, y_0) étant donné, l'algorithme ne peut démarrer que si les valeurs $(y_1, f_1), \ldots, (y_r, f_r)$ ont déjà été calculées. Ce calcul ne peut être fait que par une méthode à un pas pour (y_1, f_1), à au plus 2 pas pour (y_2, f_2), … au plus r pas pour (y_r, f_r). L'initialisation des r premières valeurs (y_i, f_i), $1 \le i \le r$, sera généralement faite à l'aide d'une méthode de Runge-Kutta d'ordre supérieur ou égal à celui de la méthode (M), ou à la rigueur un de moins (voir le début du § 1.2 sur ce point).

1.1. ERREUR DE CONSISTANCE ET ORDRE

La définition générale de l'erreur de consistance pour une méthode à $r+1$ pas est la suivante (on ne suppose pas nécessairement dans cette définition que le pas est constant).

Définition – *Soit z une solution exacte de l'équation* (E). *L'erreur de consistance e_n relative à z est l'écart*

$$e_n = z(t_{n+1}) - y_{n+1}, \quad r \le n < N,$$

obtenu en calculant y_{n+1} à partir des $r+1$ valeurs précédentes supposées exactes $y_n = z(t_n), \dots, y_{n-r} = z(t_{n-r})$.
La méthode est dite d'ordre p *si pour toute solution z il existe une constante C telle que*

$$|e_n| \le C h_n h_{\max}^p.$$

Déterminons e_n dans le cas de la méthode (M) ci-dessus. On a

$$z(t_{n+1}) = z(t_n + h) = \sum_{0\le k\le p} \frac{h^k}{k!}\, z^{(k)}(t_n) + O(h^{p+1})$$

dès que f est de classe C^p (z est alors de classe C^{p+1}). Par ailleurs

$$\begin{aligned}
y_{n-i} &= z(t_{n-i}) = z(t_n - ih) = \sum_{0\le k\le p} \frac{(-ih)^k}{k!}\, z^{(k)}(t_n) + O(h^{p+1}),\\
f_{n-i} &= f(t_{n-i}, z(t_{n-i})) = z'(t_{n-i}) = z'(t_n - ih)\\
&= \sum_{0\le k\le p-1} \frac{(-ih)^k}{k!}\, z^{(k+1)}(t_n) + O(h^p)\\
&= \sum_{0\le k\le p} k\,\frac{(-ih)^{k-1}}{k!}\, z^{(k)}(t_n) + O(h^p).
\end{aligned}$$

Il vient par conséquent

$$\begin{aligned}
e_n &= z(t_{n+1}) - y_{n+1} = z(t_{n+1}) - \sum_{0\le i\le r} (\alpha_i y_{n-i} + h\beta_i f_{n-i})\\
&= \sum_{0\le k\le p} \frac{h^k}{k!}\, z^{(k)}(t_n)\Big[1 - \sum_{0\le i\le r} (\alpha_i (-i)^k + k\beta_i (-i)^{k-1})\Big] + O(h^{p+1})\\
&= \sum_{0\le k\le p} \frac{h^k}{k!}\, z^{(k)}(t_n)\Big[1 - (-1)^k \sum_{0\le i\le r} i^k\alpha_i - k i^{k-1}\beta_i\Big] + O(h^{p+1}).
\end{aligned}$$

La méthode (M) est donc d'ordre $\ge p$ si et seulement si elle vérifie les conditions

$$\sum_{0\le i\le r} i^k\alpha_i - k i^{k-1}\beta_i = (-1)^k, \qquad 0 \le k \le p.$$

En particulier, elle est d'ordre ≥ 1 (ou consistante) si et seulement si

$$\begin{cases} \alpha_0 + \alpha_1 + \ldots + \alpha_r = 1 \\ \qquad \alpha_1 + \ldots + r\alpha_r - (\beta_0 + \ldots + \beta_r) = -1. \end{cases}$$

Pour qu'elle soit d'ordre $\geq p$, la p-ième condition s'explicite de la manière suivante :

$$\alpha_1 + 2^p\alpha_2 + \ldots + r^p\alpha_r - p(\beta_1 + 2^{p-1}\beta_2 + \ldots + r^{p-1}\beta_r) = (-1)^p.$$

1.2. Stabilité

On dit qu'une méthode à pas multiples est stable si une petite perturbation des valeurs initiales $y_0, \ldots, y_r$ et de petites erreurs ε_n dans le calcul récurrent des valeurs y_{n+1}, $r \leq n < N$, provoquent une erreur finale contrôlable. De façon précise :

Définition – *On dit qu'une méthode à $r+1$ pas est stable de constante de stabilité S si pour toutes suites y_n, $\widetilde{y}_n$ avec*

$$\begin{aligned} y_{n+1} &= \Psi(t_{n-i}, y_{n-i}, h_{n-i}), && r \leq n < N, \\ \widetilde{y}_{n+1} &= \Psi(t_{n-i}, \widetilde{y}_{n-i}, h_{n-i}) + \varepsilon_n, && r \leq n < N, \end{aligned}$$

alors

$$\max_{0\leq n\leq N} |\widetilde{y}_n - y_n| \leq S\Big(\max_{0\leq n\leq r} |\widetilde{y}_n - y_n| + \sum_{r\leq n<N} |\varepsilon_n| \Big).$$

En appliquant cette définition à $\widetilde{y}_n = z(t_n)$, on voit que l'erreur globale de la suite y_n par rapport à la solution exacte $z(t_n)$ admet la majoration

$$\max_{0\leq n\leq N} |y_n - z(t_n)| \leq S\Big(\max_{0\leq n\leq r} |y_n - z(t_n)| + \sum_{r\leq n<N} |e_n| \Big).$$

Si la méthode est d'ordre p avec $|e_n| \leq Ch_n h_{\max}^p$, alors on a $\displaystyle\sum_{r\leq n<N} |e_n| \leq CTh_{\max}^p$ car $\sum h_n = T$. Pour la phase d'initialisation, il convient donc de choisir une méthode conduisant à une erreur d'initialisation $\max_{0\leq n\leq r} |y_n - z(t_n)|$ de l'ordre de $h_{\max}^p$ au plus. Ceci conduit à choisir une méthode d'initialisation d'ordre $\geq p-1$. Il est toutefois préférable de choisir une méthode d'ordre $\geq p$, car l'erreur d'initialisation est alors bornée par $C'h_{\max}^{p+1}$ et donc négligeable.

Condition nécessaire de stabilité – On cherche à déterminer les conditions nécessaires à la stabilité de la méthode (M) décrite au §1.1. Pour cela, on considère l'équation différentielle la plus simple qui soit :

$$\text{(E)} \qquad y' = 0.$$

La suite y_n est alors définie par

$$y_{n+1} = \sum_{0\leq i\leq r} \alpha_i y_{n-i}, \qquad n \geq r.$$

L'ensemble des suites vérifiant cette relation de récurrence est un espace vectoriel de dimension $r+1$, car chaque suite est définie de manière unique par la donnée de $(y_0, y_1, \ldots, y_r)$. On a de manière évidente des solutions particulières

$$y_n = \lambda^n,$$

où λ est racine du polynôme caractéristique

$$\lambda^{r+1} - \alpha_0\lambda^r - \alpha_1\lambda^{r-1} - \ldots - \alpha_r = 0.$$

Soient λ_j les racines complexes de ce polynôme, et m_j les multiplicités correspondantes. On sait (par une théorie en tout point analogue à celle des équations différentielles linéaires à coefficients constants), qu'une base de l'espace vectoriel des suites considérées est formée des suites

$$n \mapsto n^q\lambda_j^n, \qquad 0 \le q < m_j.$$

Considérons maintenant la suite $y_n \equiv 0$ et la suite $\widetilde{y}_n = \varepsilon\lambda_j^n$, $0 \le n \le N$ avec $\varepsilon > 0$ petit (on a ici $\varepsilon_n = 0$, seule l'erreur d'initialisation intervient). Si la méthode (M) est stable, on doit avoir

$$|\widetilde{y}_N - y_N| = \varepsilon|\lambda_j|^N \le S \max_{0\le n\le r} \varepsilon|\lambda_j|^n,$$

ce qui équivaut à

$$|\lambda_j|^N \le S \ \max(1, |\lambda_j|^r).$$

Si l'on fait tendre h vers 0 et $N = \frac{T}{h}$ vers $+\infty$, ceci n'est possible que pour $|\lambda_j| \le 1$. Supposons maintenant que le polynôme caractéristique admette une racine λ_j de module $|\lambda_j| = 1$ et de multiplicité $m_j \ge 2$. En regardant la suite $\widetilde{y}_n = n\lambda_j^n$ on trouve

$$|\widetilde{y}_N - y_N| = \varepsilon N, \qquad \max_{0\le n\le r} |\widetilde{y}_n - y_n| = \varepsilon r$$

ce qui contredit la stabilité. On obtient donc la condition nécessaire suivante : $|\lambda_j| \le 1$ pour tout j, et si $|\lambda_j| = 1$ alors cette racine doit être simple.

Condition suffisante de stabilité* – On va voir que la condition nécessaire qui vient d'être trouvée est en fait suffisante.

Théorème – *On suppose que $f(t,y)$ est k-lipschitzienne en y. Alors la méthode* (M) *est stable si et seulement si*

$$\lambda^{r+1} - \alpha_0\lambda^r - \ldots - \alpha_r = 0$$

a toutes ses racines de module ≤ 1 et si les racines de module 1 sont simples.

Remarque. On observera que $\lambda = 1$ est toujours racine dès que la méthode est consistante, puisqu'alors $\alpha_0 + \ldots + \alpha_r = 1$.

Démonstration. Soient deux suites $y_n, \widetilde{y}_n$ telles que

$$\left.\begin{aligned} y_{n+1} &= \sum_{0\le i\le r} \alpha_i y_{n-i} + h\beta_i f_{n-i} \\ \widetilde{y}_{n+1} &= \sum_{0\le i\le r} \alpha_i \widetilde{y}_{n-i} + h\beta_i \widetilde{f}_{n-i} + \varepsilon_n \end{aligned}\right\} \quad r \le n < N$$

avec $\widetilde{f}_n = f(t_n, \widetilde{y}_n)$. Posons $\theta_n = \widetilde{y}_n - y_n$. Il vient $|\widetilde{f}_n - f_n| \le k|\theta_n|$ et

$$\theta_{n+1} - \sum_{0\le i\le r} \alpha_i \theta_{n-i} = h \sum_{0\le i\le r} \beta_i (\widetilde{f}_{n-i} - f_{n-i}) + \varepsilon_n.$$

Posons $\sigma_n = \theta_n - \alpha_0\theta_{n-1} - \ldots - \alpha_r\theta_{n-r-1}$. On a donc

$$|\sigma_{n+1}| \le kh \sum_{0\le i\le r} |\beta_i||\theta_{n-i}| + |\varepsilon_n|, \quad r \le n \le N. \qquad (*)$$

Pour exploiter cette inégalité, on cherche à majorer $|\theta_{n+1}|$ en fonction des $|\sigma_i|$. Pour cela, on observe que la relation de définition de σ_n équivaut à l'égalité formelle

$$\sum \sigma_n X^n = (\sum \theta_n X^n)(1 - \alpha_0 X - \ldots - \alpha_r X^{r+1})$$

avec la convention $\sigma_n = 0$, $\theta_n = 0$ pour $n < 0$. On a donc inversement

$$\sum \theta_n X^n = \frac{1}{1 - \alpha_0 X - \ldots - \alpha_r X^{r+1}} \sum \sigma_n X^n.$$

Considérons le développement en série en $X = 0$:

$$\frac{1}{1 - \alpha_0 X - \ldots - X^{r+1}} = \sum \gamma_n X^n.$$

Lemme – *Sous les hypothèses du théorème, les coefficients γ_n sont bornés.*

En effet, si les racines du polynôme caractéristique sont les complexes λ_j de multiplicité m_j, on a

$$1 - \alpha_0 X - \ldots - \alpha_r X^{r+1} = \prod_j (1 - \lambda_j X)^{m_j}.$$

Il existe par conséquent une décomposition en éléments simples

$$\frac{1}{1 - \alpha_0 X - \ldots - \alpha_r X^{r+1}} = \sum_{j, q\le m_j} \frac{c_{jq}}{(1 - \lambda_j X)^q}.$$

Par récurrence sur q et par dérivation, on vérifie facilement que

$$\frac{1}{(1 - \lambda_j X)^q} = \sum_{n=0}^{+\infty} \frac{(n+1)(n+2)\ldots(n+q-1)}{(q-1)!} \lambda_j^n X^n.$$

Les coefficients de cette dernière série admettent l'équivalent $n^{q-1}\lambda_j^n/(q-1)!$ et sont donc bornés si et seulement si $|\lambda_j| < 1$ ou bien $|\lambda_j| = 1$ et $q = 1$. ■

Notons $\Gamma = \sup_{n\in\mathbb{N}} |\gamma_n| < +\infty$. Comme $\gamma_0 = 1$, on a toujours $\Gamma \geq 1$. La relation

$$\sum \theta_n X^n = \sum \gamma_n X^n \sum \sigma_n X^n$$

équivaut à $\theta_n = \gamma_0\sigma_n + \gamma_1\sigma_{n-1} + \ldots + \gamma_n\sigma_0$, d'où

$$|\theta_n| \leq \Gamma(|\sigma_0| + |\sigma_1| + \ldots + |\sigma_n|). \qquad (**)$$

En combinant $(*)$ et $(**)$ il vient

$$|\theta_{n+1}| \leq \Gamma \sum_{0\leq j\leq n+1} |\sigma_j| \leq \Big[\sum_{r\leq j\leq n} |\sigma_{j+1}| + \sum_{0\leq j\leq r} |\sigma_j| \Big]$$

$$|\theta_{n+1}| \leq \Gamma \Big[\sum_{r\leq j\leq n} \Big(kh \sum_{0\leq i\leq r} |\beta_i||\theta_{j-i}| + |\varepsilon_j| \Big) + \sum_{0\leq j\leq r} |\sigma_j| \Big].$$

Or $\displaystyle\sum_{r\leq j\leq n} \sum_{0\leq i\leq r} |\beta_i||\theta_{j-i}| \leq \Big(\sum_{0\leq i\leq r} |\beta_i| \Big) \Big(\sum_{0\leq j\leq n} |\theta_j| \Big)$

et la relation de définition $\sigma_j = \theta_j - \alpha_0\theta_{j-1} - \ldots - \alpha_r\theta_{j-r-1}$ donne par ailleurs

$$\sum_{0\leq j\leq r} |\sigma_j| \leq \Big(1 + \sum_{0\leq i\leq r} |\alpha_i| \Big) \Big(|\theta_0| + \ldots + |\theta_r| \Big).$$

On obtient donc

$$|\theta_{n+1}| \leq \Gamma kh \sum_{0\leq i\leq r} |\beta_i|(|\theta_0| + \ldots + |\theta_n|)$$

$$+ \Gamma \left[\sum_{r\leq j\leq n} |\varepsilon_j| + \Big(1 + \sum_{0\leq i\leq r} |\alpha_i| \Big) \sum_{0\leq i\leq r} |\theta_i| \right].$$

Posons $\delta_n = |\theta_0| + \ldots + |\theta_n|$ et

$$\Lambda = \Gamma k \sum_{0\leq i\leq r} |\beta_i|,$$

$$\eta_n = \Gamma \left[\sum_{r\leq j\leq n} |\varepsilon_j| + \Big(1 + \sum_{0\leq i\leq r} |\alpha_i| \Big) \sum_{0\leq i\leq r} |\theta_i| \right].$$

La dernière inégalité s'écrit maintenant

$$\delta_{n+1} - \delta_n \leq \Lambda h\delta_n + \eta_n \Leftrightarrow \delta_{n+1} \leq (1 + \Lambda h)\delta_n + \eta_n$$

et le lemme de Gronwall discret (cf. VIII 2.3) donne

$$\delta_n \leq e^{\Lambda nh}\delta_0 + \sum_{0\leq j\leq n-1} e^{\Lambda(n-1-j)h}\eta_j.$$

Or pour $j \leq n-1$ on a $\eta_j \leq \eta_n$ et $\delta_0 = |\theta_0| \leq \eta_n$. On en déduit

$$\delta_n \leq \eta_n(1 + e^{\Lambda h} + \ldots + e^{\Lambda nh}) = \frac{e^{\Lambda(n+1)h} - 1}{e^{\Lambda h} - 1}\ \eta_n.$$

Cette inégalité entraîne aussitôt une majoration de θ_n :

$$\begin{aligned} |\theta_n| &= \delta_n - \delta_{n-1} \leq \Lambda h \delta_{n-1} + \eta_{n-1} \\ &= \Big(\Lambda h\ \frac{e^{\Lambda nh} - 1}{e^{\Lambda h} - 1} + 1\Big)\ \eta_{n-1}. \end{aligned}$$

Comme $e^{\Lambda h} - 1 \geq \Lambda h$, il vient $|\theta_n| \leq e^{\Lambda nh}\eta_{n-1}$, soit

$$|\theta_n| \leq \Gamma e^{\Lambda nh}\Big[\Big(1 + \sum_{0\leq i\leq r} |\alpha_i|\Big) \sum_{0\leq i\leq r} |\theta_i| + \sum_{r\leq j\leq n-1} |\varepsilon_j|\Big],$$

$$\max_{0\leq n\leq N} |\theta_n| \leq S'\Big[\Big(1 + \textstyle\sum_{0\leq n\leq r} |\alpha_i|\Big) \sum_{0\leq i\leq r} |\theta_n| + \sum_{r\leq n\leq N} |\varepsilon_n|\Big]$$

avec $S' = \Gamma e^{\Lambda T}$, c'est-à-dire

$$S' = \Gamma e^{\Gamma k T \sum |\beta_i|} \quad \text{où} \quad \Gamma = \sup\ |\gamma_n|.$$

Si l'erreur initiale $\max\limits_{0\leq n\leq r} |\theta_n|$ est négligeable (comme c'est souvent le cas) on prendra $S = S'$, sinon on peut prendre

$$S = (1 + r)\Big(1 + \sum_{0\leq i\leq r} |\alpha_i|\Big)S'.$$

Remarque – Pour une fonction $f(t, y)$ de constante de Lipschitz k et pour une durée d'intégration T fixées, la stabilité de la méthode dépend essentiellement de la grandeur de la constante $\Gamma \sum |\beta_i|$. On a donc intérêt à choisir une méthode pour laquelle cette constante soit la plus petite possible.

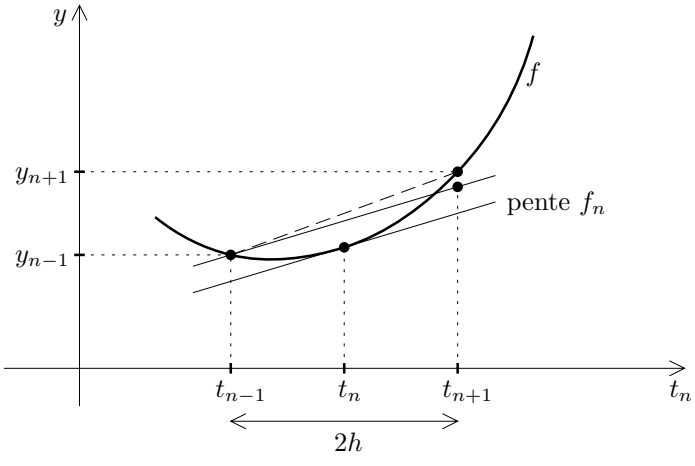
1.3. Exemples

• Méthode de Nyström

C'est la méthode à 2 pas définie par

$$y_{n+1} = y_{n-1} + 2hf_n, \quad n \geq 1.$$

On a ici $\alpha_0 = 0$, $\alpha_1 = 1$, $\beta_0 = 2$, $\beta_1 = 0$. Le principe de cette méthode est analogue à celui de la méthode du point milieu :



on approxime la pente $\frac{1}{2h}\ (y_{n+1} - y_{n-1})$ de la corde par la pente de la tangente au point milieu t_n. Les calculs du § 1.1 donnent

$$e_n = \frac{h^3}{3!}\ z^{(3)}(t_n) \cdot 2 + O(h^4) = \frac{h^3}{3}\ f^{[2]}(t_n, y_n) + O(h^4).$$

La méthode de Nyström est donc d'ordre 2. Pour la phase d'initialisation, on choisira une méthode à 1 pas d'ordre 2, par exemple la méthode du point milieu :

$$\begin{aligned} &f_0 = f(t_0, y_0)\ ; \\ &y_{1/2} = y_0 + \frac{h}{2}\ f_0\ ; \quad t_{1/2} = t_0 + \frac{h}{2}\ ; \quad f_{1/2} = f(t_{1/2}, y_{1/2})\ ; \\ &y_1 = y_0 + hf_{1/2}\ ; \quad t_1 = t_0 + h\ ; \quad f_1 = f(t_1, y_1). \end{aligned}$$

Le polynôme caractéristique est $\lambda^2 - 1$. D'après le § 1.2 la méthode est stable et

$$\frac{1}{1 - \alpha_0 X - \alpha_1 X^2} = \frac{1}{1 - X^2} = \sum X^{2n}.$$

On a donc $\Gamma = 1$, $|\beta_0| + |\beta_1| = 2$, d'où la constante de stabilité

$$S' = e^{2kt}.$$

On observe néanmoins que le polynôme caractéristique admet la racine $\lambda = -1$ située sur le cercle limite de stabilité $|\lambda| = 1$, en plus de la racine obligée $\lambda = 1$. Ceci laisse suspecter que la stabilité n'est peut-être pas très bonne. Pour mettre en évidence ce phénomène, regardons le cas de l'équation différentielle

$$y' = -y \tag{E}$$

avec donnée initiale $t_0 = 0$, $y_0 = 1$. La solution exacte est donnée par

$$z(t) = e^{-t}, \quad z(t_n) = e^{-nh}.$$

La suite y_n sera donnée par la relation de récurrence

$$y_{n+1} = y_{n-1} - 2hy_n, \qquad n \geq 1 \tag{$*$}$$

(car ici $f_n = -y_n$), avec valeurs initiales :

$$\begin{aligned} &y_0 = 1, && f_0 = -1, \\ &y_{1/2} = 1 - \frac{h}{2}, && f_{1/2} = -1 + \frac{h}{2}, \\ &y_1 = 1 - h + \frac{h^2}{2}. \end{aligned}$$

La solution générale de $(*)$ peut s'écrire

$$y_n = c_1\lambda_1^n + c_2\lambda_2^n$$

où λ_1, λ_2 sont les racines de l'équation

$$\lambda^2 + 2hk - 1 = 0,$$

à savoir $\lambda_1 = -h + \sqrt{1+h^2}$, $\lambda_2 = -h - \sqrt{1+h^2}$. Les constantes c_1, c_2 sont déterminées par

$$\begin{cases} y_0 = c_1 + c_2 = 1 \\ y_1 = c_1\lambda_1 + c_2\lambda_2 = 1 - h + \frac{h^2}{2}. \end{cases}$$

On obtient donc

$$\begin{aligned} c_1 &= \frac{1 - h + \frac{h^2}{2} - \lambda_2}{\lambda_1 - \lambda_2} = \frac{1 + \frac{h^2}{2} + \sqrt{1+h^2}}{2\sqrt{1+h^2}}, \\ c_2 &= \frac{\lambda_1 - \left(1 - h + \frac{h^2}{2}\right)}{\lambda_1 - \lambda_2} = \frac{\sqrt{1+h^2} - \left(1 + \frac{h^2}{2}\right)}{2\sqrt{1+h^2}}. \end{aligned}$$

Comme $\sqrt{1+h^2} = 1 + \frac{h^2}{2} - \frac{h^4}{8} + O(h^6)$, on voit que

$$\begin{aligned} c_1 &= 1 + O(h^4) \\ c_2 &= -\frac{1}{16}\, h^4 + O(h^6) \end{aligned}$$

Par ailleurs $\lambda_1 = 1 - h + \frac{h^2}{2} + O(h^4) = e^{-h} + O(h^3)$,

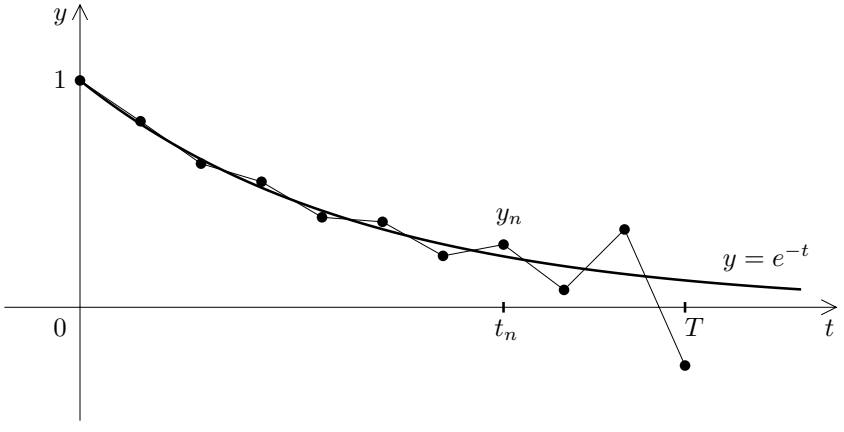
tandis que $\lambda_2 = -\left(1 + h + \frac{h^2}{2} + O(h^4)\right) = -e^h + O(h^3)$.

On voit donc que y_n est somme de deux termes

$$c_1\lambda_1^n \simeq e^{-nh}, \qquad c_2\lambda_2^n \simeq -\frac{1}{16}\, h^4(-1)^n e^{nh}.$$

Le premier de ces termes approxime bien la solution exacte, mais le second est un terme de perturbation qui diverge quand $n \to +\infty$, bien qu'il soit négligeable quand

le temps $t_n = nh$ n'est pas trop grand. Si la durée d'intégration est trop longue (plus précisément, si $h^4 e^T$ n'est plus négligeable), on va obtenir un tracé de la forme ci-dessous.



• **Méthode de Milne**

C'est la méthode à 4 pas définie par

$$y_{n+1} = y_{n-3} + h\Big(\frac{8}{3}\ f_n - \frac{4}{3}\ f_{n-1} + \frac{8}{3}\ f_{n-2}\Big).$$

On vérifie qu'elle est stable, d'ordre 4, mais comme pour la méthode de Nyström la stabilité n'est pas très bonne à cause des racines de module 1 du polynôme caractéristique $\lambda^4 - 1 = 0$.

2. Méthodes d'Adams-Bashforth

2.1. Description

On ne suppose plus ici que le pas h_n soit nécessairement constant. Si z est une solution exacte de l'équation, on écrit

$$z(t_{n+1}) = z(t_n) + \int_{t_n}^{t_{n+1}} f(t, z(t))dt.$$

Supposons que pour $0 \leq i \leq r$ on ait déjà calculé les points $z(t_{n-i})$ et les pentes $f_{n-i} = f(t_{n-i},\, z(t_{n-i}))$.

L'idée de la méthode est d'approximer la fonction $f(t, z(t))$ sur $[t_n, t_{n+1}]$ par son polynôme d'interpolation aux points $t_n, t_{n-1}, \ldots, t_{n-r}$. Considérons donc le polynôme $p_{n,r}(t)$ qui interpole les points (t_{n-i}, f_{n-i}) pour $0 \leq i \leq r$:

$$p_{n,r}(t) = \sum_{0 \leq i \leq r} f_{n-i} L_{n,i,r}(t), \quad \deg(p_{n,r}) = r,$$

où $L_{n,i,r}(t) = \prod_{\substack{0\le j\le r\\ j\neq i}} \dfrac{t-t_{n-j}}{t_{n-i}-t_{n-j}}$. On écrit maintenant :

$$\begin{aligned} z(t_{n+1}) &= z(t_n) + \int_{t_n}^{t_{n+1}} f(t,z(t))dt \\ &\simeq z(t_n) + \int_{t_n}^{t_{n+1}} p_{n,r}(t)dt \\ &= z(t_n) + h_n \sum_{0\le i\le r} b_{n,i,r} f_{n,i} \end{aligned}$$

avec

$$b_{n,i,r} = \frac{1}{h_n}\int_{t_n}^{t_{n+1}} L_{n,i,r}(t)dt.$$

L'algorithme de la méthode d'Adams-Bashforth à $r+1$ pas (en abrégé AB_{r+1}) va donc s'écrire :

$$\begin{cases} y_{n+1} = y_n + h_n \sum_{0\le i\le r} b_{n,i,r} f_{n-i}, \quad n \ge r, \\ t_{n+1} = t_n + h_n \\ f_{n+1} = f(t_{n+1}, y_{n+1}). \end{cases}$$

L'intérêt de cette méthode provient de sa relative simplicité et du fait qu'une seule évaluation de la fonction f est nécessaire à chaque étape (contrairement aux méthodes de Runge-Kutta qui en réclamaient plusieurs). Il va en résulter un gain assez important sur le temps de calcul.

Exemples

• $r = 0$: on a $p_{n,0}(t) = \text{constante} = f_n$, d'où AB_1 : $y_{n+1} = y_n + h_n f_n$. Il s'agit de la méthode d'Euler.

• $r = 1$: le polynôme $p_{n,1}$ est la fonction affine qui interpole (t_n, f_n) et (t_{n-1}, f_{n-1}), d'où les formules

$$\begin{aligned} p_{n,1}(t) &= f_n + \frac{f_n - f_{n-1}}{t_n - t_{n-1}}\,(t-t_n) \\ \int_{t_n}^{t_{n+1}} p_{n,1}(t)dt &= f_n h_n + \frac{f_n - f_{n-1}}{h_{n-1}}\,\Big[\frac{1}{2}\,(t-t_n)^2\Big]_{t_n}^{t_{n+1}} \\ &= b_n\Big(f_n + \frac{h_n}{2h_{n-1}}\,(f_n - f_{n-1})\Big). \end{aligned}$$

L'algorithme s'écrit donc

$$AB_2 \quad \begin{cases} y_{n+1} = y_n + h_n\Big(f_n + \dfrac{h_n}{2h_{n-1}}\,(f_n - f_{n-1})\Big) \\ t_{n+1} = t_n + h_n \\ f_{n+1} = f(t_n, y_n). \end{cases}$$

Dans le cas où le pas $h_n = h$ est constant, la formule de récurrence se réduit à

$$y_{n+1} = y_n + h\Big(\frac{3}{2}\ f_n - \frac{1}{2}\ f_{n-1}\Big).$$

• De manière générale, lorsque le pas est constant les coefficients $b_{n,i,r}$ ne dépendent pas de n car la méthode est invariante par translation. Pour les petites valeurs de r, les coefficients $b_{i,r}$ correspondants sont donnés par le tableau :

r	$b_{0,r}$	$b_{1,r}$	$b_{2,r}$	$b_{3,r}$	$\beta_r = \sum_i \lvert b_{i,r}\rvert$
0	1				1
1	$\frac{3}{2}$	$-\frac{1}{2}$			2
2	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$		3,66...
3	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$	6,6...

Remarque – On a toujours $\sum_{0\le i\le r} b_{n,i,r} = 1$, car pour $f_n = \ldots = f_{n-r} = 1$ on a $p_{n,r}(t) \equiv 1$, par conséquent

$$\int_{t_n}^{t_{n+1}} p_{n,r}(t)dt = h_n = h_n \sum_{0\le i\le r} b_{n,i,r} \cdot 1.$$

Comme on le verra plus loin, la quantité β_r intervient dans le calcul de la constante de stabilité S.

2.2. Erreur de consistance et ordre de la méthode AB_{r+1}

Soit z une solution exacte du problème de Cauchy. L'erreur de consistance est donnée par

$$\begin{aligned} e_n &= z(t_{n+1}) - y_{n+1} \\ &= z(t_{n+1}) - \Big(z(t_n) + \int_{t_n}^{t_{n+1}} p_{n,r}(t)dt\Big), \\ e_n &= \int_{t_n}^{t_{n+1}} \Big(z'(t) - p_{n,r}(t)\Big)dt, \end{aligned}$$

où $p_{n,r}$ est précisément le polynôme d'interpolation de la fonction $z'(t) = f(t, z(t))$ aux points t_{n-i}, $0 \le i \le r$. D'après le théorème de la moyenne, il existe un point $\theta \in]t_n, t_{n+1}[$ tel que

$$e_n = h_n\Big(z'(\theta) - p_{n,r}(\theta)\Big).$$

La formule donnant l'erreur d'interpolation (voir chapitre II, § 1.2) implique

$$z'(\theta) - p_{n,r}(\theta) = \frac{1}{(r+1)!}\ z^{(r+2)}(\xi)\pi_{n,r}(\xi),$$

où $\xi \in]t_{n-r}, t_{n+1}[$ est un point intermédiaire entre θ et les points t_{n-i}, et où

$$\pi_{n,r}(t) = \prod_{0\le i\le r} (t - t_{n-i}).$$

Si $\xi \in]t_{n-j}, t_{n-j+1}[$, $0 \le j \le r$, on a l'inégalité $|\xi - t_{n-i}| \le (1 + |j-i|)h_{\max}$, d'où

$$\begin{aligned}|\pi_{n,r}(\xi)| &\le h_{\max}^{r+1}(1+j)\dots(1+1)1(1+1)\dots(1+r-j)\\ &= h_{\max}^{r+1}(j+1)!(r-j+1)! \le h_{\max}^{r+1}(r+1)!,\end{aligned}$$

en majorant 2 par $j+2, \dots, (r-j+1)$ par $(r+1)$. On en déduit par conséquent

$$|z'(\theta) - p_{n,r}(\theta)| \le |z^{(r+2)}(\xi)|\, h_{\max}^{r+1},$$

ce qui donne la majoration cherchée de l'erreur de consistance :

$$|e_n| \le |z^{(r+2)}(\xi)| h_n h_{\max}^{r+1} \le C h_n h_{\max}^{r+1}$$

avec $C = \max\limits_{t\in[t_0,t_0+T]} |z^{(r+2)}(t)|$.

La méthode d'Adams-Bashforth à $r+1$ pas est donc d'ordre $r+1$ (le lecteur pourra vérifier à titre d'exercice que l'ordre n'est pas $\ge r+2$ en considérant le cas de la fonction $f(t,y) = t^{r+1}$).

Phase d'initialisation – De ce qui précède, il résulte qu'on choisira une méthode de Runge-Kutta d'ordre $r+1$ (ou r à la rigueur) pour initialiser les premières valeurs $y_1, \dots, y_r, f_0, f_1, \dots, f_r$.

2.3. Stabilité de la méthode AB_{r+1}

Nous allons démontrer le résultat suivant.

Théorème – *On suppose que $f(t,y)$ est k-lipschitzienne en y et que les sommes $\sum_{0\le i\le r} |b_{n,i,r}|$ sont majorées indépendamment de n par une constante β_r.*
Alors la méthode d'Adams-Bashforth à $r+1$ pas est stable, avec constante de stabilité

$$S = \exp(\beta_r k T).$$

Démonstration. Soit $\widetilde{y}_n$ la suite récurrente perturbée telle que

$$\begin{cases} \widetilde{y}_{n+1} = \widetilde{y}_n + h_n \displaystyle\sum_{0\le i\le r} b_{n,i,r}\widetilde{f}_{n-i} + \varepsilon_n, \quad r \le n < N,\\ \widetilde{f}_{n-i} = f(t_{n-i}, \widetilde{y}_{n-i}). \end{cases}$$

Posons $\theta_n = \max_{0\le i\le n} |\widetilde{y}_i - y_i|$. On a

$$|\widetilde{f}_{n-i} - f_{n-i}| \le k|\widetilde{y}_{n-i} - y_{n-i}| \le k\theta_n,$$
$$|\widetilde{y}_{n+1} - y_{n+1}| \le \theta_n + h_n \sum_{0\le i\le r} |b_{n,i,r}| \cdot k\theta_n + |\varepsilon_n|$$
$$\le (1 + \beta_r k h_n)\theta_n + |\varepsilon_n|.$$

Comme $\theta_{n+1} = \max(|\widetilde{y}_{n+1} - y_{n+1}|, \theta_n)$, on en déduit

$$\theta_{n+1} \le (1 + \beta_r k h_n)\theta_n + |\varepsilon_n|.$$

Le lemme de Gronwall implique alors

$$\theta_N \le \exp\Big(\beta_r k(t_N - t_r)\Big)\Big(\theta_r + \sum_{r\le n<N} |\varepsilon_n|\Big),$$

ce qui entraîne bien la stabilité, avec constante $S = \exp(\beta_r kT)$. D'après le tableau donné au § 2.1, on voit que la constante β_r croît assez vite quand r augmente. La stabilité devient donc de moins en moins bonne quand le nombre de pas augmente. Cette stabilité médiocre est un des inconvénients les plus sérieux des méthodes d'Adams-Bashforth lorsque r est grand. En pratique, on se limitera le plus souvent aux cas $r = 1$ ou $r = 2$.

Remarque – L'exemple AB_2 donné au § 2.1 montre que les coefficients $b_{n,i,r}$ ne sont en général bornés que si le rapport h_n/h_{n-1} de 2 pas consécutifs reste borné. Dans la pratique, il est raisonnable de supposer que

$$\frac{h_n}{h_{n-1}} \le \delta$$

avec disons $\delta \le 2$. Les formules du § 2.1 donnent dans ce cas :

$$|b_{n,i,r}| \le \max_{t\in[t_n,t_{n+1}]} |L_{n,i,r}(t)| = \prod_{1\le j\le n} \frac{t_{n+1} - t_{n-j}}{|t_{n-i} - t_{n-j}|}$$

Comme $t_{n+1} - t_{n-j} = h_{n-j} + \ldots + h_n \le (1 + \delta + \ldots + \delta^j)h_{n-j}$ et

$$h_{n-j} \le \begin{cases} |t_{n-i} - t_{n-j}| & \text{si} \quad j > i \\ \delta|t_{n-i} - t_{n-j}| & \text{si} \quad j < i \end{cases}$$

Il vient

$$|b_{n,i,r}| \le \delta^i \prod_{j\ne i}(1 + \delta + \ldots + \delta^j) \le \prod_{0\le j\le r} (1 + \delta + \ldots + \delta^j).$$

On a donc l'estimation assez grossière

$$\beta_r = \max_n \sum_{0\le i\le r} |b_{n,i,r}| \le (r+1) \prod_{0\le j\le r} (1 + \delta + \ldots + \delta^j).$$

3. Méthodes d'Adams-Moulton

3.1. Description

L'idée en est la même que celle des méthodes d'Adams-Bashforth, mais on approxime ici $f(t, z(t))$ par son polynôme d'interpolation aux points $t_{n+1}, t_n, \ldots, t_{n-r}$; le point t_{n+1} est donc pris en plus. On considère le polynôme $p^*_{n,r}(t)$ de degré $r+1$ qui interpole les points (t_{n-i}, f_{n-i}) pour $-1 \leq i \leq r$:

$$p^*_{n,r}(t) = \sum_{-1 \leq i \leq r} f_{n-i} L^*_{n,i,r}(t),$$

d'où
$$L^*_{n,i,r}(t) = \prod_{\substack{-1 \leq j \leq r \\ j \neq i}} \frac{t - t_{n,j}}{t - t_{n,i}}.$$

On obtient donc comme au §2.2

$$z(t_{n+1}) \simeq z(t_n) + h_n \sum_{-1 \leq i \leq r} b^*_{n,i,r} f_{n-i}$$

avec
$$b^*_{n,i,r} = \frac{1}{h_n} \int_{t_n}^{t_{n+1}} L^*_{n,i,r}(t) dt.$$

L'algorithme correspondant AM_{r+1} s'écrit

$$y_{n+1} - h_n b^*_{n,-1,r} f(t_{n+1}, y_{n+1}) = y_n + h_n \sum_{0 \leq i \leq r} b^*_{n,i,r} f_{n-i}.$$

On observera ici que y_{n+1} n'est pas donné explicitement en fonction des quantités y_n, f_{n-i} antérieurement calculées, mais seulement comme solution d'une équation dont la résolution n'est pas *a priori* immédiate. Pour cette raison, on dit que la méthode d'Adams-Moulton est une *méthode implicite* (la méthode d'Adams-Bashforth est dite par opposition explicite). Pour résoudre l'équation ci-dessus, on aura recours en général à une méthode itérative. Notons u_n la quantité (explicite)

$$u_n = y_n + h_n \sum_{0 \leq i \leq r} b^*_{n,i,r} f_{n-i}.$$

Le point y_{n+1} cherché est la solution x de l'équation

$$x = u_n + h_n b^*_{n,-1,r} f(t_{n+1}, x).$$

On va donc calculer la suite itérée $x_{p+1} = \varphi(x_p)$ où

$$\varphi(x) = u_n + h_n b^*_{n,-1,r} f(t_{n+1}, x).$$

Comme $\varphi'(x) = h_n b^*_{n,-1,r} f'_y(t_{n+1}, x)$, l'application φ va être contractante (avec une petite constante de Lipschitz) lorsque h_n est assez petit. Si $f(t,y)$ est k-lipschitzienne en y, il suffit que $h_n < \frac{1}{|b^*_{n,-1,r}|k}$ pour avoir convergence. La solution y_{n+1} est alors unique d'après le théorème du point fixe, et l'algorithme itératif s'écrit

$$\begin{cases} F_p = f(t_{n+1}, x_p), \\ x_{p+1} = u_n + h_n b_{n,-1,r} F_p. \end{cases}$$

On choisira une valeur initiale x_0 qui soit une approximation de y_{n+1} (la meilleure possible !), par exemple la valeur donnée par la méthode d'Adams-Bashforth :

$$x_0 = y_n + h_n \sum_{0 \le i \le r} b_{n,i,r} f_{n-i}.$$

On arrête l'itération pour $|x_{p+1} - x_p| \le 10^{-10}$ (par exemple) et on prend

$$\begin{cases} y_{n+1} = & \text{dernière valeur } x_{p+1} \text{ calculée} \\ f_{n+1} = & f(t_{n+1}, y_{n+1}) \quad (\text{ou par économie} = F_p) \end{cases}$$

Exemples

• $r = 0$: le polynôme $p^*_{n,0}$ est le polynôme de degré 1 qui interpole (t_{n+1}, f_{n+1}) et (t_n, f_n), soit

$$p^*_{n,0}(t) = f_n + \frac{f_{n+1} - f_n}{h_n}\ (t - t_n)\ ;$$
$$\int_{t_n}^{t_{n+1}} p^*_{n,0}(t)dt = h_n\Big(\frac{1}{2}\ f_{n+1} + \frac{1}{2}\ f_n\Big).$$

On obtient ainsi la méthode dite des trapèzes (ou méthode de Crank-Nicolson) :

$$y_{n+1} = y_n + h_n\Big(\frac{1}{2}\ f_{n+1} + \frac{1}{2}\ f_n\Big)$$

ou encore

$$y_{n+1} - \frac{1}{2}\ h_n f(t_{n+1}, y_{n+1}) = y_n + \frac{1}{2}\ h_n f_n.$$

• $r = 1$: le polynôme $p^*_{n,1}$ interpole les points (t_{n+1}, f_{n+1}), (t_n, f_n), (t_{n-1}, f_{n-1}), d'où les formules

$$p^*_{n,1}(t) = f_{n+1}\,\frac{(t-t_n)(t-t_{n-1})}{h_n(h_n+h_{n-1})} - f_n\,\frac{(t-t_{n-1})(t-t_{n-1})}{h_n h_{n-1}} + f_{n-1}\,\frac{(t-t_n)(t-t_{n+1})}{h_{n-1}(h_n+h_{n-1})},$$

$$y_{n+1} = y_n \int_{t_n}^{t_{n+1}} p^*_{n,1}(t)dt,$$

$$y_{n+1} = y_n + h_n\left[\frac{2h_n + 3h_{n-1}}{6(h_n+h_{n-1})}\,f_{n+1} + \frac{3h_{n-1}+h_n}{6h_{n-1}}\,f_n - \frac{h_n^2}{6h_{n-1}(h_n+h_{n-1})}\,f_{n-1}\right].$$

Dans le cas où le pas $h_n = h$ est constant, cette formule se réduit à

$$y_{n+1} = y_n + h\Big(\frac{5}{12}\ f_{n+1} + \frac{8}{12}\ f_n - \frac{1}{12}\ f_{n-1}\Big).$$

• De manière générale, les coefficients $b^*_{n,i,r}$ sont des nombres $b^*_{i,r}$ indépendants de n lorsque le pas est constant. On a la table suivante :

r	$b^*_{-1,\mathrm{r}}$	$b^*_{0,r}$	$b^*_{1,r}$	$b^*_{2,r}$	$b^*_{3,r}$	$\beta^*_r = \sum_i \lvert b^*_{i,r}\rvert$	β_{r+1}
0	$\frac{1}{2}$	$\frac{1}{2}$				1	2
1	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$			1,16...	3,66...
2	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$		1,41...	6,66...
3	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$	1,78...	12,64...

Les coefficients $b^*_{n,i,r}$ vérifient toujours $\displaystyle\sum_{-1\le i\le r} b^*_{n,i,r} = 1$.

3.2. Erreur de consistance et ordre de la méthode AM_{r+1}

Soit z une solution exacte du problème de Cauchy. Sous l'hypothèse $y_{n-i} = z(t_{n-i})$, $0 \le i \le n$, on a

$$\begin{aligned}
e_n &= z(t_{n+1}) - y_{n+1}\\
&= z(t_{n+1}) - \Big[z(t_n) + h_n \sum_{0\le i\le r} b^*_{n,i,r} f(t_{n-i}, z(t_{n-i})) + h_n b^*_{n,-1,r} f(t_{n+1}, y_{n+1})\Big]\\
&= z(t_{n+1}) - \Big[z(t_n) + h_n \sum_{-1\le i\le r} b^*_{n,i,r} f(t_{n-i}, z(t_{n-i}))\Big]\\
&\qquad + h_n b^*_{n,-1,r}[f(t_{n+1}, z(t_{n+1})) - f(t_{n+1}, y_{n+1})],\\
e_n &= \int_{t_n}^{t_{n+1}} (z'(t) - p^*_{n,r}(t))dt + h_n b^*_{n,-1,r}\Big[f(t_{n+1}, z(t_{n+1})) - f(t_{n+1}, y_{n+1})\Big].
\end{aligned}$$

Supposons que $f(t,y)$ soit lipschitzienne de rapport k en y. Alors il vient

$$\begin{aligned}
|e_n| &\le \Big|\int_{t_n}^{t_{n+1}} (z'(t) - p^*_{n,r}(t))dt\Big| + h_n b^*_{n,-1,r} k |e_n|,\\
|e_n| &\le \frac{1}{1 - h_n b^*_{n,-1,r} k}\ \Big|\int_{t_n}^{t_{n+1}} (z'(t) - p^*_{n,r}(t))dt\Big|.
\end{aligned}$$

Quand le pas h_n est suffisamment petit, on a donc

$$|e_n| = \Big| \int_{t_n}^{t_{n+1}} (z'(t) - p^*_{n,r}(t))dt \Big| (1 + O(h_n))$$

comme dans la méthode d'Adams-Bashforth. Par ailleurs la formule de la moyenne donne

$$\int_{t_n}^{t_{n+1}} (z'(t) - p^*_{n,r}(t))dt = h_n(z'(\theta) - p^*_{n,r}(\theta)), \quad \theta \in]t_n, t_{n+1}[,$$
$$z'(\theta) - p^*_{n,r}(\theta) = \frac{1}{(r+2)!}\, z^{(r+3)}(\xi)\pi^*_{n,r}(\xi), \quad \xi \in]t_{n,r}, t_{n+1}[.$$

où $\pi^*_{n,r}(t) = \prod_{-1 \le i \le r} (t - t_{n-i})$. Il résulte du § 2.2 que

$$\begin{aligned} |\pi^*_{n,r}(\xi)| &= |\xi - t_{n+1}|\, |\pi_{n,r}(\xi)| \\ &\le (r+1)h_{\max} \cdot (r+1)!\, h^{r+1}_{\max} \le (r+2)!\, h^{r+2}_{\max}, \end{aligned}$$
$$\Big| \int_{t_n}^{t_{n+1}} (z'(t) - p^*_{n,r}(t))dt \Big| \le |z^{(r+3)}(\xi)| h_n h^{r+2}_{\max},$$

par conséquent l'erreur de consistance admet la majoration

$$|e_n| \le C h_n h^{r+2}_{\max}(1 + O(h_n)),$$

avec $C = \max_{t \in [t_0, t_0+T]} |z^{(r+3)}|$.

La méthode AM_{r+1} est donc d'ordre $r+2$. On initialisera les r premières valeurs $y_1, \ldots, y_r$ à l'aide d'une méthode de Runge-Kutta d'ordre $r+2$ (ou à la rigueur $r+1$).

3.3.* STABILITÉ DE LA MÉTHODE AM_{r+1}

On suppose que les rapports h_n/h_{n-1} restent bornés, de sorte que les quantités

$$\beta^*_r = \max_n \sum_{\le i \le r} |b^*_{n,i,r}|, \qquad \gamma^*_r = \max_n |b^*_{n,-1,r}|$$

sont contrôlées. Supposons également que $f(t,y)$ soit k-lipschitzienne en y. La méthode de résolution itérative pour y_{n+1} fonctionne dès que $h_n < \frac{1}{|b^*_{n,-1,r}|k}$, en particulier dès que

$$h_{\max} < \frac{1}{\gamma^*_r k},$$

ce que nous supposons désormais. Soit $\widetilde{y}_n$ une suite perturbée telle que

$$\begin{cases} \widetilde{y}_{n+1} = \widetilde{y}_n + h_n\Big(b^*_{n,-1,r}\widetilde{f}_{n+1} + \sum_{0 \le i \le r} b^*_{n,i,r}\widetilde{f}_{n-i}\Big) + \varepsilon_n \\ \widetilde{f}_{n-i} = f(t_{n-i}, y_{n-i}), \qquad r \le n < N, \end{cases}$$

et posons $\theta_n = \max\limits_{0\le i\le n} |\widetilde{y}_i - y_i|$. Comme on a $\theta_{n+1} = \max\left(|\widetilde{y}_{n+1} - y_{n+1}|, \theta_n\right)$, il vient

$$\begin{aligned}
\theta_{n+1} &\le \theta_n + kh_n\Big(|b^*_{n,-1,r}|\theta_{n+1} + \sum_{0\le i\le r} |b^*_{n,i,r}|\theta_n\Big) + |\varepsilon_n|,\\
\theta_{n+1}(1 - |b^*_{n,-1,r}|kh_n) &\le \theta_n\Big(1 + \sum_{0\le i\le r} |b^*_{n,i,r}|kh_n\Big) + |\varepsilon_n|\\
&\le \Big(1 + \sum_{0\le i\le r} |b^*_{n,i,r}|kh_n\Big)(\theta_n + |\varepsilon_n|).
\end{aligned}$$

Or $1 - |b^*_{n,-1,r}|kh_n \ge 1 - \gamma^*_r kh_{\max} > 0$, par suite

$$\begin{aligned}
\theta_{n+1} &\le \frac{1 + \sum\limits_{0\le i\le r} |b^*_{n,i,r}|kh_n}{1 - |b^*_{n,-1,r}|kh_n}\ (\theta_n + |\varepsilon_n|),\\
\theta_{n+1} &\le \left(1 + \frac{\sum\limits_{-1\le i\le r} |b^*_{n,i,r}|kh_n}{1 - |b^*_{n,-1,r}|kh_n}\right)(\theta_n + |\varepsilon_n|),\\
\theta_{n+1} &\le (1 + \Lambda h_n)(\theta_n + |\varepsilon_n|),
\end{aligned}$$

avec $\Lambda = \beta^*_r k/(1 - \gamma^*_r kh_{\max})$. Un raisonnement analogue à la démonstration du lemme de Gronwall discret donne par récurrence sur n :

$$\theta_n \le e^{\Lambda(t_n - t_r)}\Big(\theta_r + \sum_{r\le i\le n} |\varepsilon_i|\Big),$$

d'où la constante de stabilité

$$S = e^{\Lambda T} = \exp\left(\frac{\beta^*_r kT}{1 - \gamma^*_r kh_{\max}}\right)$$

Lorsque $h_{\max}$ est assez petit devant $1/\gamma^*_r k$, on a donc sensiblement

$$S \simeq \exp\left(\beta^*_r kT\right).$$

Le tableau du § 3.1 montre qu'à ordre $r+2$ égal, la méthode AM_{r+1} est beaucoup plus stable que AB_{r+2}. Il n'en reste pas moins que malgré cet avantage important, la méthode d'Adams-Moulton est d'utilisation délicate à cause de son caractère implicite. Les méthodes de prédiction-correction que nous allons décrire permettent d'obtenir une stabilité équivalente, tout en fournissant un schéma explicite de résolution.

4. MÉTHODES DE PRÉDICTION-CORRECTION

4.1. PRINCIPE GÉNÉRAL

On se donne une méthode dite de *prédiction* (ou *prédicteur*), fournissant (explicitement) une première valeur approchée py_{n+1} du point y_{n+1} à atteindre :

$$\begin{aligned} py_{n+1} &= \text{prédiction de } y_{n+1}, \\ pf_{n+1} &= f(t_{n+1}, py_{n+1}) = \text{prédiction de } f_{n+1}. \end{aligned}$$

En substituant la valeur pf_{n+1} ainsi trouvée à f_{n+1} dans la formule d'Adams-Moulton, on obtient alors une nouvelle valeur *corrigée* y_{n+1} qui est retenue en vue des calculs ultérieurs.

De façon précise, une méthode PECE (prédiction, évaluation, correction, évaluation) à $r+1$ pas va s'écrire de la manière suivante : $y_{n-r}, f_{n-r}, \ldots, y_n, f_n$ étant déjà calculés, on pose

$$\begin{cases} \text{Prédiction :} & py_{n+1} = \ldots \ \text{(à partir des } y_{n-i}, f_{n-i}, 0 \le i \le r) \\ & t_{n+1} = t_n + h_n \\ \text{Evaluation :} & pf_{n+1} = f(t_{n+1}, py_{n+1}) \\ \text{Correction :} & y_{n+1} = y_n + h_n\Big(b^*_{n,-1,r} pf_{n+1} + \sum_{0\le i\le r} b^*_{n,i,r} f_{n-i}\Big) \\ \text{Evaluation :} & f_{n+1} = f(t_{n+1}, y_{n+1}) \end{cases}$$

Nous avons utilisé ici la méthode d'Adams-Moulton comme *correcteur*, mais cela pourrait être *a priori* n'importe quelle autre méthode implicite.

Ici encore, le démarrage de l'algorithme PECE nécessite le calcul préalable des points $y_1, \ldots, y_r$ et des pentes $f_0, \ldots, f_r$ à l'aide d'une méthode à 1 pas. On peut estimer que le coût en temps de calcul d'une méthode PECE est environ le double de celui d'une méthode d'Adams-Bashforth d'ordre égal (mais on verra que la stabilité est beaucoup meilleure). Ce temps de calcul est en général inférieur à celui des méthodes de Runge-Kutta sophistiquées (d'ordre ≥ 3).

4.2. ERREUR DE CONSISTANCE DANS PECE

Soit z une solution exacte du problème de Cauchy. L'erreur de consistance est

$$e_n = z(t_{n+1}) - y_{n+1} \quad \text{avec} \quad y_{n-i} = z(t_{n-i}), \quad 0 \le i \le r.$$

Soit y^*_{n+1} la valeur qui serait obtenue à l'aide du seul correcteur (Adams-Moulton), de sorte que

$$\begin{cases} y^*_{n+1} = y_n + h_n\Big(b^*_{n,-1,r} f^*_{n+1} + \displaystyle\sum_{0\le i\le r} b^*_{n,i,r} f_{n,i}\Big) \\ f^*_{n+1} = f(t_{n+1}, y^*_{n+1}). \end{cases}$$

L'erreur de consistance correspondante est

$$e^*_n = z(t_{n+1}) - y^*_{n+1}.$$

Le prédicteur introduit lui aussi une erreur de consistance

$$pe_n = z(t_{n+1}) - py_{n+1}.$$

Écrivons

$$e_n = (z(t_{n+1}) - y^*_{n+1}) + (y^*_{n+1} - y_{n+1})$$
$$e_n = e^*_n + (y^*_{n+1} - y_{n+1}).$$

On a par ailleurs

$$y^*_{n+1} - y_{n+1} = h_n b^*_{n,-1,r}(f^*_{n+1} - pf_{n+1}).$$

Si $f(t,y)$ est k-lipschitzienne en y, on en déduit

$$|y^*_{n+1} - y_{n+1}| \le h_n |b^*_{n,-1,r}|\, k\, |y^*_{n+1} - py_{n+1}|$$

et $y^*_{n+1} - py_{n+1} = (z(t_{n+1}) - py_{n+1}) - (z(t_{n+1}) - y^*_{n+1})$, d'où

$$|y^*_{n+1} - py_{n+1}| \le |pe_n| + |e^*_n|.$$

Il en résulte finalement

$$|y^*_{n+1} - y_{n+1}| \le |b^*_{n,-1,r}|\, kh_n(|pe_n| + |e^*_n|)$$
$$|e_n| \le |e^*_n| + |y^*_{n+1} - y_{n+1}|$$
$$|e_n| \le \big(1 + |b^*_{n,-1,r}|kh_n\big)|e^*_n| + |b^*_{n,-1,r}|\, kh_n\, |pe_n|.$$

On voit que l'influence du prédicteur est nettement moindre que celle du correcteur puisque son erreur de consistance est en facteur d'un terme $O(h_n)$. Le *correcteur AM_{r+1} étant d'ordre $r+2$* (c'est-à-dire $|e^*_n| \le Ch_n h^{r+2}_{\max}$), on voit qu'il convient de *choisir un prédicteur d'ordre $r+1$*. Les contributions de $|e^*_n|$ et $|pe_n|$ dans $|e_n|$ seront alors toutes deux $\le Ch_n h^{r+2}_{\max}$, *l'ordre global de PECE est donc $r+2$ dans ce cas.*

4.3. Exemples

• Prédicteur : Euler (ordre 1), Correcteur : AM_1 (ordre 2).

$$\begin{cases} P : & py_{n+1} = y_n + h_n f_n \\ E : & pf_{n+1} = f(t_{n+1}, py_{n+1}) \\ C : & y_{n+1} = y_n + h_n\big(\frac{1}{2}\, pf_{n+1} + \frac{1}{2}\, f_n\big) \\ E : & f_{n+1} = f(t_{n+1}, y_{n+1}) \end{cases}$$

Cet algorithme coïncide avec la méthode de Heun, qui n'est autre que la méthode de Runge-Kutta définie par

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

- **Prédicteur : Nyström (ordre 2) avec pas constant $h_n = h$, Correcteur : AM_2 (ordre 3).**

$$\begin{cases} P & : \quad py_{n+1} = y_{n-1} + 2hf_n \\ E & : \quad pf_{n+1} = f(t_{n+1}, py_{n+1}) \\ C & : \quad y_{n+1} = y_n + h\Big(\frac{5}{12}\, pf_{n+1} + \frac{8}{12}\, f_n - \frac{1}{12}\, f_{n-1}\Big) \\ E & : \quad f_{n+1} = f(t_{n+1}, y_{n+1}) \end{cases}$$

- **Prédicteur : AB_{r+1} (ordre $r+1$), Correcteur : AM_{r+1} (ordre $r+2$).**

$$\begin{cases} P & : \quad py_{n+1} = y_n + h_n \displaystyle\sum_{0\le i\le r} b_{n,i,r} f_{n-i} \\ E & : \quad pf_{n+1} = f(t_{n+1}, py_{n+1}) \\ C & : \quad y_{n+1} = y_n + h_n\Big(b^*_{n,-1,r} pf_{n+1} + \displaystyle\sum_{0\le i\le r} b^*_{n,i,r} f_{n-i}\Big) \\ E & : \quad f_{n+1} = f(t_{n+1}, y_{n+1}). \end{cases}$$

Exercice – *Vérifier que ce dernier algorithme PECE équivaut à la méthode d'Adams-Moulton dans laquelle l'algorithme itératif est arrêté à la première étape, soit $y_{n+1} = x_1$, calculé à partir de la valeur x_0 fourni par la méthode d'Adams-Bashforth.*

4.4. Stabilité de la méthode PECE

Supposons que le prédicteur soit de la forme

$$py_n = \sum_{0\le i\le r} \alpha_{n,i} y_{n-i} + h_n \sum_{0\le i\le r} \beta_{n,i} f_{n-i}$$

et notons

$$A = \max_n \sum_i |\alpha_{n,i}|, \qquad B = \max_n \sum_i |\beta_{n,i}|.$$

Soit $\widetilde{y}_n$ une suite perturbée telle que

$$\begin{cases} p\widetilde{y}_{n+1} = \displaystyle\sum_{0\le i\le r} \alpha_{n,i} \widetilde{y}_{n-i} + h_n \sum_{0\le i\le r} \beta_{n,i} \widetilde{f}_{n-i} \\ \widetilde{y}_{n+1} = \widetilde{y}_n + h_n\Big(b^*_{n,-1,r} p\widetilde{f}_{n+1} + \displaystyle\sum_{0\le i\le r} b^*_{n,i,r} \widetilde{f}_{n-i}\Big) + \varepsilon_n. \end{cases}$$

Remarque – Dans la réalité il s'introduit également une erreur d'arrondi $p\varepsilon_n$ au niveau de la prédiction, mais pour notre calcul il sera plus simple de comptabiliser cette erreur dans ε_n (ceci n'est visiblement pas restrictif).

Posons $\theta_n = \max_{0\le i\le n} |\widetilde{y}_i - y_i|$ et $p\theta_n = |p\widetilde{y}_n - py_n|$. Comme $f(t,y)$ est supposée k-lipschitzienne en y, il vient :

$$\begin{cases} p\theta_{n+1} \le A\theta_n + h_n Bk\theta_n \\ \theta_{n+1} \le \theta_n + kh_n\Big(|b^*_{n,-1,r}|p\theta_{n+1} + \sum_{0\le i\le r} |b^*_{n,i,r}|\theta_n\Big) + |\varepsilon_n|. \end{cases}$$

En substituant $p\theta_{n+1} \le \theta_n + \theta_n(A-1+Bkh_n)$ dans la deuxième ligne il vient

$$\theta_{n+1} \le \theta_n \left(1 + \Big[\sum_{-1\le i\le r} |b^*_{n,i,r}| + |b^*_{n,-1,r}|(A-1+Bkh_n)\Big]kh_n\right) + |\varepsilon_n|,$$

$$\theta_{n+1} \le \theta_n(1+\Lambda h_n) + |\varepsilon_n|,$$

avec $\Lambda = (\beta^*_r + \gamma^*_r(A-1+Bkh_{\max}))k$. Le lemme de Gronwall montre donc que la méthode PECE est stable, avec constante de stabilité

$$S = e^{\Lambda T} = \exp\Big((\beta^*_r + \gamma^*_r(A-1+Bkh_{\max})kT\Big).$$

Si $h_{\max}$ est petit, on va avoir

$$S \simeq \exp\Big((\beta^*_r + \gamma^*_r(A-1))kT\Big).$$

On voit que la stabilité du prédicteur n'a pas d'incidence sur la stabilité de la méthode PECE, seule la valeur de la constante A peut influer sur cette stabilité ; on pourrait en théorie utiliser un prédicteur instable ! La consistance du prédicteur implique $\sum_i \alpha_{n,i} = 1$. Si les coefficients $\alpha_{n,i}$ sont ≥ 0, alors on a $A = 1$ (c'est le cas des méthodes de Nyström, Milne, ou AB_{r+1}), par conséquent la constante de stabilité

$$S \simeq \exp(\beta^*_r kT)$$

sera peu différente de celle de la méthode d'Adams-Moulton seule. On obtient donc des méthodes assez stables, d'un coût modéré en temps de calcul et d'ordre aussi élevé que l'on veut. Par comparaison avec les méthodes de Runge-Kutta, elles sont un peu plus rapides mais un peu moins stables à ordre égal.

4.5. Méthodes PEC

Comme leur nom l'indique, il s'agit de méthodes de prédiction-correction dans lesquelles la dernière étape d'évaluation est omise (en vue bien sûr de gagner du temps). Ceci signifie que les pentes corrigées f_{n+1} ne sont pas calculées, il faudra donc se contenter de faire intervenir les pentes prédites pf_{n-i}. On obtient alors l'algorithme suivant :

$$\begin{cases} \text{Prédiction} : & py_{n+1} = \sum_{0\le i\le r} \alpha_{n,i} y_{n-i} + h_n \sum_{0\le i\le r} \beta_{n,i} pf_{n-i} \\ & t_{n+1} = t_n + h_n \\ \text{Evaluation} : & pf_{n+1} = f(t_{n+1}, py_{n+1}) \\ \text{Correction} : & y_{n+1} = y_n + h_n \sum_{-1\le i\le r} b^*_{n,i,r} pf_{n-i}. \end{cases}$$

Le démarrage de l'algorithme nécessite le calcul préalable des quantités $y_1, \ldots, y_r$, $pf_0, \ldots, pf_r$.

Erreur de consistance* – L'algorithme PEC n'entre pas tout à fait dans le cadre général des méthodes que nous avons considérées jusqu'à présent. Il convient de redéfinir e_n comme suit. Si z est une solution exacte, on pose

$$e_n = z(t_{n+1}) - y_{n+1}$$

où y_{n+1} est calculée à partir des valeurs antérieures $py_{n-i} = y_{n-i} = z(t_{n-i})$, $0 \leq i \leq r$. Avec cette définition, il est facile de voir que l'erreur de consistance est identique à celle de la méthode PECE, d'où

$$|e_n| \leq \Big(1 + |b^*_{n,-1,r}|kh_n\Big)|e^*_n| + |b^*_{n,-1,r}|kh_n|pe_n|.$$

Le correcteur étant d'ordre $r+2$, on choisira ici encore le prédicteur d'ordre $r+1$.

Stabilité de la méthode PEC* – Avec les notations et hypothèses du § 4.4, considérons une suite perturbée $\widetilde{y}_n$ telle que

$$\widetilde{y}_{n+1} = \widetilde{y}_n + h_n \sum_{-1 \leq i \leq r} b^*_{n,i,r} p\widetilde{f}_{n-i} + \varepsilon_n$$

et posons $\theta_n = \max\limits_{0 \leq i \leq n} |\widetilde{y}_i - y_i|$, $p\theta_n = \max\limits_{0 \leq i \leq n} |p\widetilde{y}_i - py_i|$. Il vient :

$$\begin{cases} p\theta_{n+1} \leq A\theta_n + Bkh_n p\theta_n \\ \theta_{n+1} \leq \theta_n + \beta^*_r kh_n p\theta_{n+1} + |\varepsilon_n|. \end{cases}$$

La première ligne entraîne

$$p\theta_{n+1} \leq A\theta_n + Bkh_{\max} p\theta_{n+1},$$

d'où $p\theta_{n+1} \leq \frac{A}{1 - Bkh_{\max}}\ \theta_n$ si $Bkh_{\max} < 1$. En substituant dans la deuxième ligne il vient :

$$\theta_{n+1} \leq \left(1 + \frac{\beta^*_r Ak}{1 - Bkh_{\max}}\ h_n\right)\theta_n + |\varepsilon_n|.$$

Le lemme de Gronwall donne la constante de stabilité

$$S = \exp\left(\frac{\beta^*_r AkT}{1 - Bkh_{\max}}\right).$$

Si $h_{\max}$ est assez petit, on aura

$$S \simeq \exp\left(\beta^*_r AkT\right).$$

Ceci est un peu moins bon que dans le cas de la méthode PECE, car $\gamma^*_r < \beta^*_r$. Néanmoins, pour $A = 1$ la constante de stabilité est la même :

$$S \simeq \exp\left(\beta^*_r kT\right),$$

c'est-à-dire précisément la constante de stabilité de la méthode d'Adams-Moulton seule. Par rapport à PECE, on économise donc un peu de temps de calcul, mais on perd un peu en stabilité et en précision.

5. Problèmes

5.1. On considère le problème de Cauchy $y'(t) = f(t, y(t))$, $y(t_0) = y_0$, où $f : [t_0, t_0+T] \times \mathbb{R} \to \mathbb{R}$ est une fonction de classe C^5. Pour résoudre numériquement ce problème, on se donne un entier $N \geq 2$ et on considère la subdivision $t_n = t_0+nh$, $0 \leq n \leq N$, de pas constant $h = \frac{T}{N}$.

On étudie les méthodes à 2 pas de la forme

$$\text{(M)} \qquad y_{n+1} = \alpha y_{n-1} + \alpha' y_n + h(\beta f_{n-1} + \beta' f_n + \beta'' f_{n+1}),$$

avec $f_n = f(t_n, y_n)$. La méthode est donc explicite si $\beta'' = 0$ et implicite si $\beta'' \neq 0$.

(a) Soit g une fonction de classe C^5 au voisinage de 0.

Calculer un développement limité à l'ordre 4 en $h = 0$ de la quantité

$$\Delta(h) = g(h) - [\alpha g(-h) + \alpha' g(0) + h(\beta g'(-h) + \beta' g'(0) + \beta'' g'(h))].$$

Écrire la condition nécessaire et suffisante pour que la méthode (M) soit d'ordre ≥ 1 (respectivement ≥ 2, ≥ 3, ≥ 4). Montrer que la seule méthode (M) qui soit d'ordre ≥ 4 est

$$\text{(M}^4\text{)} \qquad y_{n+1} = y_{n-1} + h\Big(\frac{1}{3}\, f_{n-1} + \frac{4}{3}\, f_n + \frac{1}{3}\, f_{n+1}\Big).$$

Quelle interprétation peut-on donner de cette méthode ?

NB : Dans les questions qui suivent, on supposera sans le repréciser chaque fois que les méthodes (M) étudiées sont d'ordre 1 au moins.

(b) On cherche à tester la stabilité de la méthode (M) en considérant l'équation différentielle triviale $y' = 0$.

Les réels y_0 et $y_1 = y_0 + \varepsilon$ étant donnés *a priori*, exprimer y_n en fonction de y_0, ε, α, n. En déduire qu'une condition nécessaire pour que la méthode (M) soit stable est que $-1 < \alpha \leq 1$.

(c) On se propose ici de montrer inversement que la méthode (M) est stable, si $0 \leq \alpha \leq 1$ et si h est assez petit. On suppose que pour tout $t \in [t_0, t_0 + T]$ la fonction $y \mapsto f(t, y)$ est k-lipschitzienne. Soient deux suite (y_n) et (z_n) telles que pour $n \geq 1$ on ait

$$\begin{aligned} y_{n+1} &= \alpha y_{n-1} + \alpha' y_n + h(\beta f(t_{n-1}, y_{n-1}) + \beta' f(t_n, y_n) + \beta'' f(t_{n+1}, y_{n+1})), \\ z_{n+1} &= \alpha z_{n-1} + \alpha' z_n + h(\beta f(t_{n-1}, z_{n-1}) + \beta' f(t_n, z_n) + \beta'' f(t_{n+1}, z_{n+1})) + \varepsilon_n. \end{aligned}$$

On pose

$$\theta_n = \max_{0 \leq i \leq n} |z_i - y_i|.$$

(α) Majorer $|z_{n+1} - y_{n+1}|$ en fonction de θ_n, θ_{n+1} et ε_n.

En déduire que si $|\beta''|kh < 1$ on a

$$\theta_{n+1} \leq \left(1 + \frac{(|\beta| + |\beta'| + |\beta''|)kh}{1 - |\beta''|kh}\right) (\theta_n + |\varepsilon_n|).$$

(β) En déduire l'existence d'une constante de stabilité $S(h)$, qui reste bornée quand h tend vers 0, telle que

$$\theta_N \leq S(h)\Big(\theta_1 + \sum_{k=1}^{N-1} |\varepsilon_n|\Big).$$

(e) Déterminer en fonction de α les méthodes (M) qui sont d'ordre 3 ; on notera celles-ci (M^3_α). A quoi correspond (M^3_0) ? Existe-t-il une méthode (M^3_α) qui soit explicite et stable ?

Montrer qu'il existe une unique méthode $(\mathrm{M}^3_{\alpha_1})$ pour laquelle $\beta = 0$.

(f) (α) Expliciter l'algorithme PECE dont le prédicteur est la méthode de Nyström et dont le correcteur est la méthode $(\mathrm{M}^3_{\alpha_1})$.

L'initialisation sera faite au moyen de la méthode de Runge-Kutta d'ordre 4 usuelle.

(β) Écrire un programme informatique mettant en œuvre l'algorithme précédent dans le cas de la fonction $f(t, y) = \sin(ty - y^2)$.

L'utilisateur fournit la donnée initiale (t_0, y_0), le pas h, et le nombre N d'itérations. L'ordinateur affichera alors les valeurs (t_n, y_n) successives pour $0 \leq n \leq N$.

5.2. L'objet de ce problème est d'étudier les méthodes d'Adams-Bashforth et d'Adams-Moulton avec pas constant.

(a) Montrer que la méthode d'Adams-Bashforth à $(r + 1)$ pas, de pas constant h, s'écrit

$$y_{n+1} = y_n + h \sum_{i=0}^{r} b_{i,r} f(t_{n-i}, y_{n-i})$$

avec

$$b_{i,r} = (-1)^i \int_0^1 \frac{s(s+1)\dots\widehat{(s+i)}\dots(s+r)}{i!(r-i)!}, \quad 0 \leq i \leq r.$$

(b) On pose

$$\gamma_r = \int_0^1 \frac{s(s+1)\dots(s+r-1)}{r!}\, ds.$$

Démontrer les formules

$$b_{i,r} - b_{i,r-1} = (-1)^i C^i_r \gamma_r, \quad 0 \leq i \leq r-1,$$
$$b_{r,r} = (-1)^r \gamma_r.$$

(c) Montrer que pour $|t| < 1$ on a

$$\int_0^1 (1-t)^{-s} ds = \sum_{r=0}^{+\infty} \gamma_r t^r.$$

En déduire la valeur de l'expression $\log(1-t) \sum_{r=0}^{+\infty} \gamma_r t^r$, puis la valeur des sommes

$$\frac{\gamma_0}{r+1} + \frac{\gamma_1}{r} + \ldots + \frac{\gamma_{r-1}}{2} + \gamma_r.$$

(d) Écrire un programme informatique mettant en œuvre la méthode d'Adams-Bashforth à un nombre arbitraire de pas. On utilisera les formules de récurrence ci-dessus pour évaluer γ_r et $b_{i,r}$. L'initialisation sera faite au moyen de la méthode de Runge-Kutta d'ordre 4.

(e) Démontrer des formules analogues à celles de (a), (b), (c) pour la méthode d'Adams-Moulton.

5.3. Soit à résoudre numériquement un problème de Cauchy

$$y' = f(t,y), \quad y(t_0) = y_0$$

où f est de classe C^2 sur $[t_0, t_0+T] \times \mathbb{R}$. On se donne une subdivision

$$t_0 < t_1 < \ldots < t_N = t_0 + T$$

de l'intervalle, et on étudie la méthode numérique suivante : si y_n est la valeur approchée de la solution au temps t_n et si $f_n = f(t_n, y_n)$ on pose

$$\text{(M)} \qquad y_{n+1} = y_{n-1} + \int_{t_{n-1}}^{t_{n+1}} p_n(t)dt$$

où p_n est le polynôme d'interpolation des pentes f_n, f_{n-1} aux temps t_n et t_{n-1}. On note $h_n = t_{n+1} - t_n$.

(a) Calculer explicitement y_{n+1}. Quelle méthode obtient-on lorsque le pas $h_n = h$ est constant ?

(b) Soit z une solution exacte de l'équation différentielle.

Déterminer un équivalent de l'erreur de consistance e_n attachée à la méthode (M). Quel est l'ordre de cette méthode ? Comment procéderiez-vous pour la phase d'initialisation ?

(c) Soit une suite $\widetilde{y}_n$ vérifiant la relation de récurrence

$$\widetilde{y}_{n+1} = \widetilde{y}_{n-1} + \int_{t_{n-1}}^{t_{n+1}} \widetilde{p}_n(t)dt + \varepsilon_n$$

où $\widetilde{p}_n$ interpole les valeurs $\widetilde{f}_n = f(t_n, \widetilde{y}_n)$ et $\widetilde{y}_{n-1}$. La quantité ε_n désigne l'erreur commise à l'étape n. On suppose que $f(t,y)$ est k-lipschizienne en y et on note $\theta_n = \max\limits_{0\le i\le n} |\widetilde{y}_i - y_i|$.

(α) Montrer que

$$\theta_{n+1} \le \left(1 + kh_n\left(1 + \max\left\{\frac{h_n}{h_{n-1}}, \frac{h_{n-1}}{h_n}\right\}\right)\right)\theta_n + \varepsilon_n.$$

(β) On suppose que le rapport de 2 pas consécutifs est majoré par une constante δ (avec disons $1 \le \delta \le 2$). Étudier la stabilité de la méthode (M).

5.4. Dans cet exercice, on se propose d'étudier une méthode de prédiction-correction de type PEPEC pour la résolution d'un problème de Cauchy

$$\begin{cases} y' &= f(t,y), \qquad t \in [t_0, t_0+T] \\ y(t_0) &= y_0. \end{cases}$$

La fonction $f(t,y)$ est supposée de classe C^4 sur $[t_0, t_0+T]\times\mathbb{R}$ et lipschitzienne de rapport k en y. Le pas est choisi constant : $h = \frac{T}{N}$, $N \in \mathbb{N}^*$.

(a) L'objet de cette question est de décrire la méthode de correction.

(α) Si z est une solution exacte du problème de Cauchy, on écrit

$$z(t_{n+1}) = z(t_n) + \int_{t_n}^{t_{n+1}} f(t, z(t))dt$$

et on approxime l'intégrale par la méthode de Simpson élémentaire. Montrer que l'algorithme correspondant s'écrit

$$\text{(C)} \qquad y_{n+1} = y_n + h(\alpha f_n + \beta f_{n+\frac{1}{2}} + \gamma f_{n+1})$$

avec des coefficients α, β, γ que l'on précisera.

(β) On suppose que les pentes f_n, $f_{n+\frac{1}{2}}$, f_{n+1} sont les dérivées exactes de z aux points t_n, $t_n+\frac{h}{2}$, t_n+h. Déterminer un équivalent de l'erreur de consistance

$$e_n^* = z(t_{n+1}) - y_{n+1}$$

en fonction de h et de la dérivée $z^{(5)}(t_n)$. Quel est l'ordre de la méthode (C) ?

(b) Pour pouvoir exploiter la formule (C), il est nécessaire de prédire des valeurs approchées $py_{n+\frac{1}{2}}$ et py_{n+1} des points $y_{n+\frac{1}{2}}$ et y_{n+1}, ainsi que les pentes correspondantes

$$pf_{n+\frac{1}{2}} = f(t_n + \frac{h}{2}, py_{n+\frac{1}{2}}), \quad pf_{n+1} = f(t_{n+1}, py_{n+1}).$$

Pour cela, on utilise comme prédicteur (P) la méthode d'Adams-Bashforth à 3 pas, de pas constant $\frac{h}{2}$, faisant intervenir les pentes prédites antérieures pf_n, $pf_{n-\frac{1}{2}}$, pf_{n-1}. La méthode (P) est utilisée une première fois pour évaluer $py_{n+\frac{1}{2}}$, puis une deuxième fois pour évaluer py_{n+1} à patir de $py_{n+\frac{1}{2}}$.

(α) Expliciter complètement l'algorithme PEPEC ainsi obtenu.

(β) Écrire en langage informatique la boucle centrale correspondant à l'itération de l'algorithme PEPEC (on précisera la signification des variables utilisées).

(γ) L'erreur de consistance de la méthode PEPEC est définie par $e_n = z(t_{n+1}) - y_{n+1}$ où l'on suppose que les points y_n et py_{n-i}, $i \in \left\{0, \frac{1}{2}, 1\right\}$ sont exacts. Si pe_n et $pe_{n+\frac{1}{2}}$ désignent les erreurs de consistance du prédicteur sur les intervalles de temps $\left[t_n, t_n + \frac{h}{2}\right]$ et $\left[t_n + \frac{h}{2}, t_{n+1}\right]$, montrer que

$$|e_n| \leq |e_n^*| + \frac{4}{6}\, kh|pe_n| + \frac{1}{6}\, kh\Big(|pe_{n+\frac{1}{2}}| + |pe_n| + \frac{23}{24}\, kh|pe_n|\Big).$$

Le choix du prédicteur est-il justifié ? Quel est l'ordre de la méthode PEPEC ? Comment procéderiez-vous pour l'initialisation de l'algorithme ?

(c) On se propose ici de déterminer une constante de stabilité de l'algorithme PEPEC. Soit $\widetilde{y}_n$ une suite perturbée telle que la formule de correction soit entachée d'une erreur ε_n, toute l'erreur étant comptabilisée au niveau de (C) :

$$\widetilde{y}_{n+1} = \widetilde{y}_n + h(\alpha\, p\widetilde{f}_n + \beta\, p\widetilde{f}_{n+\frac{1}{2}} + \gamma\, p\widetilde{f}_{n+1}) + \varepsilon_n.$$

Pour $i, n \in \mathbb{N}$ on pose

$$\theta_n = \max_{0\leq i\leq n} |\widetilde{y}_i - y_i|,$$
$$p\theta_n = \max_{0\leq i\leq n} |p\widetilde{y}_i - py_i|, \quad p\theta_{n+\frac{1}{2}} = \max_{0\leq i\leq n} |p\widetilde{y}_{i+\frac{1}{2}} - py_{i+\frac{1}{2}}|.$$

(α) Majorer $p\theta_{n+\frac{1}{2}}$ en fonction de θ_n, $p\theta_n$ et $p\theta_{n-\frac{1}{2}}$ puis $p\theta_{n+1}$ en fonction de $p\theta_{n+\frac{1}{2}}$ et $p\theta_n$. En déduire successivement (pour h assez petit) :

$$p\theta_{n+1} \leq \frac{1 + \frac{7}{6}\, kh}{1 - \frac{4}{6}\, kh}\, p\theta_{n+\frac{1}{2}} \leq \frac{1}{1 - \frac{11}{6}\, kh}\, p\theta_{n+\frac{1}{2}},$$
$$p\theta_{n+\frac{1}{2}} \leq \theta_n + kh\, \frac{\frac{11}{6}\, p\theta_{n-\frac{1}{2}}}{1 - \frac{11}{6}\, kh},$$
$$p\theta_{n+\frac{1}{2}} \leq \frac{1 - \frac{11}{6}\, kh}{1 - \frac{11}{3}\, kh}\, \theta_n.$$

En déduire une majoration de θ_{n+1} en fonction de θ_n, $|\varepsilon_n|$ et une estimation de S.

(β) Comparer la stabilité de PEPEC à la stabilité de la méthode PECE de même ordre ayant pour prédicteur Adams-Bashforth et pour correcteur Adams-Moulton.

CHAPITRE X

STABILITÉ DES SOLUTIONS ET POINTS SINGULIERS D'UN CHAMP DE VECTEURS

On se propose ici d'étudier le comportement des solutions d'une équation différentielle et des lignes intégrales d'un champ de vecteurs lorsque le temps t tend vers l'infini. On s'intéresse essentiellement au cas des équations linéaires ou « voisines » de telles équations. Dans ce cas, le comportement des solutions est gouverné par le signe de la partie réelle des valeurs propres de la matrice associée à la partie linéaire de l'équation : une solution est dite stable si les solutions associées à des valeurs voisines de la donnée initiale restent proches de la solution considérée jusqu'à l'infini. Cette notion de stabilité (dite aussi stabilité au sens de Lyapunov) ne devra pas être confondue avec la notion de stabilité d'une méthode numérique, qui concerne la stabilité de l'algorithme sur un intervalle de temps fixé. On étudie finalement les différentes configurations possibles des lignes intégrales au voisinages des points singuliers non dégénérés d'un champ de vecteurs plan.

1. STABILITÉ DES SOLUTIONS

1.1. DÉFINITIONS

On considère le problème de Cauchy associé à une équation différentielle

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

avec condition initiale $y(t_0) = z_0$. On suppose que la solution de ce problème existe sur $[t_0, +\infty[$.

Définition – *Soit $y(t, z)$ la solution maximale de (E) tel que $y(t_0, z) = z$. On dira que la solution $y(t, z_0)$ est* stable *s'il existe une boule $\overline{B}(z_0, r)$ et une constante $C \geq 0$ telles que*

(i) *Pour tout $z \in \overline{B}(z_0, r)$, $t \mapsto y(t, z)$ est définie sur $[t_0, +\infty[$;*

(ii) *Pour tous $z \in \overline{B}(z_0, r)$ et $t \geq t_0$ on a*

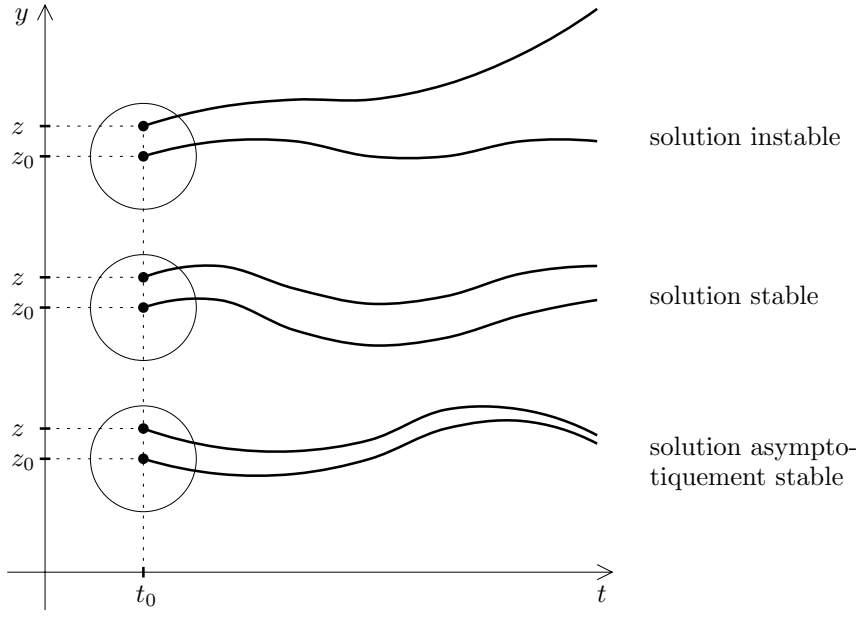
$$\|y(t, z) - y(t, z_0)\| \leq C\|z - z_0\|.$$

La solution $y(t, z_0)$ *est dite* asymptotiquement stable *si elle est stable et si la condition* (ii') *plus forte que* (ii) *est satisfaite :*

(ii') *Il existe une boule* $\overline{B}(z_0, r)$ *et une fonction* $\gamma : [t_0, +\infty[\to \mathbb{R}_+$ *continue avec* $\lim_{t \to +\infty} \gamma(t) = 0$ *telles que pour tous* $z \in \overline{B}(z_0, r)$ *et* $t \geq t_0$ *on ait*

$$\|y(t, z) - y(t, z_0)\| \leq \gamma(t)\|z - z_0\|.$$

La signification géométrique de ces notions de stabilité est illustrée par le schéma suivant.



1.2. Cas d'un système linéaire à coefficients constants

Nous étudierons d'abord le cas le plus simple, à savoir le cas d'un système linéaire sans second membre

$$\text{(E)} \qquad Y' = AY, \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}, \qquad A = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mm} \end{pmatrix}$$

avec $y_j, a_{ij} \in \mathbb{C}$; le cas réel peut bien entendu être vu comme un cas particulier du cas complexe. La solution du problème de Cauchy de condition initiale $Y(t_0) = Z$

est donnée par $Y(t,Z) = e^{(t-t_0)A} \cdot Z$. On a donc

$$Y(t,Z) - Y(t,Z_0) = e^{(t-t_0)A} \cdot (Z - Z_0)$$

et la stabilité est liée au comportement de $e^{(t-t_0)A}$ quand t tend vers $+\infty$, dont la norme $|||e^{(t-t_0)A}|||$ doit rester bornée. Distinguons quelques cas.

• $m = 1$, $A = (a)$. On a alors

$$|e^{(t-t_0)a}| = e^{(t-t_0)\,\mathrm{Re}(a)}.$$

Les solutions sont stables si et seulement si cette quantité reste bornée quand t tend vers $+\infty$, c'est-à-dire si $\mathrm{Re}(a) \leq 0$. De même, les solutions sont asymptotiquement stables si et seulement si $\mathrm{Re}(a) < 0$, et on peut alors prendre

$$\gamma(t) = e^{(t-t_0)\,\mathrm{Re}(a)} \quad \xrightarrow[t \to +\infty]{} \quad 0.$$

• m quelconque. Si A est diagonalisable, on se ramène après un changement linéaire de coordonnées à

$$\widetilde{A} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix}$$

où $\lambda_1, \ldots, \lambda_m$ désignent les valeurs prores de A. Le système se ramène aux équations indépendantes $y_j' = \lambda_j y_j$ et admet pour solution

$$y_j(t,Z) = z_j e^{\lambda_j (t-t_0)}, \quad 1 \leq j \leq m.$$

Les solutions sont donc stables si et seulement si $\mathrm{Re}(\lambda_j) \leq 0$ pour tout j et asymptotiquement stables si et seulement si $\mathrm{Re}(\lambda_j) < 0$ pour tout j.

Si A n'est pas diagonalisable, il suffit de regarder ce qui se passe pour chaque bloc d'une triangulation de A. Supposons donc

$$A = \begin{pmatrix} \lambda & & * \\ & \ddots & \\ 0 & & \lambda \end{pmatrix} = \lambda I + N$$

où N est une matrice nilpotente (triangulaire supérieure) non nulle. Il vient alors

$$\begin{aligned} e^{(t-t_0)A} &= e^{(t-t_0)\lambda I} \cdot e^{(t-t_0)N} \\ &= e^{\lambda(t-t_0)} \sum_{k=0}^{m-1} \frac{(t-t_0)^k}{k!}\, N^k, \end{aligned}$$

donc les coefficients de $e^{(t-t_0)A}$ sont des produits de $e^{\lambda(t-t_0)}$ par des polynômes de degré $\leq m-1$ non tous constants (car $N \neq 0$, donc le degré est au moins 1). Si $\mathrm{Re}(\lambda) < 0$, les coefficients tendent vers 0, et si $\mathrm{Re}(\lambda) > 0$ leur module tend vers $+\infty$ car la croissance de l'exponentielle l'emporte sur celle des polynômes. Si

$\mathrm{Re}(\lambda) = 0$, on a $|e^{\lambda(t-t_0)}| = 1$ et par suite $e^{(t-t_0)A}$ est non bornée. On voit donc que les solutions sont asymptotiquement stables si et seulement si $\mathrm{Re}(\lambda) < 0$ et sinon elle sont instables. En résumé, on peut énoncer :

Théorème – *Soient $\lambda_1, \ldots, \lambda_m$ les valeurs propres complexes de la matrice A. Alors les solutions du système linéaire $Y' = AY$ sont*

- *asymptotiquement stables si et seulement si* $\mathrm{Re}(\lambda_j) < 0$ *pour tout* $j = 1, \ldots, m$.
- *stables si et seulement si pour tout j, ou bien* $\mathrm{Re}(\lambda_j) < 0$, *ou bien* $\mathrm{Re}(\lambda_j) = 0$ *et le bloc correspondant est diagonalisable.*

1.3. Petite perturbation d'un système linéaire

On considère dans $\mathbb{K}^m = \mathbb{R}^m$ ou $\mathbb{C}^m$ un système de la forme

$$Y' = AY + g(t, Y) \tag{E}$$

où $g : [t_0, +\infty[\times \mathbb{K}^m \to \mathbb{K}^m$ est une fonction continue. On se propose de montrer que si la partie linéaire est asymptotiquement stable et si la « perturbation » g est suffisamment petite, en un sens à préciser, alors les solutions de (E) sont encore asymptotiquement stables.

Théorème – *On suppose que les valeurs propres complexes λ_j de A sont de partie réelle* $\mathrm{Re}\,\lambda_j < 0$.

(a) *S'il existe une fonction $k : [t_0, +\infty[\to \mathbb{R}_+$ continue telle que* $\lim_{t\to+\infty} k(t) = 0$ *et*

$$\forall t \in [t_0, +\infty[, \quad \forall Y_1, Y_2 \in \mathbb{K}^m, \quad \|g(t, Y_1) - g(t, Y_2)\| \leq k(t)\|Y_1 - Y_2\|,$$

alors toute solution de (E) est asymptotiquement stable.

(b) *Si $g(t, 0) = 0$ et s'il existe $r_0 > 0$ et une fonction continue $k : [0, r_0] \to \mathbb{R}_+$ telle que* $\lim_{r\to 0} k(r) = 0$ *et*

$$\forall t \in [t_0, +\infty[, \quad \forall Y_1, Y_2 \in \overline{B}(0, r), \quad \|g(t, Y_1) - g(t, Y_2)\| \leq k(r)\|Y_1 - Y_2\|$$

pour $r \leq r_0$, alors il existe une boule $\overline{B}(0, r_1) \subset \overline{B}(0, r_0)$ telle que toute solution $Y(t, Z_0)$ de valeur initiale $Z_0 \in \overline{B}(0, r_1)$ soit asymptotiquement stable.

Démonstration.* Si $\mathbb{K} = \mathbb{R}$, on peut toujours étendre le système à $\mathbb{C}^m$ en posant par exemple $\widetilde{g}(t, Y) = g(t, \mathrm{Re}(Y))$ pour $Y \in \mathbb{C}^m$. On se placera donc dans $\mathbb{C}^m$. Il existe alors une base $(e_1, \ldots, e_m)$ dans laquelle A se met sous forme triangulaire

$$A = \begin{pmatrix} \lambda_1 & a_{12} & \ldots & a_{1m} \\ & \lambda_2 & & \vdots \\ \vdots & & \ddots & a_{m-1m} \\ & \ldots & \ldots & \\ 0 & \ldots & \ldots & \lambda_m \end{pmatrix}$$

Posons $\widetilde{e}_j = \varepsilon^j e_j$ avec $\varepsilon > 0$ petit. Il vient

$$\begin{aligned} A\widetilde{e}_j &= \varepsilon^j(a_{1j}e_1 + \ldots + a_{j-1j}e_{j-1} + \lambda_j e_j) \\ &= \varepsilon^{j-1}a_{1j}\widetilde{e}_1 + \ldots + \varepsilon a_{j-1j}\widetilde{e}_{j-1} + \lambda_j\widetilde{e}_j \end{aligned}$$

de sorte que dans la base $(\widetilde{e}_j)$ les coefficients non diagonaux peuvent être rendus arbitrairement petits. On supposera donc qu'on a $|a_{ij}| \leq \varepsilon$, et on pourra choisir ε aussi petit qu'on veut. Considérons deux solutions $Y(t, Z)$ et $Y(t, Z_0)$:

$$\begin{aligned} Y'(t, Z) &= AY(t, Z) + g(t, Y(t, Z)), \\ Y'(t, Z_0) &= AY(t, Z_0) + g(t, Y(t, Z_0)) \end{aligned}$$

et cherchons à évaluer la différence $\Delta(t) = Y(t, Z) - Y(t, Z_0)$ en distinguant les deux cas (a) et (b).

(a) Observons dans ce cas que $f(t, Y) = AY + g(t, Y)$ est lipschitzienne en Y avec constante de Lipschitz $|||A||| + k(t)$. Le critère V 3.4 montre donc déjà que toutes les solutions sont globalement définies sur $[t_0, +\infty[$. Nous avons

$$\begin{aligned} &\Delta'(t) = A\Delta(t) + g(t, Y(t, Z)) - g(t, Y(t, Z_0)), \\ &\|g(t, Y(t, Z)) - g(t, Y(t, Z_0))\| \leq k(t)\|\Delta(t)\| \end{aligned}$$

Notons $(\delta_j(t))_{1\leq j\leq m}$ les composantes de $\Delta(t)$ et

$$\rho(t) = \|\Delta(t)\|^2 = \sum_{j=1}^{m} \delta_j(t)\overline{\delta_j(t)}.$$

Par différentiation on obtient

$$\begin{aligned} \rho'(t) &= \sum_{j=1}^{m} \delta_j'(t)\overline{\delta_j(t)} + \delta_j(t)\overline{\delta_j'(t)} = 2\,\mathrm{Re}\sum_{i=1}^{m} \delta_j'(t)\overline{\delta_j(t)}, \\ &= 2\,\mathrm{Re}({}^t\overline{\Delta(t)}\Delta'(t)) \\ &= 2\,\mathrm{Re}({}^t\overline{\Delta(t)}A\Delta(t)) + 2\,\mathrm{Re}\Big({}^t\overline{\Delta(t)}(g(t, Y(t, Z)) - g(t, Y(t, Z_0)))\Big). \end{aligned}$$

La deuxième partie réelle est majorée par

$$2\|\Delta(t)\| \cdot \|g(t, Y(t, Z)) - g(t, Y(t, Z_0))\| \leq 2k(t)\|\Delta(t)\|^2 = 2k(t)\rho(t).$$

On a par ailleurs

$${}^t\overline{\Delta(t)}A\Delta(t) = \sum_{j=1}^{m} \lambda_j|\delta_j(t)|^2 + \sum_{i<j} a_{ij}\overline{\delta_i(t)}\delta_j(t),$$

de sorte que

$$\mathrm{Re}({}^t\overline{\Delta(t)}A\Delta(t)) \leq \sum_{j=1}^{m}(\mathrm{Re}\,\lambda_j)|\delta_j(t)|^2 + \Big(\sum_{i<j}|a_{ij}|\Big)\|\Delta(t)\|^2.$$

Comme $\mathrm{Re}(\lambda_j) < 0$ par hypothèse et $|a_{ij}| \leq \varepsilon$, il y a un choix de ε tel que

$$\mathrm{Re}({}^t\overline{\Delta(t)}A\Delta(t)) \leq -\alpha \sum_{j=1}^{m} |\delta_j(t)|^2 = -\alpha\rho(t)$$

avec $\alpha > 0$. On obtient alors

$$\begin{aligned}
\rho'(t) &\leq -2\alpha\rho(t) + 2k(t)\rho(t),\\
\frac{\rho'(t)}{\rho(t)} &\leq -2\alpha + 2k(t),\\
\ln \frac{\rho(t)}{\rho(t_0)} &\leq -2\int_{t_0}^{t} (\alpha - k(u))du,\\
\rho(t) &\leq \|Z - Z_0\|^2 \exp\Big(-2\int_{t_0}^{t} (\alpha - k(u))du\Big)
\end{aligned}$$

car $\rho(t_0) = \|Z - Z_0\|^2$. On notera que $\rho(t) = \|\Delta(t)\|^2$ ne peut s'annuler que si les deux solutions coïncident identiquement. En prenant la racine carrée, on obtient

$$\|Y(t, Z) - Y(t, Z_0)\| \leq \gamma(t)\|Z - Z_0\|$$

avec

$$\gamma(t) = \exp\Big(-\int_{t_0}^{t} (\alpha - k(u))du\Big).$$

Comme $\lim\limits_{u\to+\infty}(\alpha - k(u)) = \alpha > 0$, l'intégrale diverge vers $+\infty$ et $\lim\limits_{t\to+\infty} \gamma(t) = 0$. Les solutions sont donc bien asymptotiquement stables.

(b) Ce cas est un peu plus délicat car on ne sait pas *a priori* si toutes les solutions sont globales ; elles ne le seront d'ailleurs pas en général si $Z_0 \notin \overline{B}(0, r_0)$, vu que les hypothèses ne concernent que ce qui se passe pour $Y \in \overline{B}(0, r_0)$. Comme $g(t, 0) = 0$, on a toutefois la solution globale $Y(t) = 0$, c'est-à-dire que $Y(t, 0) = 0$ pour $t \in [t_0, +\infty[$. De plus on a

$$\|g(t, Y(t, Z)) - g(t, Y(t, Z_0))\| \leq k(r)\|\Delta(t)\|,$$

à condition de supposer que $t \mapsto Y(t, Z)$ et $t \mapsto Y(t, Z_0)$ prennent toutes leurs valeurs dans $\overline{B}(0, r) \subset \overline{B}(0, r_0)$. Sous cette hypothèse, les mêmes calculs que précédemment donnent

$$\begin{aligned}
\rho(t) &\leq \|Z - Z_0\|^2 \ \exp\Big(-2\int_{t_0}^{t} (\alpha - k(r)du\Big),\\
\|Y(t, Z) - Y(t, Z_0)\| &\leq \exp\Big(-(t - t_0)(\alpha - k(r)\Big)\|Z - Z_0\|, \qquad (*)
\end{aligned}$$

et en particulier pour $Z_0 = 0$:

$$\|Y(t, Z)\| \leq \exp\Big(-(t - t_0)(\alpha - kr)\Big)\|Z\|.$$

Comme $\lim_{r\to 0} k(r) = 0$, on peut choisir $r_1 < r_0$ tel que $k(r_1) < \alpha$, c'est-à-dire $\alpha - k(r_1) > 0$. L'inégalité précédente montre alors que pour $Z \in B(0, r_1)$ la solution maximale $Y(t, Z)$ contenue dans la boule ouverte $B(0, r_1)$ vérifie les inégalités $\|Y(t, Z)\| \leq \|Z\| < r_1$. Cette solution maximale est nécessairement définie globalement sur $[t_0, +\infty[$. Sinon l'intervalle maximal serait un intervalle borné $[t_0, t_1[$, nécessairement ouvert à droite d'après les résultats de V 2.4. Comme la dérivée de $t \mapsto Y(t, Z)$ est majorée par

$$\|AY + g(t, Y)\| \leq (|||A||| + k(r_1))\|Y\| \leq M$$

avec $M = (|||A||| + k(r_1))r_1$, la fonction $Y(t, Z)$ vérifierait le critère de Cauchy

$$\lim_{t,t'\to t_1 - 0} \|Y(t, Z) - Y(t', Z)\| = 0.$$

Elle aurait donc une limite $Y_1 = \lim_{t\to t_1 - 0} Y(t, Z)$ avec $\|Y_1\| \leq \|Z\| < r_1$ et se prolongerait sur un voisinage à droite de t_1 en une solution entièrement contenue dans $B(0, r_1)$, contradiction. Quitte à diminuer encore un peu r_1, on voit que toute solution $Y(t, Z)$ avec $Z \in \overline{B}(0, r_1)$ est globale et entièrement contenue dans $\overline{B}(0, r_1)$. Par conséquent $(*)$ est satisfaite pour tous $t \in [t_0, +\infty[$ et $Z, Z_0 \in \overline{B}(0, r_1)$ avec la constante $\alpha - k(r_1) > 0$, ce qui démontre le théorème. ∎

2. Points singuliers d'un champ de vecteurs

2.1. Position du problème

On suppose donné un champ de vecteurs de classe C^1 dans un ouvert $\Omega \subset \mathbb{R}^2$, c'est-à-dire une application

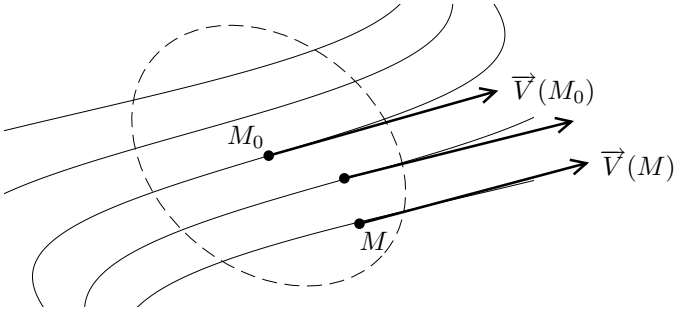
$$\Omega \to \mathbb{R}^2, \qquad M = \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \overrightarrow{V}(M) = \begin{pmatrix} f(x, y) \\ g(x, y) \end{pmatrix}$$

où f, g sont de classe C^1 sur Ω. On considère le système différentiel associé

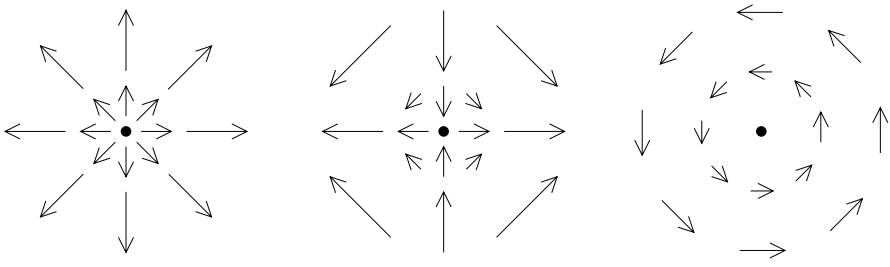
$$\frac{\overrightarrow{dM}}{dt} = \overrightarrow{V}(M) \Longleftrightarrow \begin{cases} x'(t) = f(x(t), y(t)) \\ y'(t) = g(x(t), y(t)) \end{cases}.$$

Grâce au théorème de Cauchy-Lipschitz, on sait que par tout point il passe une courbe intégrale unique. Un problème géométrique intéressant est de décrire l'allure de la famille des courbes intégrales passant au voisinage d'un point M_0 donné.

Premier cas : $\overrightarrow{V}(M_0) \neq \overrightarrow{0}$. Dans ce cas, l'angle entre $\overrightarrow{V}(M)$ et $\overrightarrow{V}(M_0)$ tend vers 0 quand M tend vers 0. Par conséquent, les tangentes aux lignes intégrables sont sensiblement parallèles les unes aux autres dans un petit voisinage de M_0. Un tel point M_0 est dit régulier :



Deuxième cas : $\overrightarrow{V}(M_0) = \overrightarrow{0}$. On voit alors facilement sur des exemples qu'il y a plusieurs configurations possibles pour le champ des tangentes :



$$\overrightarrow{V}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \qquad \overrightarrow{V}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ -y \end{pmatrix} \qquad \overrightarrow{V}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y \\ x \end{pmatrix}$$

Si $\overrightarrow{V}(M_0) = \overrightarrow{0}$, on dit que M_0 est un *point singulier* (ou *point critique*) du champ de vecteurs. Un tel point donne évidemment une solution constante $M(t) = M_0$ de (E). Pour étudier les solutions voisines, on supposera après translation éventuelle de l'origine des coordonnées que $M_0 = 0$. On a alors $f(0,0) = g(0,0) = 0$, de sorte que le système différentiel peut s'écrire

$$\begin{cases} \dfrac{dx}{dt} = f(x,y) = ax + by + o(|x| + |y|) \\ \dfrac{dy}{dt} = g(x,y) = cx + dy + o(|x| + |y|). \end{cases}$$

Introduisons la matrice

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} f'_x(0,0) & f'_y(0,0) \\ g'_x(0,0) & g'_y(0,0) \end{pmatrix}.$$

Le système différentiel considéré s'écrit maintenant

$$\frac{dM}{dt} = AM + G(M)$$

avec $G(0,0) = G'_x(0,0) = G'_y(0,0) = 0$. La fonction continue

$$k(r) = \sup_{M \in \overline{B}(0,r)} |||G'(M)|||$$

tend vers 0 quand r tend vers 0 et le théorème des accroissements finis donne

$$\|G(M_1) - G(M_2)\| \leq k(r)\|\overrightarrow{M_1M_2}\|$$

pour tous $M_1, M_2 \in \overline{B}(0,r)$. L'hypothèse (b) du théorème du § 1.3 est donc satisfaite. Dire que le point M_0 est asymptotiquement stable signifie que les lignes intégrales issues d'un point M_1 voisin de M_0 convergent toutes vers M_0 (à peu près uniformément à la même vitesse) quand le temps tend vers $+\infty$. On peut donc énoncer :

Proposition – *Pour qu'un point singulier $M_0 = (x_0, y_0)$ soit asymptotiquement stable, il suffit que les valeurs propres de la matrice jacobienne*

$$A = \begin{pmatrix} f'_x(x_0,y_0) & f'_y(x_0,y_0) \\ g'_x(x_0,y_0) & g'_y(x_0,y_0) \end{pmatrix}$$

soient de partie réelle < 0.

Remarque – Contrairement au cas d'un système linéaire, on ne peut pas décider de la nature du point critique si la matrice jacobienne a une valeur propre de partie réelle nulle. Considérons par exemple le système

$$\begin{cases} \dfrac{dx}{dt} = \alpha x^3 \\ \dfrac{dy}{dt} = \beta y^3 \end{cases} \qquad t \in [t_0, +\infty[= [0, +\infty[,$$

qui admet l'origine comme point critique avec matrice jacobienne $A = 0$. On voit facilement que la solution du problème de Cauchy est

$$x(t) = x_0(1 - 2\alpha x_0^2 t)^{-1/2}, \quad y(t) = y_0(1 - 2\,\beta y_0^2 t)^{-1/2}.$$

Par conséquent l'origine est un point asymptotiquement stable si $\alpha < 0$ et $\beta < 0$, instable dès que $\alpha > 0$ ou $\beta > 0$. Dans ce dernier cas, les solutions ne sont en fait même pas globalement définies : lorsque $\alpha > 0$, $\beta \leq 0$ et $x_0 \neq 0$, la solution maximale n'est définie que pour $t \in [0, 1/(2\,\alpha x_0^2)[$.

Si la matrice jacobienne est inversible (valeurs propres $\neq 0$), le théorème d'inversion locale montre que la fonction $M \mapsto \overrightarrow{V}(M)$ définit une bijection d'un voisinage de M_0 sur un voisinage de 0 ; en particulier, pour M assez voisin et distinct de M_0 on aura $\overrightarrow{V}(M) \neq \overrightarrow{0}$, de sorte que M_0 est un point singulier *isolé*. Ce n'est pas toujours le cas si la matrice est dégénérée : le champ $\overrightarrow{V}(x,y) = (x,0)$ admet par exemple toute une droite $x = 0$ de points singuliers. On exclura en général ces situations qui peuvent être extrêmement compliquées.

Définition – *On dira qu'un point singulier M_0 est non dégénéré si*

$$\det \begin{pmatrix} f'_x(x_0,y_0) & f'_y(x_0,y_0) \\ g'_x(x_0,y_0) & g'_y(x_0,y_0) \end{pmatrix} \neq 0.$$

Nous nous proposons maintenant d'étudier les différentes configurations possibles pour un point singulier non dégénéré. On verra au chapitre XI, § 2.3 que les courbes intégrales ont tendance à ressembler à celles du système linéaire $dM/dt = AM$ lorsqu'on se rapproche du point critique, tout au moins sur un intervalle de temps $[t_0, t_1]$ fixé ; ceci n'est pas nécessairement vrai sur tout l'intervalle $[t_0, +\infty[$ (voir § 2.3 pour des exemples). On se restreindra dans un premier temps au cas linéaire.

2.2. Cas d'un champ linéaire de vecteurs

Considérons le système

$$\frac{dM}{dt} = AM, \qquad \begin{cases} \dfrac{dx}{dt} = ax + by \\ \dfrac{dy}{dt} = cx + dy \end{cases} \quad \text{où} \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

On supposera $\det A \neq 0$, de sorte que le champ de vecteurs $\overrightarrow{V}(M) = AM$ admet l'origine pour seul point critique. Comme le champ des tangentes est invariant par les homothéties de centre O, les courbes intégrales se déduisent les unes des autres par homothéties. Distinguons maintenant plusieurs cas en fonction des valeurs propres de A.

(a) Les valeurs propres λ_1, λ_2 de A sont réelles.

• Supposons de plus $\lambda_1 \neq \lambda_2$. Dans ce cas la matrice A est diagonalisable. Après changement de base on peut supposer

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

et le système se réduit à

$$\begin{cases} \dfrac{dx}{dt} = \lambda_1 x \\ \dfrac{dy}{dt} = \lambda_2 y. \end{cases}$$

La solution du problème de Cauchy avec $M(0) = (x_0, y_0)$ est donc

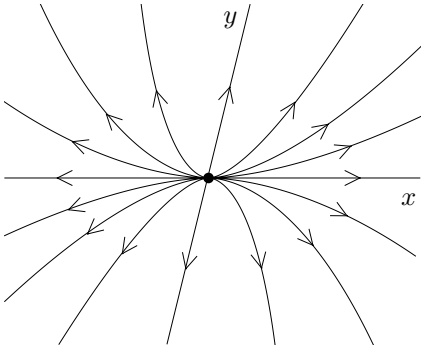
$$\begin{cases} x(t) = x_0 e^{\lambda_1 t} \\ y(t) = y_0 e^{\lambda_2 t} \end{cases}$$

de sorte que les courbes intégrales sont les courbes

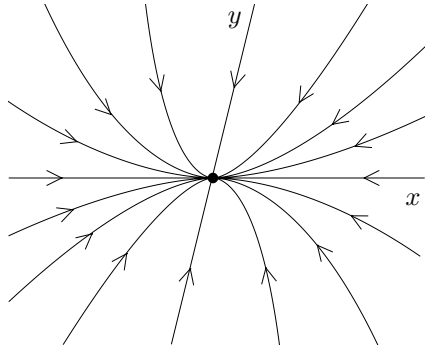
$$y = C|x|^{\lambda_2/\lambda_1}, \quad C \in \mathbb{R}$$

et la droite d'équation $x = 0$. Distinguons deux sous-cas :

∗ λ_1, λ_2 de même signe et, disons, $|\lambda_1| < |\lambda_2|$. On a alors $\lambda_2/\lambda_1 > 1$. On dit qu'on a affaire à un *nœud impropre* :

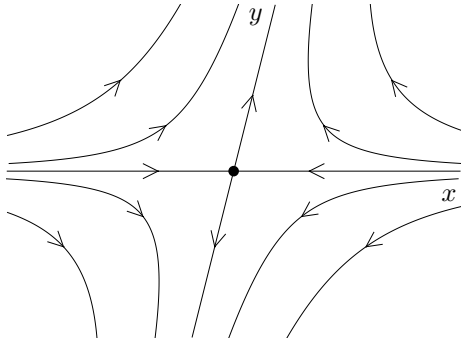


$0 < \lambda_1 < \lambda_2$
nœud impropre instable



$\lambda_2 < \lambda_1 < 0$
nœud impropre stable

∗ λ_1, λ_2 de signes opposés, par exemple $\lambda_1 < 0 < \lambda_2$. Il s'agit d'un *col* (toujours instable) :

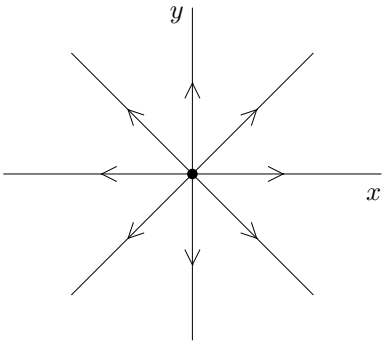


• Les valeurs propres sont confondues : $\lambda_1 = \lambda_2 = \lambda$. Deux cas sont possibles :

∗ A est diagonalisable. Alors A est en fait diagonale et les courbes intégrales sont données par

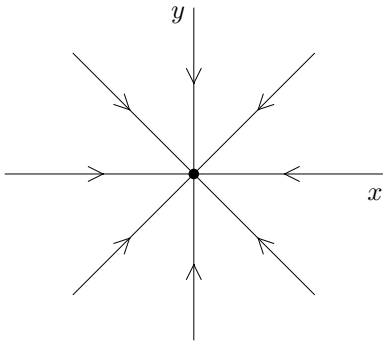
$$\begin{cases} x(t) = x_0 e^{\lambda t} \\ y(t) = y_0 e^{\lambda t}, \end{cases}$$

ce sont les droites $y = \alpha x$ et $x = 0$. On dit qu'on a affaire à un *nœud propre* :



$\lambda > 0$

Nœud propre instable



$\lambda < 0$

Nœud propre stable

$*$ A est non diagonalisable. Alors il existe une base dans laquelle la matrice A et le système s'écrivent

$$A = \begin{pmatrix} \lambda & 0 \\ 1 & \lambda \end{pmatrix}, \qquad \begin{cases} \dfrac{dx}{dt} = \lambda x \\ \dfrac{dy}{dt} = x + \lambda y. \end{cases}$$

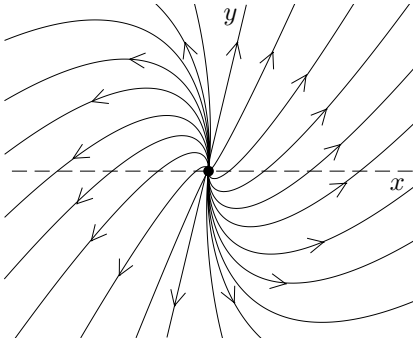
Le courbes intégrales sont données par

$$\begin{cases} x(t) = x_0 e^{\lambda t} \\ y(t) = (y_0 + x_0 t) e^{\lambda t}. \end{cases}$$

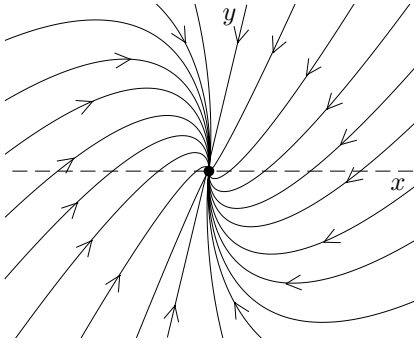
Comme toute courbe intégrale avec $x_0 \neq 0$ passe par un point tel que $|x(t)| = 1$, on obtient toutes les courbes intégrales autres que $x = 0$ en prenant $x_0 = \pm 1$, d'où

$$\begin{cases} t = \dfrac{1}{\lambda} \ \ln |x| \\ y = y_0 |x| + \dfrac{x}{\lambda} \ \ln |x| \end{cases}$$

On dit qu'il s'agit d'un *nœud exceptionnel*. Pour construire les courbes, on tracera par exemple d'abord la courbe $y = \frac{x}{\lambda} \ \ln |x|$ passant par $(x_0, y_0) = (\pm 1, 0)$. Toutes les autres s'en déduisent par homothéties.



$\lambda > 0$
Nœud exceptionnel instable



$\lambda < 0$
Nœud exceptionnel stable

(b) Les valeurs propres de A sont non réelles.

On a des valeurs propres complexes conjuguées $\alpha + i\beta$, $\alpha - i\beta$ avec disons $\beta > 0$, et il existe une base dans laquelle la matrice A et le système s'écrivent

$$A = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}, \qquad \begin{cases} \dfrac{dx}{dt} = \alpha x - \beta y \\ \dfrac{dy}{dt} = \beta x + \alpha y. \end{cases}$$

La manière la plus rapide de résoudre un tel système est de poser $z = x + iy$. On trouve alors

$$\frac{dz}{dt} = (\alpha + i\beta)x + (-\beta + \alpha i)y = (\alpha + i\beta)(x + iy) = (\alpha + i\beta)z,$$

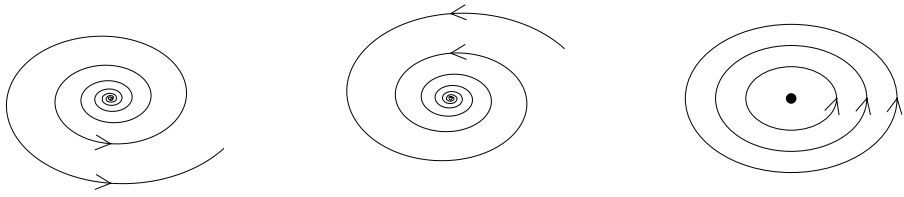
de sorte que la solution générale est

$$z(t) = z_0 e^{(\alpha+i\beta)t} = z_0 e^{\alpha t} e^{i\beta t}.$$

En coordonnées polaires $z = re^{i\theta}$, l'équation devient

$$\begin{cases} r = r_0 e^{\alpha t} \\ \theta = \theta_0 + \beta t \end{cases}, \quad \text{soit} \quad r = r_0 e^{\frac{\alpha}{\beta}(\theta - \theta_0)}.$$

Il s'agit d'une spirale logarithmique si $\alpha \neq 0$ et d'un cercle si $\alpha = 0$ (noter que ce cercle donne en général graphiquement une ellipse car la base utilisée ci-dessus n'est pas nécessairement orthonormée). On dit alors que le point singulier est un *foyer*, respectivement un *centre* :



$\alpha > 0$ Foyer instable | $\alpha < 0$ Foyer stable | $\alpha = 0$ Centre

Si $\alpha \neq 0$, le rapport d'homothétie de deux spires consécutives de la spirale est $e^{2\pi\alpha/\beta}$.

2.3. *Singularités de champs de vecteurs non linéaires*

L'objet de ce paragraphe est essentiellement de mettre en garde le lecteur contre un certain nombre d'idées fausses fréquemment rencontrées, en particulier l'idée que les courbes intégrales d'un champ de vecteurs quelconque au voisinage d'un point singulier ressemblent toujours à celles du système linéaire associé. En fait, ce n'est en général pas le cas lorsque le système linéarisé présente *un centre*, et cela peut de même ne pas être le cas lorsque celui-ci présente *un nœud*. Les deux exemples ci-dessous illustrent le phénomène.

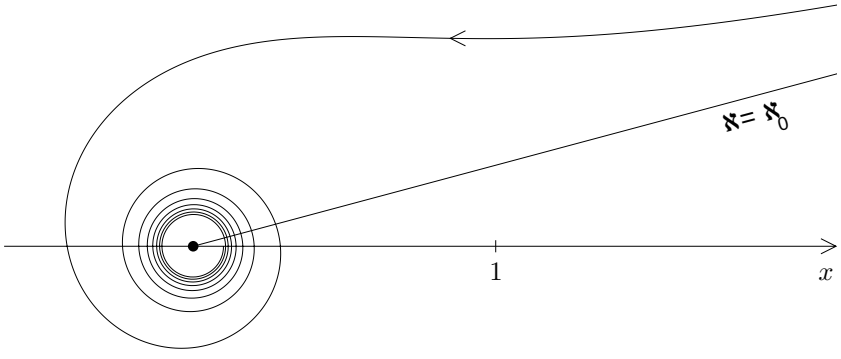
Exemple 1 – On considère le système différentiel

$$\text{(S)} \qquad \begin{cases} \dfrac{dx}{dt} = -y - x(x^2+y^2) \\ \dfrac{dy}{dt} = \quad x - y(x^2+y^2). \end{cases}$$

L'origine est un point critique non dégénéré, et le système linéaire associé $dx/dt = -y$, $dy/dt = x$ présente un centre d'après le §2.2. Si l'on passe en coordonnées polaires (r, θ) le système (S) devient

$$\begin{cases} r\,\dfrac{dr}{dt} = x\,\dfrac{dx}{dt} + y\,\dfrac{dy}{dt} = -(x^2+y^2)^2 \\ \dfrac{d\theta}{dt} = \dfrac{x\,\frac{dy}{dt} - y\,\frac{dx}{dt}}{x^2+y^2} = 1 \end{cases} \iff \begin{cases} \dfrac{dr}{r^3} = -dt \\ \\ d\theta = dt \end{cases}$$

car $r dr = x dx + y dy$ et $x dy - y dx = r^2 d\theta$ (exercice !). Les courbes intégrales de l'équation $-dr/r^3 = d\theta$ sont données par $1/2r^2 = \theta - \theta_0$, soit $r = (2(\theta - \theta_0))^{-1/2}$ pour $\theta > \theta_0$. On a ici $\theta = t + C$, $\lim_{\theta \to +\infty} r(\theta) = 0$. On voit que les courbes intégrales sont des spirales convergeant vers 0 quand $t \to +\infty$, l'origine est donc *un foyer* stable.



Exemple 2 – On considère maintenant le système différentiel

$$(\mathrm{S}) \qquad \begin{cases} \dfrac{dx}{dt} = -x - \dfrac{2y}{\ln(x^2+y^2)} \\ \dfrac{dy}{dt} = -y + \dfrac{2x}{\ln(x^2+y^2)} \end{cases}$$

sur le disque unité ouvert $x^2+y^2<1$. Observons que $2y/\ln(x^2+y^2)$ se prolonge en une fonction de classe C^1 au voisinage de $(0,0)$: elle admet en effet une limite égale à 0 à l'origine, ainsi que ses dérivées partielles

$$\frac{-4xy}{(x^2+y^2)(\ln(x^2+y^2))^2}, \quad \frac{2}{\ln(x^2+y^2)} - \frac{4y^2}{(x^2+y^2)(\ln(x^2+y^2))^2}.$$

Il en est de même pour le terme $2x/\ln(x^2+y^2)$. L'origine est donc un point singulier, et le système linéaire associé $dx/dt=-x$, $dy/dt=-y$ présente *un nœud propre*. Pour résoudre (S), on utilise de nouveau les coordonnées polaires (r,θ). Il vient

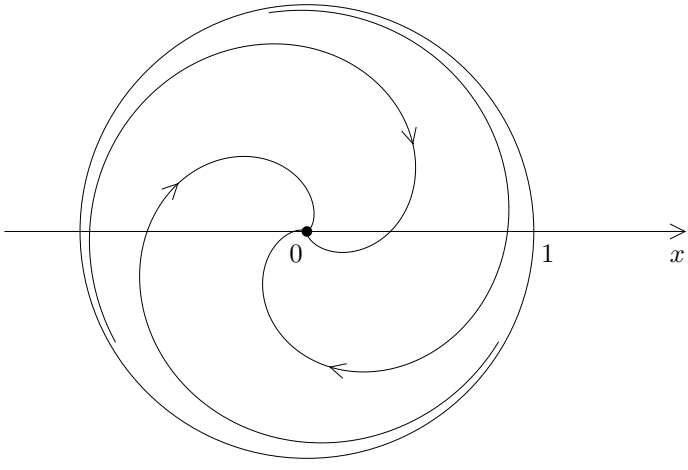
$$\begin{cases} \dfrac{dr}{dt} = -\dfrac{x^2+y^2}{r} = -r \\ \dfrac{d\theta}{dt} = \dfrac{1}{x^2+y^2}\ \dfrac{2x^2+2y^2}{\ln(x^2+y^2)} = \dfrac{1}{\ln r}. \end{cases}$$

La solution du problème de Cauchy est donnée par

$$r = r_0 e^{-t} \quad \text{avec} \quad r_0<1,$$
$$d\theta = \frac{dt}{\ln r_0 - t}, \quad \theta = \theta_0 - \ln(1 - t/\ln r_0)$$

pour une donnée initiale (r_0,θ_0) en $t=0$. La solution est définie sur $[\ln r_0, +\infty[$ et on a $\lim_{t\to+\infty} r(t) = 0$, $\lim_{t\to+\infty}\theta(t) = -\infty$. On a ici encore une spirale convergeant vers 0 (c'est peu visible sur le schéma ci-dessous car θ tend vers $-\infty$ très lentement). L'origine est donc un *foyer stable*. Comme $\lim_{t\to\ln r_0+0} r(t) = 1_-$

et $\lim_{t\to \ln r_0+0} \theta(t) = +\infty$, la courbe s'enroule en spiralant à l'intérieur du cercle $r = 1$ quand $t \to \ln r_0 + 0$.



3. PROBLÈMES

3.1. On considère sur $\mathbb{R}^2$ le champ de vecteurs

$$\overrightarrow{V}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^2 - y^2 \\ 2xy \end{pmatrix}.$$

(a) Déterminer les points critiques.

(b) En posant $z = x + iy$, calculer la solution correspondant à la donnée initiale z_0 au temps $t = 0$.

(c) En déduire que les courbes intégrales sont les deux demi-axes $0x, 0x'$ et les cercles passant par l'origine, centrés sur l'axe $y'0y$.

(d) Montrer que les solutions telles que $z_0 \in \mathbb{C} \setminus [0, +\infty[$ sont asymptotiquement stables. Qu'en est-il si $z_0 \in [0, +\infty[$?

3.2. On étudie dans $\mathbb{R}^2$ le système différentiel

$$\text{(S)} \qquad \frac{d\overrightarrow{M}}{dt} = \overrightarrow{V}(M)$$

où $\overrightarrow{V}$ désigne le champ de vecteurs qui à tout point $M(x, y) \in \mathbb{R}^2$ associe le vecteur

$$\overrightarrow{V}(M) = (-x^2 - y, -x + y^2).$$

Déterminer les points critiques du champ de vecteurs $\overrightarrow{V}$. Calculer les solutions $t \mapsto \widetilde{M}(t) = (\widetilde{x}(t), \widetilde{y}(t))$ du système différentiel obtenu en linéarisant $\overrightarrow{V}$ au voisinage de chacun des points critiques. Faire un schéma indiquant l'allure des solutions au voisinage des points critiques. Ces points sont-ils stables ?

3.3. Mêmes questions pour le champ de vecteurs $\overrightarrow{V}(x, y) = (-1 + x^2 + y^2, -x)$.

3.4. On considère le champ de vecteurs défini par

$$\overrightarrow{V}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y + x \sin\left(\frac{\pi}{x^2+y^2}\right) \exp\left(-\frac{1}{x^2+y^2}\right) \\ x + y \sin\left(\frac{\pi}{x^2+y^2}\right) \exp\left(-\frac{1}{x^2+y^2}\right) \end{pmatrix}$$

pour $(x, y) \neq (0, 0)$, et par $\overrightarrow{V}(0, 0) = \overrightarrow{0}$.

(a) Montrer que le champ $\overrightarrow{V}$ est de classe C^∞ sur $\mathbb{R}^2$; on pourra commencer par montrer que la fonction

$$t \mapsto \sin(\pi/t) \exp(-1/t), \quad t > 0,$$

se prolonge en une fonction de classe C^∞ sur $[0, +\infty[$.

(b) Montrer que le système différentiel $dM/dt = \overrightarrow{V}(M)$ se ramène à une équation de la forme $dr/d\theta = f(r)$; on ne cherchera pas à résoudre explicitement cette équation. En déduire qu'il y a une infinité de cercles concentriques $(C_k)_{k\geq 1}$ de rayons R_k décroissant vers 0, qui sont des courbes intégrales du champ.

(c) Étudier la convergence des intégrales $\int_{R_{k+1}}^{R_k} dr/f(r)$ en chacune des bornes. Montrer que la courbe intégrale issue d'un point de coordonnées polaires (r_0, θ_0) telles que $R_{k+1} < r_0 < R_k$ est une spirale admettant les cercles $r = R_k$ et $r = R_{k+1}$ comme asymptotes. Étudier de même le comportement à l'infini des courbes intégrales.

CHAPITRE XI

ÉQUATIONS DIFFÉRENTIELLES DÉPENDANT D'UN PARAMÈTRE

Étant donné une équation différentielle $y' = f(t, y, \lambda)$ dépendant d'un paramètre λ, on se propose d'étudier comment les solutions varient en fonction de λ. En particulier on montrera que, sous des hypothèses convenables, les solutions dépendent continûment ou différentiablement du paramètre λ. Outre l'aspect théorique, ces résultats sont importants en vue de la méthode dite des perturbations : il arrive fréquemment qu'on sache calculer la solution y pour une valeur particulière λ_0, mais pas pour les valeurs voisines λ ; on cherche alors un développement limité de la solution y associée à la valeur λ en fonction de $\lambda - \lambda_0$. On montrera que le coefficient de $\lambda - \lambda_0$ est obtenu en résolvant une équation différentielle linéaire, appelée équation « linéarisée » de l'équation initiale ; ce fait remarquable permet généralement de bien étudier les petites perturbations de la solution.

1. DÉPENDANCE DE LA SOLUTION EN FONCTION DU PARAMÈTRE

1.1. NOTATIONS

Soit U un ouvert de $\mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^p$ et

$$\begin{aligned} f : \quad & U \to \mathbb{R}^m \\ & (t, y, \lambda) \mapsto f(t, y, \lambda) \end{aligned}$$

une fonction continue. Pour chaque valeur de $\lambda \in \mathbb{R}^p$, on considère l'équation différentielle

$$(\mathrm{E}_\lambda) \qquad y' = f(t, y, \lambda), \qquad (t, y) \in U_\lambda$$

où $U_\lambda \subset \mathbb{R} \times \mathbb{R}^m$ est l'ouvert des points tels que $(t, y, \lambda) \in U$. Une donnée initiale (t_0, y_0) étant fixée, on note $y(t, \lambda)$ la solution maximale du problème de Cauchy relatif à (E_λ) telle que $y(t_0, \lambda) = y_0$; on supposera toujours dans la suite que les hypothèses assurant l'unicité des solutions sont vérifiées. Notre objectif est d'étudier la continuité ou la différentiabilité de $y_0(t, \lambda)$ en fonction du couple (t, λ).

Fixons un point $(t_0, y_0, \lambda_0) \in U$. Comme U est ouvert, ce point admet un voisinage (compact) contenu dans U

$$V_0 = [t_0 - T_0, t_0 + T_0] \times \overline{B}(y_0, r_0) \times \overline{B}(\lambda_0, \alpha_0).$$

On note $M = \sup_{V_0} \|f\|$. Alors pour tout $T \leq \min\left(T_0, \frac{r_0}{M}\right)$ fixé et pour tout $\lambda \in \overline{B}(\lambda_0, \alpha_0)$, le cylindre

$$C = [t_0 - T, t_0 + T] \times \overline{B}(y_0, r_0) \subset U_\lambda$$

est un cylindre de sécurité pour les solutions de E_λ, d'après les résultats du chapitre V, § 2.1. Le théorème d'existence V 2.4 implique :

Proposition – *Avec les notations précédentes, la solution $y(t, \lambda)$ est définie pour tout $(t, \lambda) \in [t_0 - T, t_0 + T] \times \overline{B}(\lambda_0, \alpha_0)$ et elle est à valeurs dans $\overline{B}(y_0, r_0)$.*

1.2. CONTINUITÉ

On suppose maintenant de plus que f est localement lipschitzienne en y, c'est-à-dire qu'après avoir éventuellement rétréci V_0, il existe une constante $k \geq 0$ telle que

$$\forall (t, \lambda) \in [t_0 - T_0, t_0 + T_0] \times \overline{B}(\lambda_0, \alpha_0), \quad \forall y_1, y_2 \in \overline{B}(y_0, r_0),$$
$$\|f(t, y_1, \lambda) - f(t, y_2, \lambda)\| \leq k\|y_1 - y_2\|.$$

Théorème – *Si f est continue sur U et localement lipschitzienne en y, alors la solution $y(t, \lambda)$ est continue sur $[t_0 - T, t_0 + T] \times \overline{B}(\lambda_0, \alpha_0)$.*

Démonstration. Remarquons d'abord que

$$\left\| \frac{d}{dt}\, y(t, \lambda) \right\| = \|f(t, y(t, \lambda), \lambda)\| \leq M$$

car $\|f\| \leq M$ sur V_0. Le théorème des accroissements finis montre alors que $y(t, \lambda)$ est M-lipschitzienne par rapport à t, c'est-à-dire que

$$\|y(t_1, \lambda) - y(t_2, \lambda)\| \leq M|t_1 - t_2|$$

pour tous (t_1, λ), $(t_2, \lambda) \in [t_0 - T, t_0 + T] \times \overline{B}(\lambda_0, \alpha_0)$. Par ailleurs, comme V_0 est compact, f y est conformément continue et il existe donc un module de continuité $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ tel que

$$\|f(t, y, \lambda_1) - f(t, y, \lambda_2)\| \leq \eta(\|\lambda_1 - \lambda_2\|)$$

avec $\lim_{u \to 0_+} \eta(u) = 0$. Alors $z_1(t) = y(t, \lambda_1)$ est la solution exacte du problème de Cauchy pour l'équation

$$(\mathrm{E}_{\lambda_1}) \qquad y' = f(t, y, \lambda_1),$$

tandis que $z_2(t) = y(t, \lambda_2)$ en est une solution ε-approchée avec $\varepsilon = \eta(\|\lambda_1 - \lambda_2\|)$. Le lemme de Gronwall V 3.1 montre que

$$\|z_1(t) - z_2(t)\| \leq \varepsilon \, \frac{e^{k|t-t_0|} - 1}{k}, \quad \text{d'où}$$
$$\|y(t, \lambda_1) - y(t, \lambda_2)\| \leq \frac{e^{kT} - 1}{k} \, \eta(\|\lambda_1 - \lambda_2\|).$$

De ces inégalités, nous déduisons

$$\begin{aligned}\|y(t_1, \lambda_1) - y(t_2, \lambda_2)\| &\leq \|y(t_1, \lambda_1) - y(t_2, \lambda_1)\| + \|y(t_2, \lambda_1) - y(t_2, \lambda_2)\| \\ &\leq M|t_1 - t_2| + \frac{e^{kT} - 1}{k} \, \eta(\|\lambda_1 - \lambda_2\|)\end{aligned}$$

et comme le second membre tend vers 0 lorsque $(t_2, \lambda_2) \to (t_1, \lambda_1)$, on voit que $y(t, \lambda)$ est bien continue sur $[t_0 - T, t_0 + T] \times \overline{B}(\lambda_0, \alpha_0)$. ∎

Remarque – La démonstration montre aussi que si $f(t, y, \lambda)$ est localement lipschitzienne en λ, alors $y(t, \lambda)$ est localement lipschitzienne en (t, λ) : dans ce cas, on peut prendre $\eta(u) = Cu$.

1.3. Différentiabilité

Afin de simplifier les notations, on suppose dans un premier temps que $\lambda \in \mathbb{R}$, c'est-à-dire que $p = 1$. Pour deviner les résultats, nous effectuons d'abord un calcul formel en supposant les fonctions f et $y(t, \lambda)$ autant de fois différentiables que nécessaire. Comme y satisfait (E_λ) par hypothèse, on a

$$\frac{\partial y}{\partial t}\,(t, \lambda) = f(t, y(t, \lambda), \lambda).$$

Différentions cette relation par rapport à λ :

$$\frac{\partial^2 y}{\partial \lambda \partial t}\,(t, \lambda) = \sum_{j=1}^{m} f'_{y_j}(t, y(t, \lambda), \lambda) \, \frac{\partial y_j}{\partial \lambda}\,(t, \lambda) + f'_\lambda(t, y(t, \lambda), \lambda).$$

En posant $u(t) = y(t, \lambda)$ et $v(t) = \frac{\partial y}{\partial \lambda}\,(t, \lambda)$ il vient

$$(\mathrm{E}'_\lambda) \qquad v'(t) = \sum_{j=1}^{m} f'_{y_j}(t, u(t), \lambda) v_j(t) + f'_\lambda(t, u(t), \lambda).$$

On observe que l'équation (E'_λ) satisfaite par v est *linéaire*. L'équation (E'_λ) s'appelle l'*équation différentielle linéarisée* associée à (E_λ). Par ailleurs v satisfait la condition initiale

$$v(t_0) = \frac{\partial y}{\partial \lambda}\,(t_0, \lambda) = 0,$$

car par hypothèse $y(t_0, \lambda) = y_0$ ne dépend pas de λ.

Théorème – *On suppose que f est continue sur U et admet des dérivées partielles f'_{y_j} et f'_λ continues sur U.*
Alors $y(t,\lambda)$ est de classe C^1 sur $[t_0-T, t_0+T]\times B(\lambda_0,\alpha_0)$ et admet des dérivées partielles secondes croisées continues

$$\frac{\partial}{\partial\lambda}\,\frac{\partial y}{\partial t} = \frac{\partial}{\partial t}\,\frac{\partial y}{\partial\lambda}.$$

De plus, si $u(t) = y(t,\lambda)$, la dérivée partielle $v(t) = \frac{\partial y}{\partial\lambda}(t,\lambda)$ est la solution de l'équation différentielle linéarisée

$$(\mathrm{E}'_\lambda) \qquad v'(t) = \sum_{j=1}^{m} f'_{y_j}(t,u(t),\lambda)v_j(t) + f'_\lambda(t,u(t),\lambda)$$

avec condition initiale $v(t_0)=0$.

Démonstration.* Les hypothèses entraînent que f est localement lipschitzienne en la variable y, donc on sait déjà que $y(t,\lambda)$ est continue. Pour λ_1 fixé posons $u_1(t) = y(t,\lambda_1)$ et soit $v_1(t)$ la solution de l'équation linéaire

$$(\mathrm{E}'_{\lambda_1}) \qquad v'_1(t) = \sum_{j=1}^{m} f'_{y_j}(t,u_1(t),\lambda_1)v_{1,j}(t) + f'_\lambda(t,u_1(t),\lambda_1).$$

avec condition initiale $v_1(t_0) = 0$. Bien entendu, on ne sait pas encore que $v_1(t) = \frac{\partial y}{\partial\lambda}(t,\lambda_1)$, c'est justement ce qu'on veut démontrer. Pour cela on compare $u(t) = y(t,\lambda)$ et $u_1(t) + (\lambda-\lambda_1)v_1(t)$ et on cherche à montrer que la différence est $o(\lambda-\lambda_1)$. Posons donc

$$w(t) = u(t) - u_1(t) - (\lambda-\lambda_1)v_1(t).$$

Par définition de u, u_1, v_1 il vient

$$w'(t) = f(t,u(t),\lambda) - f(t,u_1(t),\lambda_1) \\ -(\lambda-\lambda_1)\left(\sum_{j=1}^{m} f'_{y_j}(t,u_1(t),\lambda_1)v_{1,j}(t) + f'_\lambda(t,u_1(t),\lambda_1)\right) \qquad (*)$$

Pour chaque composante f_k, la formule des accroissements finis montre que

$$f_k(t,y,\lambda) - f_k(t,y_1,\lambda_1) = \sum_{j=1}^{m} f'_{k,y_j}(t,\widetilde{y},\widetilde{\lambda})(y_j - y_{1,j}) + f'_{k,\lambda}(t,\widetilde{y},\widetilde{\lambda})(\lambda-\lambda_1)$$

où $(\widetilde{y},\widetilde{\lambda})$ est un point appartenant au segment d'extrémités (y_1,λ_1) et (y,λ). Si η_k est un module de continuité uniforme pour les dérivées partielles f'_{k,y_j} et $f'_{k,\lambda}$ sur le compact V_0, l'écart de chaque fonction entre les points $(\widetilde{y},\widetilde{\lambda})$ et (y_1,λ_1) est majoré par

$$\eta_k(\|\widetilde{y}-y_1\| + |\widetilde{\lambda}-\lambda_1|) \le \eta_k(\|y-y_1\| + |\lambda-\lambda_1|).$$

On peut donc écrire

$$f(t,y,\lambda)-f(t,y_1,\lambda_1)=\sum_{j=1}^{m} f'_{y_j}(t,y_1,\lambda_1)(y_j-y_{1,j})+f'_\lambda(t,y_1,\lambda_1)(\lambda-\lambda_1)+g(t,y,y_1,\lambda)$$

où $g(t,y,y_1,\lambda)$ admet une majoration uniforme

$$\|g(t,y,y_1,\lambda)\| \le (\|y-y_1\|+|\lambda-\lambda_1|)\,\eta(\|y-y_1\|+|\lambda-\lambda_1|)$$

pour un certain module de continuité η. En substituant $y=u(t)$, $y_1=u_1(t)$ dans la dernière formule, on déduit de $(*)$ les relations

$$w'(t)=\sum_{j=1}^{m} f'_{y_j}(t,u_1(t),\lambda_1)(u_j(t)-u_{1,j}(t)-(\lambda-\lambda_1)v_{1,j}(t))+g(t,u(t),u_1(t),\lambda),$$

$$w'(t)=\sum_{j=1}^{m} f'_{y_j}(t,u_1(t),\lambda_1)w_j(t)+g(t,u(t),u_1(t),\lambda), \qquad (**)$$

avec la majoration uniforme

$$\|g(t,u(t),u_1(t),\lambda)\| \le (C+1)|\lambda-\lambda_1|\ \eta((C+1)|\lambda-\lambda_1|)=o(\lambda-\lambda_1)\ ;$$

en effet

$$\|u(t)-u_1(t)\|=\|y(t,\lambda)-y(t,\lambda_1)\| \le C|\lambda-\lambda_1|$$

d'après la remarque finale du § 1.2. L'équation $(**)$ est linéaire en w, et donc K-lipschitzienne avec

$$K=\sup_{V_0}\sum_{1\le j\le m}\|f'_{y_j}\|.$$

Comme $u(t_0)=u_1(t_0)=y_0$ et $v_1(t_0)=0$, on a $w(t_0)=0$, et par ailleurs $\widetilde{w}(t)\equiv 0$ est une solution ε-approchée de $(**)$ avec $\varepsilon=o(\lambda-\lambda_1)$. Le lemme de Gronwall V 3.1 montre que

$$\|w(t)\|=\|w(t)-\widetilde{w}(t)\| \le \varepsilon\,\frac{e^{KT}-1}{K}=o(\lambda-\lambda_1),$$

c'est-à-dire, par définition de w, u, u_1 :

$$\|y(t,\lambda)-y(t,\lambda_1)-(\lambda-\lambda_1)v_1(t)\|=o(\lambda-\lambda_1).$$

Ceci signifie que $\frac{\partial y}{\partial \lambda}\,(t,\lambda_1)$ existe et coïncide avec $v_1(t)$. La fonction y admet donc bien des dérivées partielles premières

$$\frac{\partial y}{\partial t}\,(t,\lambda)=f(t,y(t,\lambda),\lambda) \quad \text{et} \quad \frac{\partial y}{\partial \lambda}\,(t,\lambda).$$

La dérivée partielle $\frac{\partial y}{\partial t}$ est continue en (t,λ) puisque y l'est. Par ailleurs $v(t,\lambda)=\frac{\partial y}{\partial \lambda}\,(t,\lambda)$ est la solution avec données initiales $t_0, v_0=0$ de l'équation linéarisée

$$(\mathrm{E}'_\lambda) \qquad v'=G(t,v,\lambda)=\sum_{j=1}^{m} f'_{y_j}(t,y(t,\lambda),\lambda)v_j+f'_\lambda(t,y(t,\lambda),\lambda).$$

Ici G est continue en (t, v, λ) et localement lipschitzienne en v, puisque linéaire en cette variable. Par suite $v = \frac{\partial y}{\partial \lambda}$ est également continue en (t, λ), ce qui implique que y est de classe C^1. Enfin, on a bien

$$\begin{aligned}\frac{\partial}{\partial t}\,\frac{\partial y}{\partial \lambda} &= \frac{\partial}{\partial t}\; v(t,\lambda) \\ &= \sum_{j=1}^{m} f'_{y_j}(t, y(t,\lambda), \lambda)\;\frac{\partial y_j}{\partial \lambda}\;(t,\lambda) + f'_{\lambda}(t, y(t,\lambda), \lambda) \\ &= \frac{\partial}{\partial \lambda}\;(f(t, y(t,\lambda), \lambda) = \frac{\partial}{\partial \lambda}\;\frac{\partial y}{\partial t}\;(t,\lambda)\end{aligned}$$

et ces dérivées partielles secondes sont continues grâce à la deuxième ligne. ■

Généralisation – *Le théorème s'étend aisément au cas où $\lambda \in \mathbb{R}^p$. Il suffit en effet de fixer toutes les variables λ_i sauf une pour constater que $\frac{\partial}{\partial \lambda_i}\,\frac{\partial y}{\partial t} = \frac{\partial}{\partial t}\,\frac{\partial y}{\partial \lambda_i}$ et que $v(t) = \frac{\partial y}{\partial \lambda_i}\,(t,\lambda)$ est la solution de l'équation linéarisée*

$$(\mathrm{E}_{\lambda_i}) \qquad v'(t) = \sum_{j=1}^{m} f'_{y_j}(t, u(t), \lambda) v_j(t) + f'_{\lambda_i}(t, u(t), \lambda)$$

avec condition initiale $v(t_0) = 0$.

Le fait que $y(t, \lambda)$ soit encore de classe C^1 provient de ce que le théorème de continuité du § 1.2 est vrai globalement par rapport à l'ensemble des variables $(\lambda_1, \ldots, \lambda_p)$: celui-ci entraîne la continuité en (t, λ) des dérivées partielles $\partial y/\partial \lambda_i$. Nous étudions maintenant la différentiabilité d'ordre supérieur : nous allons voir qu'elle se déduit aussitôt par récurrence du cas de l'ordre un.

Théorème – *On suppose que f est de classe C^s et admet des dérivées partielles f'_{y_j} et f'_{λ_i} de classe C^s. Alors la solution $y(t, \lambda)$ du problème de Cauchy relatif à* (E_λ) *est de classe C^{s+1}.*

Démonstration. Par récurrence sur s. Pour $s = 0$, il s'agit du théorème précédent. Supposons le résultat déjà démontré à l'ordre $s - 1$. Alors $y(t, \lambda)$ est de classe C^s. Il en est de même de sa dérivée

$$\frac{\partial y}{\partial t}\,(t,\lambda) = f(t, y(t,\lambda), \lambda).$$

De plus $v(t, \lambda) = \frac{\partial y}{\partial \lambda_i}\,(t,\lambda)$ est la solution de l'équation linéarisée $(\mathrm{E}'_{\lambda_i})$, soit $v' = G(t, v, \lambda)$. Comme f'_{y_j} et f'_{λ_i} sont de classe C^s, on voit que G est de classe C^s ; les dérivées G'_{v_j} et G'_{λ_i} sont donc de classe C^{s-1} et l'hypothèse de récurrence implique que v est de classe C^s. Ceci montre que y est de classe C^{s+1}. ■

1.4. Dépendance de la solution en fonction de la donnée initiale

On désignera ici par $t \mapsto y(t, y_0, \lambda)$ la solution de l'équation différentielle

$$(\mathrm{E}_\lambda) \qquad y' = f(t, y, \lambda)$$

de donnée initiale (t_0, y_0). L'objectif est d'étudier la continuité ou la différentiabilité de $y(t, y_0, \lambda)$ en fonction des 3 variables t, y_0, λ. Ceci résoudra du même coup le cas particulier des solutions $y(t, y_0)$ d'une équation différentielle (E) sans paramètre.

Pour résoudre cette question, on considère la variable y_0 elle-même comme un paramètre et on pose donc $\mu = y_0$. La fonction $z(t, \mu, \lambda) = y(t, \mu, \lambda) - \mu$ satisfait alors la condition initiale $z(t_0, \mu, \lambda) = 0$ et l'équation

$$\begin{aligned}\frac{\partial z}{\partial t}(t, \mu, \lambda) = \frac{\partial y}{\partial t}(t, \mu, \lambda) &= f(t, y(t, \mu, \lambda), \lambda) \\ &= f(t, z(t, \mu, \lambda) + \mu, \lambda).\end{aligned}$$

La fonction $z(t, \mu, \lambda)$ est donc la solution de l'équation différentielle

$$(\mathrm{E}_{\mu,\lambda}) \qquad z' = f(t, z + \mu, \lambda)$$

avec donnée initiale $(t_0, 0)$. Les propriétés cherchées résultent alors des théorèmes déjà démontrés, appliqués à l'équation $(E_{\mu,\lambda})$ pour le paramètre $(\mu, \lambda) \in \mathbb{R}^{m+p}$. On peut donc énoncer :

Théorème – *Si f est de classe C^s et admet des dérivées partielles f'_{y_j} et f'_{λ_i} de classe C^s, alors $y(t, y_0, \lambda)$ est de classe C^{s+1}.*

1.5. Flot d'un champ de vecteurs

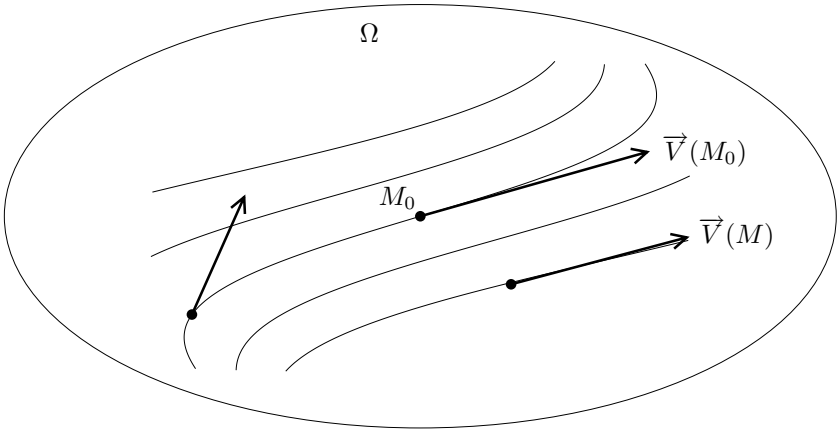
Considérons un ouvert $\Omega \subset \mathbb{R}^m$ et un champ de vecteurs $M \mapsto \overrightarrow{V}(M)$ de classe C^k défini sur Ω, $k \geq 1$. On appelle équation différentielle associée au champ de vecteurs $\overrightarrow{V}$ l'équation différentielle

$$(\mathrm{E}) \qquad \frac{\overrightarrow{dM}}{dt} = \overrightarrow{V}(M).$$

Si nous représentons le point M par ses coordonnées notées $y = (y_1, \ldots, y_m)$, et le champ $\overrightarrow{V}$ par une fonction $\overrightarrow{V}(M) = f(y)$, l'équation devient

$$(\mathrm{E}) \qquad y' = f(y),$$

il s'agit donc tout simplement d'une équation différentielle dont le second membre est indépendant du temps. L'ouvert de définition est dans ce cas l'ouvert produit $U = \mathbb{R} \times \Omega$. Résoudre (E) revient à chercher les *courbes intégrales* (ou *orbites*) du champ de vecteurs.



Dans cette situation, il y a invariance des solutions par translation dans le temps : si $t \mapsto y(t)$ est solution, il en est encore de même pour $t \mapsto y(t+a)$. Pour un point $M_0 = x \in \mathbb{R}^m$ donné, considérons la solution maximale du problème de Cauchy $y(t)$ de donnée initiale $y(0) = x$. D'après le théorème de Cauchy-Lipschitz, la solution maximale est définie un intervalle ouvert contenant le temps 0. On appelle « flot du champ de vecteurs $\overrightarrow{V}$ » l'application

$$\Phi : \mathbb{R} \times \Omega \to \Omega, \qquad (t,x) \mapsto \Phi(t,x) = y(t),$$

c'est-à-dire que $t \mapsto \Phi(t,x)$ est la solution de l'équation

$$\frac{d}{dt}\Phi(t,x) = \overrightarrow{V}(\Phi(t,x)) \qquad \text{avec } \Phi(0,x) = x.$$

La solution maximale n'est en général définie à x fixé que pour un certain intervalle ouvert de temps, de sorte que $\Phi(t,x)$ n'est défini que sur un certain voisinage ouvert Δ de $\{0\} \times \Omega$ dans $\mathbb{R} \times \Omega$. Le fait que le domaine de définition maximal Δ soit ouvert dans $\mathbb{R} \times \Omega$ résulte du § 1.1, et d'après le § 1.4, on sait que Φ est une application de classe C^k sur Δ.

Pour que les solutions soient globales et que $\Delta = U = \mathbb{R} \times \Omega$, un comportement adéquat du champ de vecteurs $\overrightarrow{V}(M)$ au voisinage du bord de Ω ; par exemple, si $\Omega = \mathbb{R}^m$, il suffit, d'après le critère de globalité des solutions du chapitre V § 3.4, que l'on ait une croissance au plus linéaire à l'infini $\|f(y)\| \leq A\|y\| + B$. Pour des raisons qui vont apparaître tout de suite, on note en général $\Phi_t(x)$ le flot plutôt que $\Phi(t,x)$. Une propriété immédiate est alors la loi de groupe

$$(*) \qquad \Phi_t \circ \Phi_s(x) = \Phi_{t+s}(x).$$

Ceci signifie précisément que si l'on suit une trajectoire à partir d'un point x pendant le temps s pour atteindre un point $\Phi_s(x)$, puis, à partir de ce point la même trajectoire pendant le temps t pour atteindre $\Phi_t(\Phi_s(x))$, cela revient au même que de suivre la trajectoire pendant le temps $s+t$. Formellement, on a besoin du théorème d'unicité de Cauchy-Lipschitz pour pouvoir conclure. On voit que $t \mapsto \Phi_t$

est un homomorphisme du groupe additif $(\mathbb{R}, +)$ dans le groupe $\mathrm{Diff}^k(\Omega)$ des C^k difféomorphismes de Ω : on a en effet

$$\Phi_t \circ \Phi_{-t} = \Phi_{-t} \circ \Phi_t = \Phi_0 = \mathrm{Id}_\Omega,$$

par suite Φ_t est un C^k-difféomorphisme d'inverse Φ_{-t}.

Dans le cas où les solutions ne sont plus globales, l'existence globale de $\Phi_t \in \mathrm{Diff}^k(\Omega)$ n'est pas assurée, mais $\Phi_t(x)$ continue à exister en temps petit, et la loi de groupe $(*)$ est encore valable là où les applications sont définies, donc au moins dans un voisinage de $\{0\} \times \Omega$. L'exercice 3.4 donne quelques critères supplémentaires pour que le flot soit globalement défini.

Le théorème de régularité du flot est l'un des points de départ de résultats plus globaux comme le théorème de Poincaré-Bendixson, que nous nos contenterons d'énoncer. Pour plus de détails, nous invitons le lecteur à approfondir la très vivante branche des mathématiques connue aujourd'hui sous le nom de théorie des systèmes dynamiques.

Théorème de Poincaré-Bendixson – *Soit $M \mapsto \overrightarrow{V}(M)$ un champ de vecteurs de classe C^1 dans un ouvert Ω du plan. On suppose qu'il existe un compact $K \subset \Omega$ ne contenant aucun zéro du champ de vecteurs $\overrightarrow{V}$ et stable par le flot Φ_t pour $t \geq 0$ [ce qui implique que $\Phi_t(x)$ est défini pour tout $x \in K$ et tout $t \geq 0$]. Alors K est constitué d'une orbite périodique du flot.*

2. Méthode des petites perturbations

2.1. Description de la méthode

On considère une équation différentielle

$$(\mathrm{E}_\lambda) \qquad y' = f(t, y, \lambda)$$

dépendant d'un paramètre λ ; on prendra pour simplifier $\lambda \in \mathbb{R}$. On suppose que f est continue ainsi que ses dérivées partielles f'_{y_j} et f'_λ. Soit $y(t, \lambda)$ la solution maximale du problème de Cauchy satisfaisant la condition initiale $y(t_0, \lambda) = y_0(\lambda)$; y_0 peut donc dépendre ici de λ ; on supposera que $y_0(\lambda)$ est de classe C^1.

On suppose connue une solution particulière $u(t) = y(t, \lambda_0)$ correspondant à une certaine valeur λ_0 du paramètre. L'objectif est d'étudier les petites perturbations de la solution, c'est-à-dire les solutions $y(t, \lambda)$ associées à des valeurs λ voisines de λ_0. Ceci correspond à une situation physique très courante, dans laquelle on connaît la solution théorique idéale d'un problème et pour laquelle on cherche la solution réelle tenant compte de petites perturbations plus ou moins complexes. En général, on ne sait pas calculer la solution exacte $y(t, \lambda)$ pour $\lambda \neq \lambda_0$. Les résultats des § 1.3, 1.4 montrent que $y(t, \lambda)$ est de classe C^1 et on cherche donc une approximation au premier ordre

$$\begin{aligned} y(t, \lambda) &= y(t, \lambda_0) + (\lambda - \lambda_0) \, \frac{\partial y}{\partial \lambda} \, (t, \lambda_0) + o(\lambda - \lambda_0) \\ &= u(t) + (\lambda - \lambda_0) v(t) + o(\lambda - \lambda_0) \end{aligned}$$

où $v(t) = \frac{\partial y}{\partial \lambda}(t, \lambda_0)$ est à déterminer. On sait que v est la solution de l'équation linéarisée

$$(\mathrm{E}'_{\lambda_0}) \qquad v'(t) = \sum_{j=1}^{m} f'_{y_j}(t, u(t), \lambda_0) v_j(t) + f'_\lambda(t, u(t), \lambda_0),$$

la condition initiale s'écrivant ici

$$v(t_0) = \frac{\partial y}{\partial \lambda}(t_0, \lambda_0) = y'_0(\lambda_0).$$

Comme $(\mathrm{E}'_{\lambda_0})$ est linéaire, la résolution est en principe plus facile que celle de (E_λ). Nous allons maintenant donner des exemples concrets de la méthode.

2.2. *Perturbation d'un champ de vecteurs*

Considérons dans le plan le champ de vecteurs

$$M = (x, y) \mapsto \overrightarrow{V}(M) = (-y, x).$$

Les lignes intégrales du champ $t \mapsto (x(t), y(t))$ sont les solutions du système différentiel

$$\frac{d\overrightarrow{M}}{dt} = \overrightarrow{V}(M) \quad \begin{cases} \dfrac{dx}{dt} = -y \\ \dfrac{dy}{dt} = x. \end{cases}$$

On résout ce système comme au chapitre X, § 2.2(b) en posant $z = x+iy$. Le système se ramène alors à l'équation $dz/dt = iz$, de sorte que la solution du problème de Cauchy avec donnée initiale $z_0 = x_0 + iy_0$ est $z = z_0 e^{it}$. Les lignes de champ sont les cercles centrés en 0.

On suppose maintenant que le champ de vecteurs subit une petite perturbation de la forme

$$M = (x, y) \mapsto \overrightarrow{V_\lambda}(M) = (-y, x) + \lambda(a(x, y), b(x, y))$$

où a, b sont de classe C^1 et où $\lambda \in \mathbb{R}$ est petit. On aboutit donc au système différentiel

$$(\mathrm{E}_\lambda) \qquad \begin{cases} \dfrac{dx}{dt} = -y + \lambda\, a(x, y) \\ \dfrac{dy}{dt} = x + \lambda\, b(x, y) \end{cases} \iff \frac{dz}{dt} = iz + \lambda A(z) \qquad (*)$$

avec $A(z) = a(x, y) + ib(x, y)$. Notons $z(t, \lambda)$ la solution telle que $z(0, \lambda) = z_0$. On sait que $z(t, 0) = z_0 e^{it}$ et on aimerait avoir une approximation de $z(t, \lambda)$ quand λ est petit. Écrivons

$$z(t, \lambda) = z(t, 0) + \lambda \frac{\partial z}{\partial \lambda}(t, 0) + o(\lambda).$$

Grâce à une différentiation de $(*)$ par rapport à λ, il vient

$$\begin{aligned}\frac{\partial}{\partial t}\,\frac{\partial z}{\partial \lambda} &= \frac{\partial}{\partial \lambda}\,\frac{\partial z}{\partial t} = \frac{\partial}{\partial \lambda}\,(iz + \lambda A(z)) \\ &= i\,\frac{\partial z}{\partial \lambda} + A(z) + \lambda\,\frac{\partial}{\partial \lambda}\,(A(z(t,\lambda))).\end{aligned}$$

Par conséquent $v(t) = \frac{\partial z}{\partial \lambda}\,(t,0)$ satisfait l'équation

$$(\mathrm{E}'_0) \qquad \frac{dv}{dt} = iv + A(z(t,0)) = iv + A(z_0 e^{it})$$

avec condition initiale $v(0) = \frac{\partial z}{\partial \lambda}\,(0,0) = 0$. Cette équation se résout par variation des constantes en posant $v(t) = C(t)e^{it}$. Il vient

$$C'(t)e^{it} + iC(t)e^{it} = iC(t)e^{it} + A(z_0 e^{it}),$$

soit $C'(t) = e^{-it}A(z_0 e^{it})$. Comme $C(0) = v(0) = 0$, on obtient

$$\begin{aligned}C(t) &= \int_0^t e^{-iu}A(z_0 e^{iu})du, \\ v(t) = C(t)e^{it} &= \int_0^t e^{i(t-u)}A(z_0 e^{iu})du, \\ z(t,\lambda) = z_0 e^{it} + \lambda &\int_0^t e^{i(t-u)}A(z_0 e^{iu})du + o(\lambda).\end{aligned}$$

Ceci se calcule facilement dès que $a(x,y)$ et $b(x,y)$ sont des polynômes par exemple. Néanmoins, même dans ce cas, il est généralement impossible d'expliciter la solution exacte de l'équation non linéarisée.

2.3. Courbes intégrales d'un champ de vecteurs au voisinage d'un point singulier

Soit $M \mapsto \overrightarrow{V}(M)$ un champ de vecteurs de classe C^s, $s \geq 1$, sur un ouvert Ω du plan. On se place au voisinage d'un point singulier qu'on suppose choisi comme origine des coordonnées pour simplifier. On a donc $\overrightarrow{V}(0) = \overrightarrow{0}$, et comme au chapitre X § 2.1, le système différentiel va s'écrire

$$(\mathrm{E}) \qquad \frac{dM}{dt} = \overrightarrow{V}(M), \qquad \begin{cases} \dfrac{dx}{dt} = ax + by + g(x,y) \\ \dfrac{dy}{dt} = cx + dy + h(x,y) \end{cases}$$

où $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ est la matrice des dérivées partielles $(\overrightarrow{V'_x}(0,0), \overrightarrow{V'_y}(0,0))$ et où g, h s'annulent ainsi que leurs dérivées partielles premières au point $(0,0)$. Les fonctions g, h sont par hypothèse définies sur une certaine boule $\overline{B}(0,r_0)$.

On cherche à comparer l'allure des courbes intégrales situées près de l'origine lorsqu'on effectue des « grossissements successifs » (observation à l'œil nu, à la loupe puis au microscope...). Autrement dit, on veut comparer la courbe issue d'un point M_0 et la courbe issue du point λM_0 (avec λ petit), lorsque ces courbes sont ramenées à la même échelle. Pour grossir la deuxième courbe dans le rapport $1/\lambda$, on effectue le changement de coordonnées $X = x/\lambda$, $Y = y/\lambda$. Dans ces nouvelles coordonnées, l'équation du système devient

$$(\mathrm{E}_\lambda) \qquad \begin{cases} \dfrac{dX}{dt} = aX + bY + \dfrac{1}{\lambda}\, g(\lambda X, \lambda Y) \\ \dfrac{dY}{dt} = cX + dY + \dfrac{1}{\lambda}\, h(\lambda X, \lambda Y). \end{cases}$$

La solution de (E_λ) de valeur initiale (X_0, Y_0) en $t = 0$, correspond à la solution de (E) de valeur initiale $(x_0, y_0) = (\lambda X_0, \lambda Y_0)$. Vu les hypothèses, les fonctions

$$G(X, Y, \lambda) = \frac{1}{\lambda}\, g(\lambda X, \lambda Y)$$
$$H(X, Y, \lambda) = \frac{1}{\lambda}\, h(\lambda X, \lambda Y)$$

sont définies et de classe C^s sur $\overline{B}(0, r_0)\times]0, 1]$ et se prolongent par continuité en $\lambda = 0$ en posant $G(X, Y, 0) = H(X, Y, 0) = 0$. Ce prolongement est en fait de classe C^{s-1} sur $\overline{B}(0, r_0) \times [0, 1]$ car

$$G(X, Y, \lambda) = \int_0^1 (X g'_x(u\lambda X, u\lambda Y) + Y g'_y(u\lambda X, u\lambda Y))du = \left[\frac{1}{\lambda}\, g(u\lambda X, u\lambda Y)\right]_{u=0}^{u=1}.$$

Si $s = 1$, les fonctions G, H sont bien lipschitziennes en X, Y (avec la même constante de Lipschitz que g et h). La solution $(X(t, \lambda), Y(t, \lambda))$ du problème de Cauchy est donc continue, et pour $s \geq 1$ elle est de classe C^{s-1} par rapport à (t, λ), borne $\lambda = 0$ incluse. En particulier :

Proposition – *Lorsque λ tend vers 0, la solution de (E_λ) de valeur initiale (X_0, Y_0) converge vers la solution du système linéaire*

$$(\mathrm{E}) \qquad \begin{cases} \dfrac{dX}{dt} = aX + bY \\ \dfrac{dY}{dt} = cX + dY. \end{cases}$$

Bien entendu, les méthodes précédentes permettent aussi d'avoir un développement limité à l'ordre 1 de la solution de (E_λ) au voisinage de $\lambda = 0$, comme on l'a vu au § 2.2.

2.4. Oscillations du pendule simple

On étudie les oscillations d'un pendule ponctuel de masse m suspendu à un fil de longueur l. L'équation du mouvement a été établie au paragraphe VI 4.2(c) :

$$\theta'' = -\frac{g}{l}\ \sin\,\theta,$$

où θ désigne l'angle fait par le fil avec la verticale. On suppose ici qu'on lache le pendule au temps $t = 0$ à partir de l'élongation maximale $\theta = \theta_m$ avec vitesse initiale nulle. Lorsque θ_m est petit, on fait habituellement l'approximation $\sin\,\theta \simeq \theta$, d'où

$$\theta'' = -\omega^2\theta \quad \text{avec} \quad \omega^2 = \frac{g}{l}.$$

La solution cherchée est alors classiquement

$$\theta(t) = \theta_m \cos\,\omega t.$$

Néanmoins, ceci n'est pas rigoureusement exact, et on aimerait connaître l'erreur commise. Pour cela, on pose $\theta(t) = \theta_m y(t)$, de sorte que y est la solution de l'équation

$$y'' = -\omega^2\ \frac{\sin\,\theta_m y}{\theta_m}$$

avec condition initiales $y(0) = 1$, $y'(0) = 0$. Le développement en série de $\sin\,\theta_m y$ donne

$$\frac{\sin\,\theta_m y}{\theta_m} = y - \frac{1}{6}\ \theta_m^2 y^3 + \frac{1}{120}\ \theta_m^4 y^5 - \ldots + (-1)^n\ \frac{1}{(2n+1)!}\ \theta_m^{2n} y^{2n+1} + \ldots = \varphi(y,\lambda)$$

où $\lambda = \theta_m^2$ et où

$$\varphi(y,\lambda) = y - \frac{1}{6}\ \lambda y^3 + \ldots + (-1)^n\ \frac{1}{(2n+1)!}\ \lambda^n y^{2n+1} + \ldots$$

est une fonction de classe C^∞. La solution $y(t,\lambda)$ de l'équation

$$(\mathrm{E}_\lambda) \qquad y'' = -\omega^2\varphi(y,\lambda)$$

est donc de classe C^∞ (on notera que tous les résultats du paragraphe 1 sont encore vrais pour des équations d'ordre ≥ 2, puisque ces équations sont équivalentes à des systèmes d'ordre 1). On a $\varphi(y,0) = y$ et $y(t,0) = \cos\,\omega t$. Pour obtenir un développement limité de $y(t,\lambda)$ lorsque λ est petit, on dérive (E_λ) par rapport à λ, ce qui donne

$$(\mathrm{E}'_\lambda) \qquad \frac{\partial^2}{\partial t^2}\left(\frac{\partial y}{\partial \lambda}\right) = \frac{\partial}{\partial \lambda}\,\frac{\partial^2 y}{\partial t^2} = \frac{\partial}{\partial \lambda}\ (-\omega^2\varphi(y,\lambda))$$
$$= -\omega^2\Big(\varphi'_y(y,\lambda)\ \frac{\partial y}{\partial \lambda} + \varphi'_\lambda(y,\lambda)\Big).$$

On a $\varphi'_y(y,0) = 1$ et $\varphi'_\lambda(y,0) = -\frac{1}{6}\ y^3$, de sorte que la fonction $v(t) = \frac{\partial y}{\partial \lambda}(t,0)$ satisfait l'équation

$$\begin{aligned} v''(t) &= -\omega^2\Big(v(t) - \frac{1}{6}\ y(t,0)^3\Big) \\ &= -\omega^2 v(t) + \frac{1}{6}\ \omega^2 \cos^3 \omega t. \end{aligned}$$

Comme $y(0,\lambda) = 1$ et $\partial y/\partial t(0,\lambda) = 0$ pour tout λ, les conditions initiales sont $v(0) = v'(0) = 0$. La solution générale du système sans second membre est

$$v(t) = \alpha \cos \omega t + \beta \sin \omega t.$$

On applique alors la méthode de variation des constantes avec

$$v(t) = \alpha(t) \cos \omega t + \beta(t) \sin \omega t.$$

D'après le chapitre VII § 3.3, ceci conduit au système

$$\begin{cases} \alpha'(t) \cos \omega t + \beta'(t) \sin \omega t = 0 \\ \alpha'(t)(-\omega \sin \omega t) + \beta'(t)\, \omega \cos \omega t = \dfrac{1}{6}\, \omega^2 \cos^3 \omega t \end{cases}$$

d'où

$$\begin{cases} \alpha'(t) = -\dfrac{\omega}{6} \cos^3 \omega t \ \sin \omega t \\ \beta'(t) = \dfrac{\omega}{6} \cos^4 \omega t = \dfrac{\omega}{48}\,(3 + 4 \cos 2\omega t + \cos 4\omega t), \end{cases}$$

$$\begin{cases} \alpha(t) = \alpha_0 + \dfrac{1}{24} \cos^4 \omega t \\ \beta(t) = \beta_0 + \dfrac{1}{48}\,(3\omega t + 2 \sin 2\omega t + \dfrac{1}{4} \sin 4\omega t). \end{cases}$$

Les conditions initiales $v(0) = \alpha(0) = 0$ et $v'(0) = \alpha'(0) + \omega\beta(0) = 0$ donnent $\alpha(0) = \beta(0) = 0$, donc $\alpha_0 = -\frac{1}{24}$, $\beta_0 = 0$ et

$$\begin{aligned} v(t) &= \frac{1}{24}\,(\cos^4 \omega t - 1) \cos \omega t + \frac{1}{48}\,(3\omega t + 2 \sin 2\omega t + \frac{1}{4} \sin 4\omega t) \sin \omega t \\ &= \frac{1}{16}\ \Big(\omega t + \frac{1}{6} \sin 2\omega t\Big) \sin \omega t, \\ y(t,\lambda) &= \cos \omega t + \frac{\lambda}{16}\,(\omega t + \frac{1}{6} \sin 2\omega t) \sin \omega t + O(\lambda^2). \end{aligned}$$

Cherchons maintenant à partir de là l'influence de l'élongation maximale θ_m sur la période des oscillations [on a vu au chapitre VI, 2.4 c) que les solutions de faible amplitude étaient périodiques ; ceci n'est pas contradictoire avec le fait que le développement limité de $y(t,\lambda)$ ci-dessus soit non périodique, à cause du terme $O(\lambda^2)$ dépendant de t et lui-même non périodique]. Soit $T(\lambda)$ la période, de sorte que $\frac{1}{4}\ T(\lambda)$ correspond au plus petit $t > 0$ tel que $y(t,\lambda) = 0$, c'est-à-dire $y\Big(\frac{1}{4}\, T(\lambda), \lambda\Big) = 0$. Le théorème des fonctions implicites montre que cette équation définit une fonction $T(\lambda)$ de classe C^∞ pour λ petit, car

$$T(0) = \frac{2\pi}{\omega}, \qquad \frac{\partial y}{\partial t}\ \Big(\frac{1}{4}\ T(0), 0\Big) = -\omega \sin \omega t\Big|_{t=\frac{1}{4}\ T(0)} = -\omega \neq 0.$$

Par différentiation de l'équation en $\lambda = 0$, on trouve de plus

$$\frac{1}{4}\, T'(0)\, \frac{\partial y}{\partial t}\, \Big(\frac{1}{4}\, T(0), 0\Big) + \frac{\partial y}{\partial \lambda}\, \Big(\frac{1}{4}\, T(0), 0\Big) = 0$$

avec

$$\frac{\partial y}{\partial \lambda}\, \Big(\frac{1}{4}\, T(0), 0\Big) = v\Big(\frac{\pi}{2\omega}\Big) = \frac{\pi}{32},$$

d'où

$$\frac{1}{4}\, T'(0)(-\omega) + \frac{\pi}{32} = 0, \qquad T'(0) = \frac{\pi}{8\omega},$$
$$T(\lambda) = T(0) + \lambda T'(0) + O(\lambda^2)$$
$$= \frac{2\pi}{\omega}\, \Big(1 + \frac{1}{16}\, \lambda + O(\lambda^2)\Big).$$

On retrouve ainsi l'approximation bien connue

$$T(\theta_m^2) = 2\pi \sqrt{\frac{l}{g}}\, \left(1 + \frac{1}{16}\, \theta_m^2 + O(\theta_m^4)\right).$$

Cette approximation pourrait également se retrouver de manière directe à partir de la relation exacte

$$T = \sqrt{\frac{8l}{g}} \int_0^{\theta_m} \frac{d\varphi}{\sqrt{\cos\varphi - \cos\theta_m}},$$

qui résulte des formules obtenues au chapitre VI 4.2 c). Exercice pour le lecteur !

3. Problèmes

3.1. On considère une équation différentielle d'ordre p

$$(\mathrm{E}_\lambda) \qquad y^{(p)} = f(t, y, y', \dots, y^{(p-1)}, \lambda)$$

dépendant d'un paramètre λ, où f est définie et continue sur un ouvert U de $\mathbb{R} \times (\mathbb{R}^m)^p \times \mathbb{R}^q$.

(a) On suppose que $f(t, Y, \lambda)$ est de classe C^k sur U et qu'elle admet des dérivées partielles de classe C^k par rapport à chacune des composantes de Y et λ. On note

$$y(t, y_0, y_1, \dots, y_{p-1}, \lambda)$$

la solution de (E_λ) satisfaisant les conditions initiales

$$y(t_0) = y_0, \quad y'(t_0) = y_1, \dots, y^{(p-1)}(t_0) = y_0.$$

Montrer que cette solution y est de classe C^{k+1} par rapport à l'ensemble des variables t, y_j, λ, de même que ses dérivées partielles $\partial y/\partial t, \dots, \partial^{p-1} y/\partial t^{p-1}$.

[*Indication :* se ramener au cas d'un système d'ordre 1].

(b) On suppose ici que $\lambda \in \mathbb{R}$. Soit $u(t,\lambda)$ la solution satisfaisant la condition initiale

$$\frac{\partial^k u}{\partial t^k}(t_0,\lambda) = y_k(\lambda), \quad 0 \leq k \leq p-1$$

où $y_0(\lambda), \ldots, y_{p-1}(\lambda)$ sont de classe C^1 au moins en λ. Montrer que $v(t,\lambda) = \partial u/\partial\lambda(t,\lambda)$ est la solution d'une équation différentielle linéaire d'ordre p dont on précisera les conditions initiales.

(c) Application : Écrire l'équation du (b) pour $y'' = e^{\lambda t}y^2 + \lambda y'$, avec les conditions initiales $y(0) = e^{-\lambda}$, $y'(0) = \cosh 2\lambda$.

3.2. On s'intéresse ici au comportement des courbes intégrales passant par un point voisin de l'origine, pour le système (S) de l'exercice X 3.2. Pour tout $\lambda > 0$, on note

$$t \mapsto M(t,\lambda) = (x(t,\lambda), y(t,\lambda))$$

la solution maximale de (S) qui passe par le point de coordonnées $(\lambda, 0)$ au temps $t = 0$.

(a) Déterminer la solution $t \mapsto (\widetilde{x}(t), \widetilde{y}(t))$ du système *linéarisé* au voisinage de l'origine, qui passe par le point $(1,0)$ au temps $t = 0$.

(b) A l'aide d'une homothétie convenable et des méthodes du chap. XI, montrer que $M(t,\lambda)$ admet un développement limité de la forme

$$M(t,\lambda) = (\lambda\widetilde{x}(t) + \lambda^2 u(t) + O(\lambda^3),\ \lambda\widetilde{y}(t) + \lambda^2 v(t) + O(\lambda^3))$$

où u, v sont des fonctions que l'on explicitera.

3.3. L'objet de ce problème est d'étudier les lignes de champ créées dans un plan par un dipôle électrique (par exemple une molécule polarisée telle que le chlorure d'hydrogène).

Le plan est rapporté au repère orthonormé direct $(O\,;\vec{i},\vec{j})$, où O est la position du dipôle (supposé ponctuel), et $(O\,;\vec{i})$ l'axe du dipôle. Si M est un point quelconque distinct de O, on note (r,θ) les coordonnées polaires de M relativement au repère $(O\,;\vec{i},\vec{j})$. On associe à M le vecteur radial $\vec{u} = \cos\theta\cdot\vec{i} + \sin\theta\cdot\vec{j}$ et le vecteur orthoradial $\vec{v} = -\sin\theta\cdot\vec{i} + \cos\theta\cdot\vec{j}$. On admettra que le potentiel électrique $V(M)$ créé par le dipôle en tout point $M \neq O$ est donné par

$$V(M) = \frac{\cos\theta}{r^2}.$$

(a) On rappelle les formules

$$d\overrightarrow{M} = dr\cdot\vec{u} + r d\theta\cdot\vec{v},$$
$$\overrightarrow{\operatorname{grad}}\, V = \frac{\partial V}{\partial r}\cdot\vec{u} + \frac{1}{r}\,\frac{\partial V}{\partial\theta}\cdot\vec{v}.$$

Évaluer le champ électrique $\overrightarrow{E} = -\overrightarrow{\text{grad}}\, V$ créé par le dipôle.

Déterminer l'équation $r = \varphi(\theta)$ de la ligne de champ (courbe intégrale du champ de vecteurs $\overrightarrow{E}$) passant par le point de coordonnées polaires (r_0, θ_0) avec $r_0 > 0$ et $\theta_0 = \frac{\pi}{2}$.

(b) Le dipôle est supposé placé dans un champ électrique ambiant constant $\overrightarrow{E}_0 = \lambda\vec{j}$, d'intensité très faible par rapport à son champ propre $\overrightarrow{E}$.

(α) Écrire l'équation différentielle $\frac{dr}{d\theta} = f(r, \theta, \lambda)$ des lignes de champ relatives au champ $\overrightarrow{E} + \overrightarrow{E}_0$. Calculer le développement limité à l'ordre 1 de $f(r, \theta, \lambda)$ en fonction de λ.

(β) On note $r = \psi(\theta, \lambda)$ l'équation polaire de la ligne de champ passant par le point $\left(r_0, \frac{\pi}{2}\right)$ [on ne cherchera pas à évaluer ψ].

Montrer que $w(\theta) = \frac{\partial \psi}{\partial \lambda}\,(\theta, 0)$ satisfait à une équation différentielle linéaire.

En déduire le développement limité à l'ordre 1 de $\psi(\theta, \lambda)$ en fonction de λ.

3.4. Soit Ω un ouvert de $\mathbb{R}^m$ et $M \mapsto \overrightarrow{V}(M)$ un champ de vecteurs de classe C^1 sur Ω. On considère le flot associé à l'équation différentielle

$$\text{(E)} \qquad \frac{\overrightarrow{dM}}{dt} = \overrightarrow{V}(M).$$

Montrer que toute solution maximale de (E) est globale dans les trois cas suivants :

(a) $\Omega = \mathbb{R}^m$ et $\overrightarrow{V}$ satisfait à l'infini la condition $\overrightarrow{V}(M) \cdot \overrightarrow{OM} \leq \|\overrightarrow{OM}\|\, \varphi(\|\overrightarrow{OM}\|)$ où $\varphi : [0, +\infty[\to \mathbb{R}$ est une fonction continue, croissante et positive telle que $\int_1^{+\infty} dt/\varphi(t) = +\infty$.
Indication: majorer $v(\|\overrightarrow{OM}\|)$ où $v(t) = \int_1^t du/\varphi(u)$ à l'aide d'un raisonnement de type lemme de Gronwall.

(b) Ω est un ouvert borné et on a la condition $\|\overrightarrow{V}(M)\| \leq \varphi(d(M, \partial\Omega))$ avec $\varphi :]0, +\infty[\to \mathbb{R}$ continue, croissante et positive telle que $\int_0^1 dt/\varphi(t) = +\infty$ (et donc telle que $\lim_{t \to 0_+} \varphi(t) = 0$).
Indication: majorer $v(d(M, \partial\Omega))$ où $v(t) = \int_t^1 du/\varphi(u)$.

(c) Ω est un ouvert borné à bord régulier de classe C^1, et $M \mapsto \overrightarrow{V}(M)$ se prolonge en champ de vecteurs de classe C^1 sur $\overline{\Omega}$ tel que $\overrightarrow{V}(M)$ soit tangent à $\partial\Omega$ en tout point du bord.

Références

Arnold V. (1974) – Equations différentielles ordinaires, Editions de Moscou.

Artigue M., Gautheron V. (1983) – Systèmes différentiels, Etude Graphique, Cedic/Fernand Nathan, Paris.

Cartan H. (1977) – Cours de calcul différentiel, nouvelle édition refondue et corrigée, Hermann, Paris.

Coddington E.A., Levinson N. (1955) – Theory of Ordinary Differential Equations, Mac Graw-Hill, New-York.

Crouzeix M., Mignot A.L. (1984) – Analyse numérique des équations différentielles, Masson, Paris.

Crouzeix M., Mignot A.L. (1986) – Exercices d'analyse numérique des équations différentielles, Masson, Paris.

Dieudonné J. (1968) – Calcul infinitésimal, Hermann, Paris.

Fröberg C.E. (1965) – Introduction to numerical analysis, Addison-Wesley, Reading.

Gastinel N. (1966) – Analyse numérique linéaire, Hermann, Paris.

Hartmann P. (1964) – Ordinary differential equations, John Wiley, New-York.

Hildebrand F.B. (1956) – Introduction to numerical analysis, Mac Graw-Hill, New York.

Hirsch W., Smale S. (1974) – Differential equations, dynamical systems and linear algebra, Academic Press, New York.

Reinhard H. (1989) – Equations différentielles. Fondements et applications, Dunod, Paris, deuxième édition.

Roseau M. (1976) – Equations différentielles, Masson, Paris.

Rouche N., Mawhin J. (1973) – Equations différentielles ordinaires, tomes 1, 2, Masson, Paris.

Sibony M., Mardon J.L. (1982) – Analyse numérique, tomes 1, 2, Hermann, Paris.

Stiefel E. (1965) – An introduction to numerical Mathematics, Academic Press, New York.

Index terminologique

Index des notations

Formulaire et principaux résultats

Chapitre I : Calculs numériques approchés

Soit ε la précision relative de l'expression approchée des nombres réels sur le calculateur utilisée. Les erreurs d'arrondi sur les opérations arithmétiques vérifient

$$\Delta(x+y) \leq \Delta x + \Delta y + \varepsilon(|x|+|y|),$$
$$\Delta(xy) \leq |x|\Delta y + |y|\Delta x + \varepsilon|xy|.$$

Si la sommation $s_n = x_1 + x_2 + \ldots + x_n$ est calculée par ordre croissant d'indice et si $\Delta x_i = 0$, alors

$$\Delta(s_n) \leq \varepsilon(|x_n| + 2|x_{n-1}| + \ldots + (n-1)|x_2| + (n-1)|x_1|),$$
$$\Delta(x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}) \leq \varepsilon(|\alpha_1| + \ldots + |\alpha_n| - 1)|x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}|.$$

Chapitre II : Approximation polynomiale des fonctions numériques

Polynôme d'interpolation de Lagrange de $f : [a,b] \to \mathbb{R}$ en des points $x_0, x_1, \ldots, x_n \in [a,b]$:

$$p_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x), \qquad l_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

Formule d'erreur :

$$f(x) - p_n(x) = \frac{1}{(n+1)!}\ \pi_{n+1}(x) f^{(n+1)}(\xi_x)$$

avec $$\pi_{n+1}(x) = \prod_{i=0}^{n} (x - x_i), \quad \xi_x \in \,]\min(x, x_i), \max(x, x_i)[.$$

Différences divisées :

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0},$$

$$p_n(x) = f(x_0) + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k](x - x_0) \ldots (x - x_{k-1}).$$

Formule de Newton (pas constant $h = \frac{b-a}{n}$) :

$$x_i = a + ih, \quad f_i = f(x_i), \quad \Delta^k f_i = \Delta^{k-1} f_{i+1} - \Delta^{k-1} f_i,$$

$$p_n(x) = \sum_{k=0}^{n} \Delta^k f_0 \, \frac{s(s-1)\ldots(s-k+1)}{k!} \quad \text{où} \quad x = a + sh.$$

Polynômes de Tchebychev :

$$t_n(x) = \cos(n \operatorname{Arc\,cos} x), \quad x \in [-1, 1],$$

$$\begin{cases} t_0(x) = 1, \quad t_1(x) = x \\ t_{n+1}(x) = 2x\, t_n(x) - t_{n-1}(x), \quad n \geq 1. \end{cases}$$

Points d'interpolation de Tchebychev (= racines de t_{n+1}) :

$$x_i = \cos \frac{2i+1}{2n+2}\,\pi, \qquad 0 \leq i \leq n.$$

Estimation de $|\pi_{n+1}(z)|$, $z \in \mathbb{C}$ avec $x_i = a + ih$, $h = \frac{b-a}{n}$:

On pose $\delta_n(z) = \min |z - x_i|$ et $A(z) = \exp\left(\frac{1}{b-a} \int_a^b \ln|z - x| dx\right)$. Alors il existe des constantes $C_1, C_2 > 0$ indépendantes de n telles que

$$C_1 \delta_n(z) A(z)^n \leq |\pi_{n+1}(z)| \leq C_2 n \delta_n(z) A(z)^n.$$

Meilleure approximation uniforme : si $f \in \mathcal{C}([a, b])$, le polynôme q_n de degré n minimisant la distance uniforme $\|f - q_n\|$ existe et est unique. Il est caractérisé par la propriété que $f - q_n$ équioscille sur au moins $n + 2$ points de $[a, b]$.

Théorème de Jackson : soit $f \in \mathcal{C}([a, b])$. Si ω_f est le module de continuité de f et si p_n est le polynôme d'approximation de Jackson de f de degré n, on a

$$\|f - p_n\| \leq 3\,\omega_f\left(\frac{b-a}{n}\right).$$

Constante de Lebesgue : $\Lambda_n = \sup_{x \in [a,b]} \sum_{i=0}^{n} |l_i(x)|$.

Si $L_n(f) = p_n = \sum_{i=0}^{n} f(x_i) l_i$, on a

$$\|f - L_n(f)\| \leq (1 + \Lambda_n)\|f - q_n\|.$$

Si les x_i sont équidistants, on a $\Lambda_n \sim \frac{2^{n+1}}{en \ln(n)}$.
Si les x_i sont les points de Tchebychev, on a $\Lambda_n \sim \frac{2}{\pi} \ln(n)$.

Polynômes orthogonaux. Formule de récurrence :

$$p_n(x) = (x - \lambda_n)p_{n-1}(x) - \mu_n p_{n-2}(x), \quad n \geq 2$$

avec $\lambda_n = \langle x p_{n-1}, p_{n-1}\rangle / \|p_{n-1}\|_2^2$, $\mu_n = \|p_{n-1}\|_2^2 / \|p_{n-2}\|_2^2$, $1 \leq j \leq l$.

Polynômes de meilleure approximation quadratique :

$$r_n(x) = \sum_{k=0}^{n} \frac{\langle f, p_k\rangle}{\|p_k\|_2^2} \, p_k(x).$$

Chapitre III : Intégration numérique

Méthodes de Newton-Cotes d'indice l :

$$\int_\alpha^\beta f(x)dx \simeq \sum_{i=0}^{k-1} (\alpha_{i+1} - \alpha_i) \sum_{j=0}^{l} \omega_j f(\xi_{i,j})$$

avec $\xi_{i,j} = \alpha_i + j \cdot \frac{\alpha_{i+1} - \alpha_i}{l}$, $1 \leq j \leq l$.

- $l = 1$: méthodes des trapèzes $\omega_0 = \omega_1 = \frac{1}{2}$ (ordre 1).
  erreur : $-\frac{1}{12} h^2 f''(\xi)(\beta - \alpha)$ si le pas est constant.
- $l = 2$: méthode de Simpson $\omega_0 = \omega_2 = \frac{1}{6}$, $\omega_1 = \frac{4}{6}$ (ordre 3).
  erreur : $-\frac{1}{2880} h^4 f^{(4)}(\xi)(\beta - \alpha)$.
- $l = 4$: méthode de Boole-Villarceau
  $\omega_0 = \omega_4 = \frac{7}{90}$, $\omega_1 = \omega_3 = \frac{16}{45}$, $\omega_2 = \frac{2}{15}$ (ordre 5).
  erreur : $-\frac{1}{1\,935\,360} h^6 f^{(6)}(\xi)(\beta - \alpha)$.

Formule de Taylor avec reste intégral :

$$f(x) = \sum_{k=0}^{N} \frac{1}{k!} f^{(k)}(\alpha)(x - \alpha)^k + \int_\alpha^\beta \frac{1}{N!} (x - t)_+^N f^{(N+1)}(t)dt.$$

Noyau de Peano d'une méthode d'ordre N :

Si $E(f) = \displaystyle\int_\alpha^\beta f(x)w(x)dx - \sum_{j=0}^{l} \lambda_j f(x_j)$ est l'erreur d'intégration, alors

$$K_N(t) = E(x \mapsto (x-t)_+^N), \quad t \in [\alpha, \beta],$$
$$E(f) = \frac{1}{N!} \int_\alpha^\beta K_N(t) f^{(N+1)}(t)dt.$$

Si K_N est de signe constant, alors

$$E(f) = \frac{1}{N!}\, f^{(N+1)}(\xi) \int_\alpha^\beta K_N(t)dt, \quad \xi \in]\alpha, \beta[.$$

Noyau de Peano d'une méthode composée. Si la méthode élémentaire est d'ordre N et admet k_N pour noyau de Peano sur $[-1, 1]$, le noyau de Peano composé est donné par

$$K_N(t) = \left(\frac{h_j}{2}\right)^{N+1} k_N\left(\frac{2}{h_j}\left(t - \frac{\alpha_j + \alpha_{j+1}}{2}\right)\right), \quad t \in [\alpha_j, \alpha_{j+1}].$$

Méthodes de Gauss :

$$\int_\alpha^\beta f(x)w(x)dx \simeq \sum_{j=0}^{l} \lambda_j f(x_j)$$

avec $x_0, \ldots, x_l \in]\alpha, \beta[$ racines du polynôme orthogonal p_{l+1}

$$\lambda_i = \int_\alpha^\beta L_i(x)w(x)dx, \quad L_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

La méthode est d'ordre $N = 2l + 1$, et l'erreur est donnée par

$$E(f) = \frac{f^{(2l+2)}(\xi)}{(2l+2)} \int_\alpha^\beta \pi_{l+1}(x)^2 w(x)dx.$$

Polynômes de Bernoulli : ils sont définis par récurrence par

$$\begin{cases} B_1(x) = x - \frac{1}{2} \quad \text{si} \quad x \in [0, 1[, \\ B_p' = pB_{p-1}(x) \quad \text{sur} \quad [0, 1[\quad \text{pour} \quad p \geq 2, \\ \displaystyle\int_0^1 B_p(x)dx = 0, \\ B_p \quad \text{périodique de période 1 sur } \mathbb{R}. \end{cases}$$

$$B_p(x) = -p! \sum_{n \in \mathbb{Z}^*} \frac{e^{2\pi i n x}}{(2\pi i n)^p}.$$

Nombres de Bernoulli :

$$b_0 = 1, \quad b_1 = -\frac{1}{2}, \quad b_p = B_p(0) \quad \text{si} \quad p \geq 2.$$

$$\begin{cases} b_{2k} = \dfrac{2(-1)^{k-1}(2k)!}{(2\pi)^{2k}} \displaystyle\sum_{n=1}^{+\infty} \frac{1}{n^{2k}} & \text{si} \quad k \geq 1, \\ b_{2k+1} = 0 & \text{si} \quad k \geq 1. \end{cases}$$

$$|B_{2k}(x)| \leq |b_{2k}|,$$

$$B_p(x) = \sum_{m=0}^{p} C_p^m b_m x^{p-m}, \quad B_p(1-x) = (-1)^p B_p(x),$$

$$C_{2k+1}^{2k} b_{2k} + C_{2k+1}^{2k-2} b_{2k-2} + \ldots + C_{2k+1}^{2} b_2 + C_{2k+1}^{1} b_1 + 1 = 0,$$

$$b_2 = \frac{1}{6}, \quad b_4 = -\frac{1}{30}, \quad b_6 = \frac{1}{42}, \quad b_8 = -\frac{1}{30}, \quad b_{10} = \frac{5}{66}.$$

Formule d'Euler-Maclaurin : Pour $f \in C^{2k}([\alpha, \beta])$, avec $\alpha, \beta \in \mathbb{Z}$, on a

$$\frac{1}{2}\, f(\alpha) + f(\alpha+1) + \ldots + f(\beta-1) + \frac{1}{2}\, f(\beta) = \int_\alpha^\beta f(x)dx + \sum_{m=1}^{k} \frac{b_{2m}}{(2m)!} \left(f^{(2m-1)}(\beta) - f^{(2m-1)}(\alpha) \right) - \int_\alpha^\beta \frac{B_{2k}(x)}{(2k)!}\, f^{(2k)}(x)dx.$$

Formule du développement asymptotique. Soit $f \in C^\infty([\alpha, +\infty[)$, $\alpha \in \mathbb{Z}$. Si $\lim_{x \to +\infty} f^{(m)}(x) = 0$ et si $f^{(m)}(x)$ est de signe constant sur $[x_0, +\infty[$ pour $m \geq m_0$ alors $\forall n \geq x_0$ et $\forall k > \frac{m_0}{2}$ on a

$$f(\alpha) + f(\alpha+1) + \ldots + f(n) = C + \frac{1}{2}\, f(n) + \int_\alpha^n f(x)dx + \sum_{m=1}^{k-1} \frac{b_{2m}}{(2m)!}\, f^{(2m-1)}(n) + \theta\, \frac{b_{2k}}{2k!}\, f^{(2k-1)}(n), \qquad 0 \leq \theta \leq 1.$$

Extrapolation de Richardson

Pour calculer la limite de $A(t) = a_0 + a_1 t + \ldots + a_k t^k + O(t^{k+1})$ en $t = 0$, on pose

$$A_{m,0} = A(r^{-m} t_0),$$
$$A_{m,n} = \frac{r^n A_{m,n-1} - A_{m-1,n-1}}{r^n - 1}.$$

Méthode de Romberg : pour évaluer $\displaystyle\int_\alpha^\beta f(x)dx$, on calcule

$$A_{m,0} = h\Big(\frac{1}{2}\, f(\alpha) + f(\alpha+h) + \ldots + f(\beta-h) + \frac{1}{2}\, f(\beta) \Big)$$

où $h = \frac{\beta-\alpha}{2^m}$, puis

$$A_{m,n} = \frac{4^n A_{m,n-1} - A_{m-1,n-1}}{4^n - 1}.$$

Chapitre IV : Méthodes itératives pour la résolution d'équations

Méthode de Newton : pour résoudre $f(x) = 0$, on itère

$$\varphi(x) = x - \frac{f(x)}{f'(x)}.$$

Si $f(a) = 0$ et $M = \max\limits_{x\in[a-r,a+r]} \left|\frac{f''(x)}{f'(x)}\right|$, alors pour $h = \min\left(1, \frac{1}{M}\right)$ on a

$$(\forall x \in [a-h, a+h]) \qquad |\varphi(x) - a| \leq M|x-a|^2.$$

Variante (méthode de la sécante) : A partir de valeurs initiales x_0, x_1, on pose

$$\begin{cases} \tau_p = \frac{f(x_p)-f(x_{p-1})}{x_p - x_{p-1}} \\ x_{p+1} = x_p - \frac{f(x_p)}{\tau_p} \end{cases}, \quad (\forall p \geq 1).$$

Méthode de Newton-Raphson. Pour résoudre $f(x) = 0$ où $f : \mathbb{R}^m \to \mathbb{R}^m$, on itère la fonction

$$\varphi(x) = x - f'(x)^{-1} \cdot f(x).$$

Chapitre V : Équations différentielles. Résultats fondamentaux

Étant donné un ouvert $U \subset \mathbb{R} \times \mathbb{R}^m$ et une fonction continue $f : U \to \mathbb{R}^m$, il passe par tout point $(t_0, y_0) \in U$ au moins une solution locale de l'équation différentielle

$$y' = f(t, y). \tag{E}$$

De plus toute solution peut être prolongée en une solution maximale ; l'intervalle de définition d'une solution maximale est ouvert.

Méthode d'Euler. Partant d'un point initial (t_0, y_0), on pose

$$\begin{cases} y_{n+1} = y_n + h_n f(t_n, y_n) \\ t_{n+1} = t_n + h_n \end{cases}$$

et on construit une solution approchée $\underline{y}(t)$ linéaire par morceaux en joignant les points (t_n, y_n). Si $C = [t_0 - T, t_0 + T] \times \overline{B}(y_0, r_0)$ est un cylindre de sécurité tel que

$$T \leq \min\left(T_0, \frac{r_0}{M}\right), \quad \text{où} \quad M = \sup_{[t_0-T_0, t_0+T_0]\times\overline{B}(y_0,r_0)} \|f\|,$$

la solution approchée $(t, y(t))$ reste contenue dans C. L'erreur vérifie

$$\|y'(t) - f(t, y(t))\| \leq \omega_f((M+1)h_{\max})$$

où ω_f est un module de continuité pour f.

Théorème de Cauchy-Lipschitz. Si f est localement lipschitzienne en y, il passe par tout point (t_0, y_0) une solution maximale unique et toute suite de solutions ε_p-approchées avec $\lim \varepsilon_p = 0$ converge vers la solution exacte : le lemme de Gronwall montre que l'écart entre 2 telles solutions approchées y_1, y_2 vérifie

$$\|y_1(t) - y_2(t)\| \leq (\varepsilon_1 + \varepsilon_2)\ \frac{e^{k|t-t_0|} - 1}{k}$$

où k est la constante de Lipschitz.

Équations différentielles d'ordre p : $y^{(p)} = f(t, y, y', \ldots, y^{(p-1)})$.
Si f est continue (resp. continue et localement lipschitzienne en toutes les variables autres que t), il existe au moins une solution (resp. une unique solution) satisfaisant la condition initiale

$$y(t_0) = y_0, \quad y'(t_0) = y_1, \ldots, y^{(p-1)}(t_0) = y_{p-1}.$$

Chapitre VI : Méthodes de résolution explicite des équations différentielles

Équations à variables séparées : $y' = f(x)g(y)$.
Écrire $\frac{dy}{g(y)} = f(x)dx$.

Équations linéaires du premier ordre : $y' = a(x)y + b(x)$.
Solution générale : $y(x) = \lambda e^{A(x)} + y_{(1)}(x)$ où A est une primitive de a et où $y_{(1)}(x)$ s'obtient par la méthode de variation des constantes, $y_{(1)}(x) = \lambda(x)e^{A(x)}$.

Équations de Bernoulli : $y' + p(x)y + q(x)y^\alpha = 0$ où $\alpha \in \mathbb{R} \setminus \{1\}$.
Diviser par y^α et poser $z(x) = y(x)^{1-\alpha}$. Alors z satisfait une équation linéaire.

Équations de Riccati : $y' = a(x)y^2 + b(x)y + c(x)$.
Si une solution particulière $y_{(1)}$ est connue, poser $y = y_{(1)} + z$. Alors $\frac{1}{z}$ satisfait une équation linéaire.

Équations homogènes : $y' = f\left(\frac{y}{x}\right)$.
Poser $y(x) = xz(x)$, ou passer en coordonnée polaires.

Équations non résolues en y' : regarder si on peut trouver une paramétrisation simple de l'équation.

Équations de Lagrange : $y = a(y')x + b(y')$.
Choisir $p = y'$ comme nouvelle variable et $x(p)$ comme nouvelle fonction inconnue. Calculer dy et écrire $dy = p\,dx$.

Équations du second ordre : $y'' = f(y, y')$: choisir y comme nouvelle variable et $v = y'$ comme nouvelle fonction inconnue de y. On a alors $y'' = v\,dv/dy$.

Équations linéaires homogènes du second ordre : $a(x)y'' + b(x)y' + c(x)y = 0$.
Si une solution particulière $y_{(1)}$ est connue, utiliser la méthode de variation des constantes : $y(x) = \lambda(x)y_{(1)}(x)$. Alors $\mu = \lambda'$ est solution d'une équation du premier ordre.

Chapitre VII : Systèmes différentiels linéaires

Systèmes différentiels linéaires à coefficients constants dans $\mathbb{R}^m$
Solution du problème de Cauchy $Y(t_0) = V_0$ pour

$$Y' = AY \ : \qquad Y(t) = e^{(t-t_0)A} \cdot V_0$$

$$Y' = AY + B(t) \ : \qquad Y(t) = e^{(t-t_0)A} \cdot V_0 + \int_{t_0}^{t} e^{(t-u)A} B(u) du$$

On a $\det(e^A) = \exp(\operatorname{tr} A)$.

Équations d'ordre p : $a_p y^{(p)} + \ldots + a_1 y' + a_0 y = 0$, $a_j \in \mathbb{C}$.
Si le polynôme caractéristique $a_p\lambda^p + \ldots + a_1\lambda + a_0$ admet pour racines complexes $\lambda_1, \ldots, \lambda_s$ de multiplicités $m_1, \ldots, m_s$, alors l'espace des solutions admet pour base

$$t \mapsto t^q e^{\lambda_j t}, \quad 1 \leq j \leq s, \quad 0 \leq q < m_j.$$

Systèmes différentiels linéaires quelconques $Y' = A(t)Y + B(t)$.
Si $R(t, t_0)$ désigne la résolvante, alors la solution du problème de Cauchy $Y(t_0) = V_0$ est donnée par

$$Y(t) = R(t, t_0) \cdot V_0 + \int_{t_0}^{t} R(t, u)B(u)du,$$

et on a $\det(R(t, t_0)) = \exp\Big(\int_{t_0}^{t} \text{tr } A(u)du\Big)$.

Chapitre VIII : Méthodes numériques à un pas

Ces méthodes peuvent s'écrire de manière générale

$$y_{n+1} = y_n + h_n\Phi(t_n, y_n, h_n), \quad 0 \leq n < N.$$

Si z est une solution exacte, l'*erreur de consistance* relative à z est par définition $e_n = z(t_{n+1}) - y_{n+1}$ pour $y_n = z(t_n)$. La méthode est dite *d'ordre* $\geq p$ s'il existe une constante C indépendante de n et $h_{\max}$ telle que $|e_n| \leq Ch_n h_{\max}^p$. Pour qu'il en soit ainsi, il faut et il suffit que

$$\frac{\partial^l \Phi}{\partial h^l}\,(t, y, 0) = \frac{1}{l+1}\, f^{[l]}(t, y), \quad 0 \leq l \leq p-1.$$

La méthode est dite *stable de constante de stabilité* S si pour toute suite perturbée $\widetilde{y}_n$ telle que

$$\widetilde{y}_{n+1} = \widetilde{y}_n + h_n\Phi(t_n, \widetilde{y}_n, h_n) + \varepsilon_n, \quad 0 \leq n < N,$$

alors $\max\limits_{0\leq n\leq N} |\widetilde{y}_n - y_n| \leq S\left(|\widetilde{y}_0 - y_0| + \sum\limits_{0\leq n<N} |\varepsilon_n|\right)$.
Sous cette hypothèse, l'*erreur globale* admet la majoration

$$\begin{aligned}\max_{0\leq n\leq N} |y_n - z(t_n)| &\leq S\left(|y_0 - z(t_0)| + \sum_{0\leq n<N} |e_n|\right), \\ &\leq S\left(|y_0 - z(t_0)| + CTh_{\max}^p\right).\end{aligned}$$

Lemme de Gronwall. Si $\theta_{n+1} \leq (1 + \Lambda h_n)\theta_n + |\varepsilon_n|$ avec $h_n, \theta_n \geq 0$ et $\varepsilon_n \in \mathbb{R}$, alors

$$\begin{aligned}\theta_n &\leq e^{\Lambda(t_n - t_0)}\theta_0 + \sum_{0\leq i\leq n-1} e^{\Lambda(t_n - t_{i+1})}|\varepsilon_i| \\ &\leq e^{\Lambda(t_n - t_0)}\Big(\theta_0 + \sum_{0\leq i\leq n-1} |\varepsilon_i|\Big).\end{aligned}$$

Si $\Phi(t,y,h)$ est Λ-lipschitzienne en y, la méthode est stable avec constante de stabilité $S = \exp(\Lambda T)$, $T = t_N - t_0$.

Méthode du point milieu :

$$\begin{cases} y_{n+\frac{1}{2}} = y_n + \dfrac{h_n}{2}\, f(t_n, y_n) \\ p_n = f\Big(t_n + \dfrac{h_n}{2}, y_{n+\frac{1}{2}}\Big) \\ y_{n+1} = y_n + h_n p_n \\ t_{n+1} = t_n + h_n \end{cases}$$

du point milieu modifiée :

$$\begin{cases} y_{n+\frac{1}{2}} = y_n + \dfrac{h_n}{2}\, p_{n-1} \\ p_n = f\Big(t_n + \dfrac{h_n}{2}, y_{n+\frac{1}{2}}\Big) \\ y_{n+1} = y_n + h_n p_n \\ t_{n+1} = t_n + h_n \end{cases}$$

Ces méthodes sont d'ordre 2 (celle du point milieu est à 1 pas, mais la méthode modifiée est une méthode à 2 pas).

Méthodes de Runge-Kutta :

$$\begin{array}{c|cccccc} c_1 & 0 & 0 & \dots & 0 & 0 \\ c_2 & a_{21} & 0 & \dots & 0 & 0 \\ \vdots & & & & & \\ c_q & a_{q1} & a_{q2} & \dots & a_{qq-1} & 0 \\ \hline & b_1 & b_2 & \dots & b_{q-1} & b_q \end{array}$$

$$\begin{cases} \left[\begin{array}{l} t_{n,i} = t_n + c_i h_n \\ y_{n,i} = y_n + h_n \displaystyle\sum_{1\le j<i} a_{ij} p_{n,j} \\ p_{n,i} = f(t_{n,i}, y_{n,i}) \end{array}\right] 1 \le i \le q \\ t_{n+1} = t_n + h_n \\ y_{n+1} = y_n + h_n \displaystyle\sum_{1\le j\le q} b_j p_{n,j} \end{cases}$$

On a toujours $\displaystyle\sum_{1\le j<i} a_{ij} = c_i$, $\displaystyle\sum_{1\le j\le q} b_j = 1$.

$*$ Constante de stabilité : $S = \exp(\Lambda T)$ où k = constante de Lipschitz de f et

$$\Lambda = k \sum_{1\le j\le q} |b_j| \left(1 + (\alpha k h_{\max}) + \ldots + (\alpha k h_{\max})^{j-1}\right), \quad \alpha = \max_i \sum_j |a_{ij}|.$$

$*$ L'ordre est $\ge p$ si et seulement si les coefficients satisfont les relations

$p \ge 2$: $\quad \displaystyle\sum b_j c_j = \frac{1}{2}$

$p \ge 3$: $\quad \displaystyle\sum b_j c_j = \frac{1}{2}$; $\displaystyle\sum b_j c_j^2 = \frac{1}{3}$; $\displaystyle\sum_{i,j} b_i a_{ij} c_j = \frac{1}{6}$

$p \ge 4$: $\quad \displaystyle\sum b_j c_j^3 = \frac{1}{4}$; $\displaystyle\sum_{i,j} b_i a_{ij} c_j^2 = \frac{1}{12}$; $\displaystyle\sum_{i,j} b_i c_i a_{ij} c_j = \frac{1}{8}$; $\displaystyle\sum_{i,j,k} b_i a_{ij} a_{jk} c_k = \frac{1}{12}$

en plus des conditions des ordres 2 et 3.

∗ Exemples :

Ordre 2 :

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}$$

$\alpha = \frac{1}{2}$: point milieu

$\alpha = 1$: méthode de Heun

Ordre 4 :

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$$

Chapitre IX : Méthodes à pas multiples

Pour une méthode à $r+1$ pas de la forme

$$y_{n+1} = \Psi(t_n, y_n, h_n\,; \ldots ;\, t_{n-r}, y_{n-r}, h_{n-r}), \quad r \le n < N,$$

l'erreur de consistance relative à une solution exacte z est

$$e_n = z(t_{n+1}) - y_{n+1}, \qquad \text{avec} \quad y_{n-i} = z(t_{n-i}), \qquad 0 \le i \le r.$$

La méthode est *stable avec constante* S si pour toute suite $(\widetilde{y}_n)$ telle que

$$\widetilde{y}_{n+1} = \Psi(t_n, \widetilde{y}_n, h_n\,; \ldots ;\, t_{n-r}, \widetilde{y}_{n-r}, h_{n-r}) + \varepsilon_n$$

alors $\theta_N \le S\left(\theta_r + \sum\limits_{r \le n < N} |\varepsilon_n|\right)$ où $\theta_n = \max\limits_{0 \le i \le n} |\widetilde{y}_i - y_i|$.

∗ La méthode à pas constant $h_n = h$:

$$y_{n+1} = \sum_{0 \le i \le r} \alpha_i y_{n-i} + h \sum_{0 \le i \le r} \beta_i f_{n-i}, \quad f_n = f(t_n, y_n)$$

est d'ordre $\ge p$ si et seulement si

$$\sum_{0 \le i \le r} i^l \alpha_i - l i^{l-1} \beta_i = (-1)^l, \quad 0 \le l \le p.$$

Elle est stable si et seulement si l'équation $\lambda^{r+1} - \alpha_0 \lambda^r - \ldots - \alpha_r = 0$ a toutes ses racines de module ≤ 1, celles de module 1 étant simples. Dans ce cas, si

$(1-\alpha_0 X-\ldots-\alpha_r X^{r+1})^{-1}=\sum\gamma_n X^n$ et si $\Gamma=\sup|\gamma_n|<+\infty$, la constante de stabilité vaut

$$S=\Gamma\exp(\Gamma kT\sum|\beta_i|).$$

Méthode de Nyström (ordre 2) : $y_{n+1}=y_{n-1}+2hf_n$.

Méthode de Milne (ordre 4) : $y_{n+1}=y_{n-3}+h\Big(\frac{8}{3}\,f_n-\frac{4}{3}\,f_{n-1}+\frac{8}{3}\,f_{n-2}\Big)$.

∗ **Méthodes d'Adams-Bashforth** AB_{r+1} :

$$y_{n+1}=y_n+\int_{t_n}^{t_{n+1}}p_{n,r}(t)dt=y_n+h_n\sum_{0\le i\le r}b_{n,i,r}f_{n-i},$$

avec

$$p_{n,r}(t)=\sum_{0\le i\le r}f_{n-i}L_{n,i,r}(t),\quad L_{n,i,r}(t)=\prod_{\substack{j\ne i\\0\le j\le r}}\frac{t-t_{n-j}}{t_{n-i}-t_{n-j}},$$

$$b_{n,i,r}=\frac{1}{h_n}\int_{t_n}^{t_{n+1}}L_{n,i,r}(t)dt.$$

Erreur de consistance : $|e_n|\le|z^{(r+2)}(\xi)|h_nh_{\max}^{r+1}$ avec $\xi\in\,]t_{n-r},t_{n+1}[$.

Constante de stabilité : $S=\exp(\beta_r kT)$ avec $\beta_r=\max_n\sum_i|b_{n,i,r}|$.

∗ **Méthodes d'Adams-Moulton** AM_{r+1} **:**

$$y_{n+1}=y_n+\int_{t_n}^{t_{n+1}}p^*_{n,r}(t)dt=y_n+h_n\sum_{-1\le i\le r}b^*_{n,i,r}f_{n-i}$$

où $p^*_{n,r}$ interpole les points $(t_{n+1},f_{n+1})\ldots(t_{n-r},f_{n-r})$.

Erreur de consistance : $|e^*_n|\le|z^{(r+3)}(\xi)|h_nh_{\max}^{r+2}(1+O(h_n))$, $\xi\in\,]t_{n-r},t_{n+1}[$.

Constante de stabilité : $S=\exp\Big(\frac{\beta^*_r kT}{1-\gamma^*_r kh_{\max}}\Big)$, avec

$$\beta^*_r=\max_n\sum_{-1\le i\le r}|b^*_{n,i,r}|,\qquad\gamma^*_r=\max_n|b^*_{n,-1,r}|.$$

Si le pas $h_n=h$ est constant, les coefficients sont donnés par :

r	$b_{0,r}$	$b_{1,r}$	$b_{2,r}$	$b_{3,r}$	$b^*_{-1,r}$	$b^*_{0,r}$	$b^*_{1,r}$	$b^*_{2,r}$	$b^*_{3,r}$
0	1				$\frac{1}{2}$	$\frac{1}{2}$			
1	$\frac{3}{2}$	$-\frac{1}{2}$			$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$		
2	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$		$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$	
3	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$

∗ **Méthodes de prédiction-correction PECE et PEC**

$$\begin{cases} P : & py_{n+1} = \sum_{0\le i\le r} \alpha_{n,i} y_{n,i} + h_n \sum_{0\le i\le r} \beta_{n,i} f_{n-i} \\ E : & pf_{n+1} = f(t_{n+1}, py_{n+1}) \\ C : & y_{n+1} = y_n + h_n (b^*_{n,-1,r} pf_{n+1} + \sum_{0\le i\le r} b^*_{n,i,r} f_{n-i}) \\ E : & f_{n+1} = f(t_{n+1}, y_{n+1}) \end{cases}$$

$$\begin{cases} P : & py_{n+1} = \sum_{0\le i\le r} \alpha_{n,i} y_{n,i} + h_n \sum_{0\le i\le r} \beta_{n,i} pf_{n-i} \\ E : & pf_{n+1} = f(t_{n+1}, py_{n+1}) \\ C : & y_{n+1} = y_n + h_n \sum_{-1\le i\le r} b^*_{n,i,r} pf_{n-i} \end{cases}$$

Erreur de consistance :

$$|e_n| \le \big(1 + |b^*_{n,-1,r}|\, kh_n\big)|e^*_n| + b^*_{n,-1,r}|\, kh_n\, |pe_n|$$

où e^*_n est l'erreur de consistance de AM_{r+1} et pe_n l'erreur de consistance du prédicteur.

Constante de stabilité :

méthode PECE : $S = \exp((\beta^*_r + \gamma^*_r(A-1)Bkh_{\max})kT)$

méthode PEC : $S = \exp\left(\dfrac{\beta^*_r AkT}{1 - Bkh_{\max}}\right)$

où $A = \max_n \sum_i |\alpha_{n,i}|$, $B = \max_n \sum_i |\beta_{n,i}|$.

Chapitre X : Stabilité des solutions et points singuliers d'un champ de vecteurs

Stabilité des solutions. Une solution de valeur initiale z_0 en $t = t_0$ est dite stable (resp. asymptotiquement stable) si l'écart entre cette solution et la solution de valeur initiale z voisine de z_0 reste majorée sur tout l'intervalle $[t_0, +\infty[$ par $C\|z - z_0\|$ où C est une constante (resp. par $\gamma(t)\|z - z_0\|$ où $\lim_{t\to t_0} \gamma(t) = 0$).

Un système linéaire $Y' = AY$ est asymptotiquement stable (resp. stable) si les valeurs propres complexes de A sont toutes de partie réelle < 0 (resp. sont ou bien de partie réelle < 0, ou bien de partie réelle 0 et le bloc caractéristique correspondant est diagonal).

Courbes intégrales d'une équation $\frac{d\overrightarrow{M}}{dt} = \overrightarrow{V}(M)$ associée à un champ de vecteurs $\overrightarrow{V}(x,y) = (f(x,y), g(x,y))$ dans le plan.

On dit que $M_0 = (x_0, y_0)$ est un point singulier si $\overrightarrow{V}(M_0) = \overrightarrow{0}$, régulier sinon. Pour qu'un point singulier soit asymptotiquement stable, il suffit que la matrice

$$A = \begin{pmatrix} f'_x(x_0,y_0) & f'_y(x_0,y_0) \\ g'_x(x_0,y_0) & g'_y(x_0,y_0) \end{pmatrix}$$

ait ses valeurs propres de partie réelle < 0. On dit qu'un point singulier est non dégénéré si $\det A \neq 0$; si λ_1 et λ_2 sont les valeurs propres de A, on a alors les différentes configurations possibles suivantes :

$*$ λ_1, λ_2 réelles, distinctes, de même signe : nœud impropre,
de signe opposé : col.

$*$ $\lambda_1 = \lambda_2$ réelles et A diagonalisable : nœud propre (si champ linéaire),
et A non diagonalisable : nœud exceptionnel.

$*$ λ_1, λ_2 non réelles de partie réelle non nulle : foyer,
de partie réelle nulle : centre (si champ linéaire).

Chapitre XI : Équations différentielles dépendant d'un paramètre

Dépendance de la solution. Étant donné une équation

$$(\mathrm{E}_\lambda) \qquad y' = f(t,y,\lambda), \quad y \in \mathbb{R}^m,$$

où f dépend d'un paramètre $\lambda \in \mathbb{R}$ et est continue en (t,y,λ), la solution $y(t,y_0,\lambda)$ de (E_λ) de valeur initiale y_0 en $t = t_0$ est :

$*$ continue si f est localement lispchitzienne en y

$*$ de classe C^1 si f admet des dérivées partielles f'_{y_j} et f'_λ continues en (t,y,λ).

$*$ de classe C^{k+1} si f est de classe C^k et admet des dérivées partielles f'_{y_j}, f'_λ de classe C^k en (t,y,λ).

Dans ces deux derniers cas, la dérivée partielle $v(t) = \frac{\partial}{\partial\lambda}\, y(t,y_0,\lambda_0)$ est la solution de l'équation différentielle linéarisée

$$(\mathrm{E}'_{\lambda_0}) \qquad v'(t) = \sum_{j=1}^{m} f'_{y_j}(t,u(t),\lambda_0)v_j(t) + f'_\lambda(t,u(t),\lambda_0)$$

avec condition initiale $v(t_0) = 0$ et avec $u(t) = y(t,y_0,\lambda_0)$.

Si la valeur initiale est elle-même une fonction $y_0(\lambda)$ de classe C^1 en λ, la dérivée partielle $v(t) = \frac{\partial}{\partial \lambda} y(t, y_0(\lambda), \lambda)$ en $\lambda = \lambda_0$ satisfait la même équation linéarisée $(\mathrm{E}'_{\lambda_0})$, avec condition initiale $v(t_0) = y'_0(\lambda_0)$.

Méthode des petites perturbations. Si la solution $u(t) = y(t, y_0, \lambda_0)$ est connue et si on peut calculer la solution $v(t)$ de $(\mathrm{E}'_{\lambda_0})$, on obtient un développement limité au premier ordre

$$y(t, y_0, \lambda) = u(t) + (\lambda - \lambda_0)v(t) + o(\lambda - \lambda_0).$$

Table des matières