Introduction aux méthodes numériques

Deuxième édition

Springer
Paris
Berlin
Heidelberg
New York
Hong Kong
Londres
Milan
Tokyo

Franck Jedrzejewski

Introduction aux méthodes numériques

Deuxième édition

Springer

Franck Jedrzejewski
CEA Saclay - INSTN / UERTI
91191 Gif-sur-Yvette Cedex

ISBN-10 : 2-287-25203-7 Paris Berlin Heidelberg New York

ISBN-13 : 978-2-287-25203-7 Paris Berlin Heidelberg New York

Imprimé en France
Springer-Verlag France est membre du groupe Springer Science + Business Media

ISBN : 2-287-59711-5

SPIN : 114 03500

Maquette de couverture : Jean-François MONTMARCHÉ

Préface

C'est avec un très grand plaisir que j'accepte de présenter au lecteur cet ouvrage de mon jeune collègue et ami, Franck Jedrzejewski.

Issu de ses enseignements à la formation d'ingénieurs en Génie Atomique de l'Institut National des Sciences et Techniques Nucléaires et au DEA « Modélisation et Instrumentation en Physique » de l'Université Paris-VI, voici un livre qui deviendra indispensable à tout ingénieur ou chercheur voulant développer ou utiliser des modélisations en physique et leur concrétisation dans des codes de calcul pour la résolution des équations associées ; ce pourra être aussi une référence de base pour les étudiants de deuxième ou troisième cycles universitaires qui se destinent à la R&D en physique ou en chimie.

La physique des réacteurs enseignée au Génie Atomique – neutronique, thermohydraulique, thermomécanique, transport des rayonnements et optimisation des protections, mécanique appliquée aux réacteurs... – est un exemple typique d'une science à cheval entre la recherche de base et les applications industrielles. À ce dernier titre, elle nécessite le développement de gros logiciels scientifiques destinés à prévoir de façon précise le comportement des réacteurs, en situation normale pour optimiser les performances technico-économiques, ou en situation incidentelle pour vérifier que la sûreté est bien prise en compte : la maîtrise des méthodes numériques est une nécessité absolue pour la construction de ces logiciels.

Certes, d'autres approches peuvent être envisagées. En neutronique, par exemple, on oppose l'approche déterministe et l'approche Monte-Carlo. Cette dernière consiste à expliciter la grandeur recherchée comme l'espérance mathématique d'une variable aléatoire et à construire, avec l'ordinateur, un « jeu » générant cette variable : la réalisation d'un grand nombre de « scores » permet d'estimer cette espérance mathématique. En neutronique, le « jeu » choisi est très naturellement celui auquel « jouent » les vrais neutrons d'un réacteur dont le cheminement a un caractère stochastique (on est parfois amené à « biaiser » ce jeu de façon à réduire l'écart type des estimations). Dans d'autres cas, le jeu peut avoir un caractère plus artificiel. La technique Monte-Carlo, grosse consommatrice de calculs informatiques, est aujourd'hui de plus en plus utilisée. L'exposé de cette

technique nécessiterait à lui seul un ouvrage complet, et n'est pas abordé dans ce livre de Franck Jedrzejewski.

L'autre approche, dite par opposition « déterministe », reste cependant celle qui est la plus couramment mise en œuvre, car elle est généralement la seule qui permette des calculs d'un volume plus réduit. Elle consiste à expliciter le problème physique sous la forme d'équations mathématiques (différentielles, intégrales...), puis à tenter de résoudre numériquement le plus précisément possible ces équations. Le choix et l'optimisation des algorithmes numériques mis en pratique sont absolument cruciaux tant pour les calculs de type industriel souvent très répétitifs et devant donc pouvoir être exécutés en un temps très court, que pour les calculs de référence pour lesquels la seule limite est la patience de celui qui les fait. Par exemple, en neutronique, en laissant tourner une station de travail pendant quelques jours, les physiciens résolvent des systèmes frisant le milliard d'inconnues.

L'expérience montre qu'entre une approche numérique standard et une approche soigneusement réfléchie et optimisée un gain de temps de calcul d'un facteur 100, voire davantage, est souvent observé. Il est clair qu'on peut passer ainsi, grâce à cet effort, d'un calcul totalement déraisonnable à un calcul parfaitement banal : tout l'enjeu des méthodes numériques est là !

C'est dire l'importance pour le chercheur et pour l'ingénieur de bien connaître ces méthodes, leurs avantages et leurs limites. Dans la plupart des domaines scientifiques – non seulement la physique des réacteurs prise comme exemple, mais l'avionique, la météorologie, la thermique, etc. – tout calcul passera par l'exploitation de techniques de représentation des fonctions et des algorithmes de calcul de dérivées et d'intégrales, de résolution d'équations différentielles, aux dérivées partielles et/ou intégrales, de localisation de zéros, de recherche d'éléments propres de matrices...

Ces différents problèmes sont traités successivement dans cet ouvrage ; l'avant-dernier chapitre est dévolu à la méthode des éléments finis d'application très courante dans les différents domaines qui viennent d'être évoqués. La présentation est claire et progressive ; à noter la présence d'exercices en fin de chaque chapitre permettant au lecteur de vérifier ou de consolider l'assimilation des notions introduites, d'un index thématique, ainsi que d'une très complète bibliographie permettant au lecteur qui le souhaite d'approfondir certains aspects ou de retrouver les sources.

En un mot, un livre, fort pédagogique, qui prendra place sur le bureau des physiciens, des chercheurs et des ingénieurs, confirmés ou débutants.

Paul REUSS
Professeur à l'INSTN
Expert Senior au CEA
20 juillet 2000

Table des matières

Introduction

Ce livre est une introduction aux méthodes numériques considérées tant du point de vue pratique que de celui de leur mise en application. Il s'adresse à des physiciens ou à des ingénieurs, mais il peut aussi servir d'introduction à des mathématiciens qui souhaiteraient étudier l'analyse numérique. Il se fonde sur un cours donné pendant presque dix années et couvre toutes les notions élémentaires impliquées dans le traitement numérique, qu'il soit matriciel ou équationnel. Il doit permettre au lecteur d'acquérir une base technique suffisante pour aborder des ouvrages de mathématiques plus compliqués et fournir une connaissance des grands principes qui se trouvent mis en pratique dans le développement de grands codes de calcul.

Le premier chapitre traite non seulement des concepts premiers du calcul numérique, essentiellement de la stabilité et de la convergence, mais aussi de problèmes qui intéressent plus particulièrement les informaticiens comme la complexité algorithmique et l'optimisation. L'accélération de la convergence est illustrée par le procédé de Richardson et l'erreur commise par la perturbation d'un système matriciel introduit le conditionnement. Les problèmes bien posés sont aussi exposés de manière à sensibiliser le lecteur sur la nécessité de bien spécifier un problème avant d'en proposer une réalisation informatique.

Le deuxième chapitre introduit les problèmes d'approximation, qui, bien qu'ils se formulent souvent de manière simple, cachent en réalité de réelles difficultés. L'accent est mis sur l'approximation polynomiale et les solutions apportées au problème par Lagrange, Hermite et Tchebychev. L'étude plus générale d'une meilleure approximation montre les difficultés du problème qui n'admet pas toujours de solution. Le phénomène de Runge illustre

l'intérêt du concept de convergence uniforme. Les polynômes orthogonaux sont introduits, car ils sont à la base des techniques d'intégration par les méthodes de Gauss. La dialectique du local et du global est mise à profit dans le paragraphe sur les fonctions splines et les courbes de Bézier. L'approximation de Padé est aussi une belle illustration d'une approximation locale.

Le troisième chapitre est une brève présentation des techniques de résolution des équations algébriques. La mise en pratique de ces résolutions soulève deux problèmes essentiels : comment déterminer le nombre de racines d'une équation et comment savoir si une racine existe dans une région donnée. Le premier problème a des résonances importantes dans la théorie mathématique, puisqu'il a été résolu dans le cas polynomial par le théorème de d'Alembert qui fonde l'algèbre moderne et ses aboutissants comme la théorie de Galois, tandis que le second problème, celui de la localisation des racines, est illustré par des algorithmes applicables à des polynômes. Il dissimule l'important théorème de Rolle qui permet aussi de mesurer l'erreur des développements asymptotiques.

Les techniques d'intégration numérique sont présentées dans le quatrième chapitre qui se divise en deux parties. La première partie traite des méthodes composées, dans lesquelles la fonction à intégrer est remplacée par une approximation polynomiale ; elle couvre grosso modo les méthodes "historiques". La seconde partie traite des méthodes d'intégration de Gauss à l'aide des polynômes orthogonaux dont les propriétés sont présentées dans le deuxième chapitre.

L'analyse numérique matricielle occupe les cinquième et sixième chapitres. On présente tout d'abord les techniques de résolution des systèmes linéaires par les trois grandes catégories de méthodes classiques, à savoir les méthodes directes (méthodes de Gauss, de Cholesky, de Householder), les méthodes itératives (méthodes de Jacobi, de Gauss-Seidel, de relaxation) et les méthodes projectives (méthode de la plus profonde descente et méthodes du gradient conjugué). Certaines méthodes servent aussi à calculer l'inverse d'une matrice. On présente également des problèmes spécifiquement numériques que l'on rencontre, comme le problème des pivots de Gauss. Le sixième chapitre traite non seulement du calcul des valeurs propres et des vecteurs propres d'une matrice, mais aussi des techniques de calcul du polynôme caractéristique.

Le septième chapitre introduit les équations différentielles ordinaires, les concepts et principaux résultats associés, qui pour la plupart illustrent la différence entre local et global. Les problèmes de stabilité, de points critiques et d'hyperbolicité sont présentés, ainsi que la notion de bifurcation qui a été appliquée sur de nombreux cas physiques et a donné lieu à plusieurs publications importantes liées au problème du chaos.

Les derniers chapitres traitent des équations différentielles aux dérivées partielles. Le huitième chapitre présente les résultats fondamentaux. Après un rappel des principales définitions concernant les distributions et les

transformées de Fourier, on présente les opérateurs pseudo-différentiels et les espaces de Sobolev qui se situent entre les espaces de fonctions de classe C^k et les espaces de Lebesgue. Les propriétés des espaces de Sobolev permettent de démontrer que des problèmes sont bien posés et qu'ils admettent des solutions régulières. Plusieurs notions importantes sont esquissées, la notion de solution faible, les paramétrixes, la variété des caractéristiques et la formulation variationnelle, qui est à la base de la méthode des éléments finis présentée dans le dernier chapitre. Les grands types d'équations forment chacun un chapitre séparé. Les méthodes de résolution des équations elliptiques sont illustrées par l'équation de Poisson. Les propriétés spectrales de l'opérateur de Laplace, qui intervient aussi dans des équations paraboliques ou hyperboliques sont présentées brièvement. L'accent est mis sur les résultats liés au principe du maximum. En ce qui concerne les équations paraboliques, les résultats portent sur l'équation de la chaleur qui est un cas modèle. Enfin, les problèmes hyperboliques sont illustrés par plusieurs exemples issus de l'équation des ondes et de l'équation de J.M. Burgers. Le solveur de Riemann est introduit comme un concept central dans la résolution des équations hyperboliques. Les méthodes de viscosité numérique sont illustrées par un exemple. La stabilité des schémas numériques des équations aux dérivées partielles est démontrée par des techniques d'analyse de Fourier qui débouchent sur les conditions de Courant-Friedrichs-Lewy. Enfin, pour approfondir ce cours introductif, on pourra se reporter à la bibliographie qui termine l'ouvrage. Lors de sa rédaction, de nombreux collègues et étudiants ont bien voulu me faire part de leurs avis et suggestions pour améliorer la lisibilité du texte. Je tiens à les remercier très sincèrement pour cette fructueuse collaboration. Cette seconde édition a été rendue possible grâce à la relecture minutieuse de Luc Albert et de Christian Lebœuf. Je voudrais tout particulièrement les remercier pour ce travail, ainsi que Nicolas Puech qui a coordonné ce projet.

1
Problèmes numériques

L'analyse numérique traite de nombreux problèmes de sciences physiques, biologiques, technologiques ou des problèmes issus de modèles économiques et sociaux. Elle intervient dans le développement de codes de calcul (météorologie, physique des particules...), mais aussi dans les problèmes de simulations (aéronautique, industrie nucléaire...) ou d'expérimentations mathématiques. Elle entretient des liens étroits avec l'informatique. Si sa partie théorique relève plus des mathématiques, sa mise en pratique aboutit généralement à l'implémentation d'algorithmes sur ordinateur. Ses méthodes se fondent à la fois sur la recherche de solutions exactes comme dans le cas de l'analyse matricielle ou du calcul symbolique, sur des solutions approchées qui résultent le plus souvent de processus de discrétisation comme dans le traitement des équations différentielles. Récemment, l'analyse numérique s'est enrichie des techniques probabilistes comme les méthodes de Monte-Carlo (non traitées ici).

1.1 Erreurs et précision

Pour évaluer la précision d'un résultat, le numéricien doit connaître parfaitement les erreurs qui ont été commises. Donnons trois exemples.

Les *erreurs d'arrondi* sont imposées par le calculateur. La représentation d'un nombre en mémoire de l'ordinateur étant finie, tout nombre réel n'est connu qu'avec une précision donnée de n chiffres significatifs. Pour un

nombre quelconque compris entre 0 et 1, la machine écrira par exemple

$$x = 0, a_1 a_2 a_3 ... a_n$$

Lors de la manipulation de ces nombres, la machine devra choisir entre la troncature ou l'arrondi à la décimale la plus proche. Pour effectuer l'addition des nombres $x = 0,1234$ et $y = 0,5678$ avec seulement trois chiffres significatifs, on obtiendra soit 0,690 lorsque y est approché par 0,567 soit 0,691 lorsque y est approché par 0,568. On comprend comment, à plus grande échelle, ces erreurs peuvent induire des problèmes de précision.

Les *erreurs de troncature* sont liées à la précision de l'algorithme utilisé. Elles peuvent être contrôlées par l'algorithme lui-même. Si une fonction est approchée par son développement de Taylor, l'erreur de troncature sera obtenue par une évaluation du reste du développement. Son contrôle sera obtenu par une majoration de ce reste. Au voisinage d'un point a, si une fonction f admet un développement de Taylor de la forme

$$f(x) = f(a) + ... + \frac{(x-a)^{n-1}}{(n-1)!} f^{(n-1)}(x) + \int_a^x \frac{(x-t)^{n-1}}{(n-1)!} f^{(n)}(t) dt$$

et si la dérivée n-ième de f est majorée par une constante M, le reste sera majoré par

$$\left| \int_a^x \frac{(x-t)^{n-1}}{(n-1)!} f^{(n)}(t) dt \right| \leq M \frac{|x-a|^n}{n!}$$

Les *erreurs de méthode* se produisent lorsqu'une expression est mal équilibrée et mélange des valeurs dont la différence est importante. C'est un problème de calibration numérique qui est sensible aux erreurs d'arrondi. Dans la plupart des cas, l'algorithme doit être modifié. Considérons l'équation du second degré

$$10^{-8} x^2 - 0,8x + 10^{-8} = 0$$

Cette équation admet deux racines $r_1 \simeq 0,8.10^8$ et $r_2 \simeq 1,25.10^{-8}$. Si on ne s'intéresse qu'à la plus petite racine, certains calculateurs et en particulier les calculatrices de poche donnent des valeurs erronées. Cela provient du fait que lors du calcul du discriminant la soustraction $\Delta = 0,64 - 4.10^{-16}$ n'est pas toujours correctement effectuée car le terme 4.10^{-16} est négligé devant $0,64$. Pour obtenir une valeur exacte on doit modifier l'algorithme en proposant par exemple de calculer la racine r_2 par la relation donnant le produit des racines $r_2 = 1/r_1$. Remarquons que si on multiplie l'équation par 10^8 le problème reste entier.

Dans les processus récurrents ou itératifs, les erreurs s'ajoutent, ce qui a pour effet d'amplifier l'erreur globale et de diminuer la précision du calcul.

La propagation des erreurs dans diverses parties du calcul a pour conséquence d'ajouter de l'imprécision là où elle n'était pas nécessairement attendue. Dans les calculs itératifs, l'erreur se propage d'une étape à l'autre.

Exemple. Dans le calcul numérique des termes de la suite définie par la relation de récurrence

$$x_{n+1} = \frac{1}{n} + ax_n$$

l'erreur qui est donnée par

$$\Delta x_{n+1} \simeq a\Delta x_n$$

évolue exponentiellement. À l'étape n, l'erreur est multipliée par a^n. D'une étape à l'autre l'erreur se propage et peut conduire à l'explosion de l'algorithme.

1.2 Convergence et stabilité

Les méthodes numériques utilisées pour résoudre un problème approché conduisent à un résultat qui est toujours entaché d'erreur. Cette erreur doit être suffisamment petite pour que la solution numérique converge vers la solution réelle. Dans ce cas l'algorithme (ou la méthode) est dit *convergent*. Si un raisonnement mathématique permet de montrer qu'une méthode diverge, elle ne pourra en aucun cas être utilisée sur un calculateur. En revanche, si la méthode converge il se peut qu'en pratique elle diverge.

La *vitesse de convergence* est un facteur important de la qualité des algorithmes. Si la vitesse de convergence est élevée, l'algorithme converge rapidement et le temps de calcul est moindre. Ces préoccupations de rapidité de convergence ont conduit à diversifier les *modes de convergence* et à chercher des processus *optimaux*.

La *stabilité* garantit que les erreurs ne s'amplifient pas au cours du déroulement de l'algorithme et que la méthode reste stable. À côté de cette stabilité numérique, il y a aussi la *stabilité des solutions* qui intervient dans les problèmes équationnels et qui est bien mise en évidence par les techniques perturbatives. Lorqu'un problème (P) admet une solution, il est intéressant d'envisager le problème perturbé, noté (P_ε), où ε est un petit paramètre et de se demander si les solutions du système perturbé sont voisines de la solution du système non perturbé. Il n'existe pas de théorème général qui réponde à cette question.

Donnons quelques définitions. Soit $u : I \longrightarrow R$ une fonction à valeurs réelles définie sur un intervalle $I = [a, b]$ et une subdivision $a = x_0 < x_1 < x_2$ ...

$< x_n = b$. On note $h_i = x_i - x_{i-1}$ et h est la plus grande valeur des pas de la subdivision $h = \sup_i(h_i)$. On suppose que la fonction u est dotée d'une réalisation numérique (méthode, processus ou schéma de discrétisation) qui s'exprime sous la forme

$$u_{i+1} = \phi(h_1, ..., h_i, u_1, ...u_i)$$

On appelle *erreur de consistance* relative à la fonction $u(x)$ la quantité $e_i = u(x_i) - u_i$ et *erreur globale* l'expression

$$e = \sup_{0 \leq i \leq n} |u(x_i) - u_i|$$

On dit que la méthode *converge* si l'erreur globale tend vers 0 lorsque le pas de la subdivision h tend vers 0.

La méthode est dite *consistante* si la somme $\sum_{i=0}^{n} |e_i|$ des erreurs de consistance relatives à la fonction u tend vers 0 quand h tend vers 0.

La méthode est d'*ordre* p si la limite

$$\lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^p}$$

existe quand n tend vers l'infini. On dit que l'erreur de consistance est en h^p, et on note

$$e_i = O(h^p) \quad \forall\, i = 0, ..., n$$

La méthode est *stable* si pour toutes suites voisines u_{i+1} et v_{i+1} vérifiant

$$u_{i+1} = \phi(h_1, ..., h_i, u_1, ...u_i)$$
$$v_{i+1} = \phi(h_1, ..., h_i, v_1, ...v_i) + \varepsilon_i$$

il existe une constante S appelée *constante de stabilité* satisfaisant l'inégalité

$$\sup_{0 \leq i \leq n} |v_i - u_i| \leq S \sum_{i=0}^{n} |e_i|$$

On démontre que tout processus stable et consistant converge.

En effet, la méthode est stable, donc pour les suites u_{i+1} et $v_{i+1} = u(x_{i+1})$, on a

$$\sup_{0 \leq i \leq n} |v_i - u_i| \leq S \sum_{i=0}^{n} |e_i|$$

La méthode étant consistante, le membre de droite tend vers 0 lorsque h tend vers 0.

1.3 Accélération de la convergence

Le procédé d'extrapolation de Richardson illustre l'accélération de la convergence d'une méthode numérique. Proposée en 1927 par Lewis Fry Richardson (1881-1953), l'extrapolation à la limite consiste à calculer plusieurs fois la même quantité avec un maillage différent. Soit $r > 1$ un réel fixé et u_h une approximation de u. Si u est du premier ordre et si le calcul est fait deux fois, on a

$$\begin{aligned} u_h &= u + \alpha h + O(h^2) \\ u_{h/r} &= u + \alpha \tfrac{h}{r} + O(h^2) \end{aligned}$$

Ainsi, en combinant le résultat u_h avec un résultat issu d'un maillage plus fin, on obtient

$$r u_{h/r} - u_h = (r-1)u + O(h^2)$$

Plus généralement, si u est approché à l'ordre n

$$u(h) = a + bh^n + ch^{n+1} + \cdots + eh^{n+l} + o(h^{n+l})$$

En prenant deux pas quelconques h_1 et h_2, si le pas h_2 est plus petit que le pas h_1, $u(h_2)$ est une meilleure approximation que $u(h_1)$. On obtient une approximation encore meilleure en supprimant le terme en h^n, en prenant

$$u(h_1, h_2) = \frac{h_1^n u(h_2) - h_2^n u(h_1)}{h_1^n - h_2^n}$$

En particulier, pour $h_1 = h$ et $h_2 = h/r$ avec $r > 1$, on obtient

$$\begin{aligned} u_h &= u + ah^n + O(h^{n+1}) \\ u_{h/r} &= u + a\tfrac{h^n}{r^n} + O(h^{n+1}) \end{aligned}$$

d'où la relation usuelle

$$\frac{r^n u_{h/r} - u_h}{r^n - 1} = u + O(h^{n+1}) \quad \forall n \geq 1$$

1.4 Complexité

Les problèmes traités sur un calculateur se répartissent en deux grandes catégories selon qu'une valeur numérique est attendue (problèmes de calcul) ou qu'une réponse par oui ou non est souhaitée (problème de décision). Les propriétés des algorithmes ont été étudiées dans les années 1930 par le mathématicien Alan Turing (1912-1954) qui inventa la machine qui porte son nom. Turing, en démontrant que les problèmes qui ne pouvaient pas être résolus par sa machine symbolique n'avaient pas d'algorithme, fixa les limites de la calculabilité. Depuis, on classe les problèmes en deux grandes

catégories : les problèmes pour lesquels il n'existe pas d'algorithme et les problèmes pour lesquels un algorithme existe. Parmi ces derniers, on mesure l'efficacité de l'algorithme selon la croissance de la durée de leur exécution en fonction de la taille du problème. Pour un problème de taille n, on considère comme efficaces les algorithmes dont la croissance est polynomiale et inefficaces ou difficilement exploitables les algorithmes dont la croissance est exponentielle. On dit qu'un algorithme n'est pas *résoluble* ou *décidable* s'il n'est pas justiciable d'une solution à l'aide d'un algorithme. On distingue plusieurs classes.

La classe P (*Polynomial*) représente la classe des langages décidables en un temps polynomial : ce sont les problèmes qui admettent une solution sur une machine de Turing en temps polynomial. La résolution d'un problème est obtenue en un temps inférieur à une puissance donnée de la taille n du problème : si la taille n du problème augmente, le nombre d'étapes de l'algorithme reste toujours plus petit qu'une certaine puissance de n.

La classe NP (*Non deterministic polynomial*) représente la classe des langages décidables en temps non déterministe polynomial. Ce sont des problèmes pour lesquels, si une solution est proposée, on peut vérifier que cette solution répond bien au problème en un temps polynomial. Pour certains problèmes de cette classe, on ne connaît aucun algorithme polynomial. On sait que la classe P est contenue dans la classe NP et on conjecture que $P \neq NP$. Le coloriage d'une carte est un problème de la classe NP. En 1975, Kenneth Appel et Wolfgang Haken ont "démontré" sur ordinateur qu'il suffit de quatre couleurs pour colorier une carte en évitant que deux pays voisins aient la même couleur.

La classe NP-complet représente les problèmes de la classe NP qui sont liés : si un problème de cette classe peut être résolu par un algorithme en temps polynomial, alors tous les problèmes de la classe NP seront solubles par un algorithme efficace. Si on trouve un tel algorithme, on aura alors identité des classes P et NP. Le problème du voyageur de commerce, qui consiste à trouver le chemin le plus court reliant une série de villes, est un problème NP-complet. Le *problème du sac à dos* : étant donné un sous-ensemble S de l'ensemble des entiers naturels et m un nombre positif, peut-on trouver une partie A de S telle que la somme de ses éléments soit égale à l'entier m, est un problème NP-complet.

La complexité des algorithmes se mesure en ne retenant que des ordres de grandeurs. Si $T(n)$ désigne le nombre d'instructions élémentaires exécutées par une machine formelle, on dira que le temps d'exécution est en $O(T(n))$ ou que la complexité de l'algorithme est proportionnelle à $f(n)$ si en notation de Landau

$$T(n) = O(f(n))$$

c'est-à-dire s'il existe deux constantes c et n_0 telles que

$$T(n) \leq cf(n) \qquad \forall n \geq n_0$$

Dans une méthode à accès direct, une donnée est localisée en $O(1)$ opérations. L'accès dans un arbre de recherche est en $O(logn)$. Une addition polynomiale est en $O(n)$. Un tri récursif ou une transformée de Fourier rapide sont en $O(nlogn)$. Une multiplication matricielle en $O(n^2)$. Donnons un exemple simple de calcul de complexité.

Exemple. Considérons l'algorithme récursif du calcul de n !

```
        FAIRE
(1)       SI (n ≤ 1) ALORS
(2)          fact = 1
(3)       SINON  fact(n) = n * fact(n − 1)
        FIN FAIRE
```

Les lignes (1) et (2) ont une complexité en $O(1)$, la ligne (3) est en $O(1)$ $+T(n-1)$. Par conséquent, si le nombre c désigne le nombre d'opérations en $O(1)$ et si $n > 1$, alors $T(n) = c + T(n-1)$. En définitive, $T(n)$ $= c(n-1) + T(1)$, d'où $T(n) = O(n)$. L'algorithme récursif du calcul de factoriel n est donc en $O(n)$.

1.5 Optimisation

En pratique, le choix d'un algorithme n'est pas toujours un problème simple. On cherchera l'algorithme qui donne la meilleure *précision* sur les résultats obtenus et qui minimise l'*encombrement mémoire* et le *temps de calcul*. L'optimisation cherche à réduire le nombre d'opérations et en premier lieu le nombre de multiplications. Donnons deux exemples dans lesquels on cherche à diminuer le nombre de multiplications, quitte à les remplacer par des additions, moins coûteuses en temps de calcul.

Exemple 1. Le produit de deux nombres complexes $z_1 = a+ib$ et $z_2 = c+id$ nécessite l'évaluation de quatre quantités ac, bd, ad et bc. En écrivant :

$$ac - bd = (a+b)c - b(c+d)$$
$$ad + bc = (a-b)d - b(c+d)$$

on diminue le calcul à l'évaluation de trois quantités : $(a+b)c$, $(a-b)d$ et $b(c+d)$. Le gain vient du fait qu'une multiplication est beaucoup plus lente qu'une addition.

Exemple 2. Dans un produit matriciel, on peut réduire le nombre de multiplications en augmentant le nombre d'additions. Le produit de deux matrices à deux lignes et deux colonnes nécessite sept multiplications et non huit :

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$$

On calcule successivement

$$\begin{aligned} p_1 &= ae \\ p_2 &= bg \\ p_3 &= (a-c)(h-f) \\ p_4 &= (c+d)(f-e) \\ p_5 &= (a+b-c-d)h \\ p_6 &= d(e-f+g-h) \\ p_7 &= (c-a+d)(e-f+h) \end{aligned}$$

Ces sept valeurs suffisent pour déterminer le produit des deux matrices

$$\begin{aligned} \alpha &= p_1 + p_2 \\ \beta &= p_1 + p_4 + p_5 + p_7 \\ \gamma &= p_1 + p_3 + p_6 + p_7 \\ \delta &= p_1 + p_3 + p_4 + p_7 \end{aligned}$$

Lorsque plusieurs utilisateurs ou plusieurs fragments de calculs veulent utiliser un même résultat, il est possible d'organiser le calcul de manière parallèle de sorte que chaque calcul réutilise un ensemble de données préalablement calculées : c'est le *préconditionnement*. Par exemple, pour évaluer la valeur en x d'un polynôme du quatrième degré, on calculera au préalable les quantités α, β, γ et δ définies par :

$$ax^4 + bx^3 + cx^2 + dx + e = [(x+\alpha)x + \beta]\,[(x+\alpha)x + (x+\gamma)] + \delta$$

Les coefficients étant donnés par les relations

$$\begin{aligned} \alpha &= (b-a)/2a \\ \beta &= (d\alpha c/a^2) + \alpha^2(\alpha+1) \\ \gamma &= (c/a) - \alpha(\alpha+1) - \beta \\ \delta &= e - a\beta\gamma \end{aligned}$$

Par l'évaluation de ces quantités et leur mise à disposition dans d'autres calculs, le calcul du polynôme ne nécessite plus que trois multiplications. On peut généraliser ce processus. Strassen a montré que le produit de deux matrices $2n \times 2n$ se ramène au produit de sept matrices $n \times n$.

La *règle de Horner* permet l'évaluation d'un polynôme en un point donné en un nombre optimal d'opérations. Dans un article publié en 1819, William Horner (1786-1837) a indiqué une méthode pour évaluer la valeur d'un

polynôme en un point x_0. La méthode usuelle qui consiste à calculer d'abord x^2 puis x^3, ..., puis x^n nécessite $(2n-1)$ multiplications et n additions. Pour calculer le polynôme

$$P(x_0) = a_n x_0^n + a_{n-1} x_0^{n-1} + ... + a_1 x_0 + a_0$$

Horner propose de factoriser $P(x)$ sous la forme :

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

et d'évaluer successivement les quantités

$$\begin{aligned} b_n &= a_n \\ b_{n-1} &= a_{n-1} + x_0 b_n \\ &... \\ b_1 &= a_1 + x_0 b_2 \\ b_0 &= a_0 + x_0 b_1 \end{aligned}$$

Au terme de ce calcul b_0 donne la valeur du polynôme P au point x_0. À chaque étape, on effectue une multiplication et une addition, de sorte que la méthode de Horner pour évaluer la valeur d'un polynôme de degré n en un point donné nécessite n multiplications et n additions, ce qui réalise une économie par rapport à la méthode usuelle et par conséquent un gain de temps si le degré du polynôme est élevé. On démontre que la méthode de Horner est optimale et que c'est la seule méthode optimale. L'extension de la règle de Horner à des systèmes de polynômes ou à des polynômes de plusieurs variables est aussi optimale.

1.6 Problèmes bien posés, problèmes raides

Les équations différentielles offrent des exemples variés de problèmes numériques. Nous adopterons les définitions suivantes : Un problème (P) est *mathématiquement bien posé* si le problème (P) admet une solution unique qui est stable au sens de Hadamard, c'est-à-dire qui dépend continûment des données initiales. Un problème numérique est dit *numériquement bien posé* si la continuité de la solution est suffisamment bonne par rapport aux conditions initiales pour que la solution ne soit pas perturbée par une erreur initiale ou de petites erreurs d'arrondi.

Exemple 1. Le problème de Neumann pour une fonction $u(x)$ définie sur un intervalle $[a, b]$:

$$\begin{cases} u''(x) = 0 \\ u'(a) = u_0 \\ u'(b) = v_0 \end{cases}$$

admet aucune solution si $u_0 \neq v_0$ et une infinité de solutions si $u_0 = v_0$ de la forme $u(x) = u_0 x + c$, où c est une constante arbitraire. Le problème, qui n'admet pas de solution unique, est donc mal posé.

Exemple 2. Le problème de Cauchy pour $x \geq 0$

$$\begin{cases} u'(x) = 2\sqrt{|u(x)|} \\ u(0) = 0 \end{cases}$$

admet une infinité de solutions de la forme :

$$\begin{cases} u(x) = 0 & x \in [0, a] \\ u(x) = (x-a)^2 & x \in [a, \infty[\end{cases}$$

La quantité a étant arbitraire, le problème est mal posé.

Exemple 3. Considérons le problème suivant

$$\begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0 \qquad x \in [0,1]\,, y \in [0,1] \\ \\ u(x,0) = 0 \\ \\ \dfrac{\partial u}{\partial y}(x,0) = \dfrac{\pi}{n}\sin(\pi n x) \qquad n = 1, 2, ... \end{cases}$$

Les fonctions

$$u_n(x,y) = \frac{1}{n^2}\sin(\pi n x) sh(\pi n y)$$

sont solutions du système précédent. Or, pour chaque valeur de n, on peut trouver un nombre $x_n \in [0,1]$ tel que $\sin(\pi n x_n) = 1$ vérifiant :

$$\lim_{n\to\infty} \sup |u_n(x,y)| = \infty$$

Ce qui prouve que les solutions de (P) ne dépendent pas continûment des données initiales.

Exemple 4. Le problème de Cauchy

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} + \dfrac{\partial u}{\partial x} = 0 \\ \\ u(x,0) = 0 \\ \\ \dfrac{\partial u}{\partial t}(x,0) = \dfrac{1}{n} e^{-n^2 x} \end{cases}$$

admet une solution de la forme

$$u(x,t) = \frac{1}{n^2}\frac{e^{nt} - e^{-nt}}{2} e^{-n^2 x}$$

Si n tend vers l'infini, en $t = 0$, u tend uniformément vers 0 ainsi que sa dérivée partielle en temps alors que $u(x,t)$ diverge dans chaque région pour $t > 0$: le problème est instable au sens de Hadamard.

Exemple 5. Considérons l'équation différentielle :

$$\begin{cases} u'(x) = u(x) - 1 \\ u(0) = 0 \end{cases}$$

Cette équation admet comme solution $u(x) = e^x - 1$. Si la condition initiale est donnée par $u(0) = \varepsilon$, la solution est alors $v(x) = (1+\varepsilon)e^x - 1$. De sorte que la différence s'écrit : $v(x) - u(x) = \varepsilon e^x$. Si x varie dans l'intervalle $[0, 30]$, on a $v(30) - u(30) = \varepsilon e^{30} \simeq 10^{13}\varepsilon$. Si la précision des calculs est de 10^{-10}, le problème est numériquement mal posé, bien que mathématiquement bien posé.

Un problème est dit *numériquement raide* si sa solution par une méthode numérique ne peut être donnée en un temps raisonnable. Dans un problème raide (*stiff*, en anglais) les variables calculées évoluent lentement, malgré un pas d'intégration petit. Ce phénomène est courant dans les méthodes explicites du type Runge-Kutta ou Adams. De plus, il apparaît souvent des instabilités, sous la forme d'oscillations qui font diverger l'algorithme. Lorsque le système est raide, il faut suivre la plus petite échelle de temps du système pour s'assurer de la stabilité de la méthode, comme le montre l'exemple suivant.

Exemple 6. Soit λ un réel $\lambda > 2$. Considérons le système d'équations différentielles

$$\begin{cases} u' = (\lambda - 2)u + 2(\lambda - 1)v \\ v' = (1 - \lambda)u + (2\lambda - 1)v \end{cases}$$

avec pour conditions initiales, $u(0) = 1$ et $v(0) = 0$. Ce système admet comme solutions exactes $u(x) = 2e^{-x} - e^{-\lambda x}$ et $v(x) = -e^{-x} + e^{-\lambda x}$. Il a deux échelles de temps données par les deux exponentielles. Si on choisit $\lambda = 10000$, on constate que bien que le terme $e^{-\lambda x}$ soit négligeable dans la solution exacte, il introduit une instabilité dans la solution numérique. Pour que la méthode soit stable, il faut que le pas de temps choisi soit très inférieur à $1/\lambda$. Ce phénomène est caractéristique des systèmes raides.

1.7 Conditionnement

Le conditionnement mesure l'influence des erreurs d'arrondi sur la solution d'un problème donné. Il est mis en évidence par une légère perturbation des données initiales. C'est une notion générale qui s'applique aussi bien aux racines d'un polynôme vis-à-vis de la variation de ses coefficients

qu'aux valeurs propres ou vecteurs propres d'une matrice vis-à-vis de la perturbation de ses éléments. Considérons le système linéaire $Ax = b$ suivant

$$\begin{pmatrix} 23 & 9 & 12 \\ 12 & 10 & 1 \\ 14 & -12 & 25 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 44 \\ 23 \\ 27 \end{pmatrix}$$

qui admet la solution $(1, 1, 1)$. Remarquons que la matrice A est inversible et admet trois valeurs propres $\lambda_1 \simeq 36,16$ $\lambda_2 \simeq 0,056$ et $\lambda_3 \simeq 21,79$. Considérons le problème suivant dans lequel le vecteur b est légèrement perturbé

$$\begin{pmatrix} 23 & 9 & 12 \\ 12 & 10 & 1 \\ 14 & -12 & 25 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 44,44 \\ 22,77 \\ 26,73 \end{pmatrix}$$

La solution du système perturbé est $(6,23 \quad 4,73 \quad 4,69)$. Remarquons qu'une erreur de $1/100$ sur les données entraîne une erreur relative de l'ordre de 5 sur la solution, les composantes du vecteur solution sont multipliées par 5. De même, si on perturbe les éléments de la matrice

$$\begin{pmatrix} 23,23 & 9,09 & 12,12 \\ 12,12 & 9,9 & 1,01 \\ 14,14 & -11,88 & 25,25 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 44 \\ 23 \\ 27 \end{pmatrix}$$

Une erreur de $1/100$ sur les données provoque une erreur de l'ordre de 6. L'amplification des erreurs relatives est d'environ 600. La solution du système est $(6,89 \quad 5,56 \quad 5,40)$.

Envisageons ce problème du point de vue algébrique. Soit A une matrice inversible et $Ax = b$ un système linéaire. Étudions la perturbation

$$(A + \delta A)(x + \delta x) = b + \delta b$$

où δA et δb sont les perturbations sur A et b dues aux erreurs d'arrondi et δx l'erreur commise sur la résolution du système linéaire. Comme $Ax = b$, il vient

$$(A + \delta A)(\delta x) = \delta b - \delta A.A^{-1}.b$$

Si la matrice $I + A^{-1}.\delta A$ est inversible, alors :

$$\delta x = (I + A^{-1}.\delta A)^{-1}.A^{-1}.(\delta b - \delta A.A^{-1}.b)$$

D'où la majoration

$$\|\delta x\| \leq \frac{\left\|A^{-1}\right\| . \left(\|\delta b\| + \|\delta A\| . \left\|A^{-1}.b\right\|\right)}{1 - \|A^{-1}\| . \|\delta A\|}$$

Comme $\left\|A^{-1}.b\right\| = \|x\| \geq \frac{\|b\|}{\|A\|}$, on a une majoration de l'erreur relative

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right)$$

si la constante $\kappa(A) = \|A\| \, . \left\|A^{-1}\right\|$ vérifie $\kappa(A)\frac{\|\delta A\|}{\|A\|} < 1$

Soit A une matrice inversible, on appelle *conditionnement* de A, le nombre

$$\kappa(A) = \|A\| \, . \left\|A^{-1}\right\|$$

Ce nombre dépend du choix de la norme : il y a autant de définitions du conditionnement que de normes matricielles. La norme standard, appelée la *1-norme,*

$$\begin{aligned} \|A\|_1 &= \sup_j \sum_i |a_{ij}| \\ cond_1(A) &= \|A\|_1 \left\|A^{-1}\right\|_1 \end{aligned}$$

est définie par le maximum de la somme des modules des éléments d'une ligne. La 2-norme définie par

$$\|A\| = \|A^*\| = \sqrt{\rho(AA^*)} = \sqrt{\rho(A^*A)}$$

où $\rho(A)$ est le rayon spectral de A, c'est-à-dire le plus grand des modules des valeurs propres de A. Le conditionnement est noté

$$cond_2(A) = \|A\|_2 \left\|A^{-1}\right\|_2$$

La 2-norme vérifie les inégalités

$$\|A\|_2 \leq \|A\|_e \leq \sqrt{n} \, \|A\|_2$$

Pour la norme euclidienne ou norme de Frobenius

$$\|A\|_e = \sqrt{\sum_{i,j} |a_{ij}|^2}$$

Le nombre de conditionnement est de la forme

$$cond_e(A) = \|A\|_e \left\|A^{-1}\right\|_e$$

Pour une matrice carrée A d'ordre n, le conditionnement pour la 2-norme a les propriétés suivantes :

(1) Le conditionnement est un nombre positif

$$cond_2(A) \geq 0$$

(2) Le conditionnement de la matrice ne varie pas lorsqu'on multiplie la matrice par un scalaire

$$\forall \alpha \in \mathbb{C}, \quad cond_2(\alpha A) = cond_2(A)$$

(3) Le conditionnement est invariant par transformation unitaire. Pour toute matrice unitaire U, on a

$$cond_2(AU) = cond_2(UA) = cond_2(A)$$

(4) Soit $\sigma^2_{\max}$ la plus grande et $\sigma^2_{\min}$ la plus petite des valeurs propres de la matrice A^*A, on a alors

$$cond_2\,(A) = \frac{\sigma_{\max}}{\sigma_{\min}}$$

(5) Si A est une matrice hermitienne et si $\lambda_{\max}$ et $\lambda_{\min}$ désignent respectivement la plus grande et la plus petite des valeurs propres de A en valeur absolue, on a :

$$cond_2\,(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

La vérification de ces propriétés est immédiate. La propriété (1) s'établit en considérant le produit $I = A.A^{-1}$

$$\|A\| \,.\, \left\|A^{-1}\right\| \geq \left\|AA^{-1}\right\| = \|I\| = 1$$

La propriété (2) est une conséquence directe des axiomes de définition d'une norme. La propriété (3) résulte de

$$cond_2(U.A) = \|U.A\| \,.\, \left\|A^{-1}U^*\right\| = \|A\| \,.\, \left\|A^{-1}\right\| = cond_2(A)$$

Pour démontrer la propriété (4) remarquons que

$$\|A\|^2 = \rho(A^*A) = \sup_i \lambda_i(A^*A) = \sigma^2_{\max}$$

les matrices A^*A et AA^* étant semblables

$$\begin{aligned} \left\|A^{-1}\right\|^2 &= \rho((AA^*)^{-1}) = \rho((A^*A)^{-1}) \\ &= \sup_i \lambda_i(A^*A)^{-1} = \frac{1}{\inf_i \lambda_i(A^*A)} = \frac{1}{\sigma^2_{\min}} \end{aligned}$$

La propriété (5) résulte de l'égalité $\|A\| = \rho(A)$ vérifiée pour les matrices normales.

Le conditionnement mesure l'*éparpillement relatif* $(\lambda_{\max}/\lambda_{\min})$ des valeurs propres. Dans l'exemple qui précéde, les valeurs numériques donnent un conditionnement égal à $cond_2(A) \simeq 645$. Si on pose

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad \delta x = \begin{pmatrix} 5.23 \\ -5.73 \\ -5.69 \end{pmatrix} \qquad b = \begin{pmatrix} 44 \\ 23 \\ 27 \end{pmatrix} \qquad \delta b = \begin{pmatrix} 0.44 \\ -0.23 \\ -0.27 \end{pmatrix}$$

Les calculs montrent que

$$\frac{\|\delta x\|}{\|x\|} \simeq \frac{9.62}{1.73} \simeq 5.55 \text{ et } \kappa(A)\frac{\|\delta b\|}{\|b\|} \simeq 645 * 0.01 \simeq 6.45$$

l'égalité est presque satisfaite.

Une matrice est dite *bien équilibrée* si ses vecteurs lignes et ses vecteurs colonnes ont une norme de l'ordre de grandeur de l'unité. Une matrice est dite *bien conditionnée* si son conditionnement est de l'ordre de grandeur de l'unité. Remarquons qu'une matrice équilibrée peut être mal conditionnée. La matrice suivante, matrice carrée d'ordre 100, écrite sous sa forme de Jordan

$$A = \begin{pmatrix} 1/2 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1/2 & 1 & 0 & \dots & \dots & 0 \\ \vdots & & \ddots & \ddots & & & \vdots \\ \vdots & & & \ddots & \ddots & & \vdots \\ 0 & & & 0 & 1/2 & 1 & 0 \\ 0 & \dots & \dots & \dots & 0 & 1/2 & 1 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1/2 \end{pmatrix}$$

est équilibrée et mal conditionnée. En effet, sa plus petite valeur propre est $1/2$. Son inverse est formé d'éléments b_{ij}. L'élément $b_{1,100} = 2^{100} \geq 10^{30}$ montre que $\left\|A^{-1}\right\| > 10^{30}$ et comme $\|A\| \geq \frac{1}{\sqrt{n}} \|A\|_e > 1,1$ on en déduit que le conditionnement de A excède 10^{30}.

On appelle *matrice de Hilbert* une matrice symétrique d'ordre n dont les éléments sont donnés par :

$$h_{ij} = \frac{1}{i+j-1}$$

Pour les ordres 2 et 3, les matrices de Hilbert s'écrivent

$$H_2 = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix} \qquad H_3 = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$$

Les matrices de Hilbert sont des matrices mal conditionnées. Selon l'ordre de la matrice et le choix de la norme, on a les valeurs suivantes du conditionnement.

n	$cond_1$	$cond_2$	$cond_e$
2	27	19.281	19.3
3	748	524.06	526.2
4	28 375	$0.6.10^{-4}$	15613.8
5	943 656	$0.21.10^{-5}$	480849.1
6	29 070 279	$0.66.10^{-7}$	15 118 987.1

1.8 Exercices

1. Calculer la complexité de l'algorithme suivant

   For $i = 1$ to $(n-1)$ do
   For $j = i+1$ to n do
   For $k = 1$ to j do
   $A(k) = 1$
   Endo Endo Endo

2. Calculer la complexité de l'algorithme suivant

   For $i = 1$ to $(n-1)$ do
   For $j = n$ downto $(i+1)$ do
   If $A(j-1) > A(j)$ then
   Begin
   $b = A(j-1)$
   $A(j-1) = A(j)$
   $A(j) = b$
   End Endo Endo

3. Le problème $y'(x) = 3y(x)^{2/3}$, et $y(0) = 0$ est-il un problème bien posé ?

4. Calculer le conditionnement de la matrice

$$A = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

5. Montrer que le conditionnement pour la 2-norme de la matrice $(a_{i,j})$ qui est nul partout sauf pour $a_{i,i} = 1$ et $a_{i,i+1} = 2$ est toujours supérieur ou égal à 2^n.

6. On considère la matrice

$$A = \begin{pmatrix} 1+\varepsilon\cos(2/\varepsilon) & -\varepsilon\sin(2/\varepsilon) \\ -\varepsilon\sin(2/\varepsilon) & 1-\varepsilon\cos(2/\varepsilon) \end{pmatrix}$$

où ε est un réel compris entre 0 et 1. Calculer le conditionnement de A. Étudier le cas où ε tend vers 0. Soit P la matrice de passage formée des vecteurs propres de A. Montrer que si ε tend vers 0, la matrice P n'a pas de limite. Calculer le conditionnement de P. Commenter.

2
Approximation et interpolation

Dans les problèmes numériques, on substitue très souvent une fonction $f(x)$ connue en un nombre fini de points $x_0, x_1, ..., x_n$ par une fonction $P(x)$ plus simple et facilement calculable : c'est l'*approximation*. En termes mathématiques, l'approximation consiste à minimiser la distance qui sépare les fonctions $f(x)$ et $P(x)$. L'*interpolation* impose de plus que les fonctions $f(x)$ et $P(x)$ coïncident aux points x_j. Lorsque la fonction $P(x)$ représente la fonction $f(x)$ décrite par un ensemble de points expérimentaux $(x_j, f(x_j))$, on parle de *lissage*. L'approximation d'une fonction est liée aux problèmes de représentation des fonctions comme limites de fonctions plus simples (développements en série, développements en série de Fourier, représentations intégrales, etc.). En pratique, on cherche à construire une suite de fonctions $f_n(x)$ qui converge vers la fonction de base $f(x)$. Lorsque les fonctions $f_n(x)$ sont des polynômes, on parle d'*approximation polynomiale*. L'approximation polynomiale est une des plus utilisées, car il est facile de rendre l'erreur d'approximation arbitrairement petite en augmentant le degré du polynôme. Elle se fonde sur le théorème de Weierstrass (1866) qui affirme que toute fonction continue sur un intervalle $[a, b]$ est limite uniforme d'une suite de fonctions polynomiales.

2.1 Interpolation de Lagrange

L'approximation polynomiale, fondée en général sur le développement en série de Taylor, permet d'approcher une fonction f suffisamment

régulière par un polynôme de degré n. Rappelons que la série de Taylor d'une fonction peut ne pas converger et que, si elle converge, elle peut converger vers une quantité différente de la fonction initiale (par exemple, $f(x) = exp(-1/x^2)$ au voisinage de l'origine). Publiée pour la première fois par Brook Taylor (1685-1731) en 1715, puis reprise par Joseph-Louis Lagrange (1736-1813), et démontrée avec reste intégral par Augustin-Louis Cauchy (1789-1857), la formule de Taylor conduit à une estimation de l'erreur dans l'approximation d'une fonction par un polynôme de Lagrange.

Soit f une fonction continue d'un intervalle $[a, b]$ dans $\mathbb{R}$ et $x_0, x_1, ..., x_n$ $(n + 1)$ points distincts de l'intervalle $[a, b]$. Considérons les polynômes de degré n définis par

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

Ces polynômes sont appelés les *polynômes de Lagrange*. Il ne faut pas les confondre avec le polynôme d'*interpolation* de Lagrange P_n (voir ci-dessous). En posant

$$\pi_n(x) = \prod_{j=0}^{n} (x - x_j)$$

Les polynômes de Lagrange s'écrivent de manière plus simple, sous la forme

$$\forall x \neq x_i \qquad l_i(x) = \frac{\pi_n(x)}{(x - x_i)\pi'_n(x_i)}$$

On démontre le résultat suivant : Toute fonction continue sur un intervalle borné et connue en $(n + 1)$ points distincts peut être approchée par un polynôme qui coïncide avec cette fonction en ces $(n+1)$ points. Si $f : [a, b] \rightarrow R$ est une fonction continue et si $x_0, x_1, ..., x_n$ sont $(n+1)$ points distincts de l'intervalle $[a, b]$, alors il existe un unique polynôme P_n de degré n appelé *polynôme d'interpolation de Lagrange*, dont la valeur coïncide avec f aux points x_i, c'est-à-dire vérifiant $P_n(x_i) = f(x_i)$, et qui est donné par la formule

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

Il est facile de vérifier ce résultat. En effet, le polynôme P_n vérifie l'égalité $P_n(x_i) = f(x_i)$, car les polynômes de Lagrange satisfont $l_i(x_j) = \delta_{ij}$ où δ_{ij} est le symbole de Kronecker ($\delta_{ij} = 0$ si $i \neq j$ et $\delta_{ii} = 1$), ce qui prouve l'existence de ce polynôme. Pour montrer l'unicité, supposons qu'il existe un polynôme Q_n de degré n vérifiant $Q_n(x_i) = P_n(x_i) = f(x_i)$. Chaque

valeur x_i est racine du polynôme $Q_n - P_n$. Le polynôme $Q_n - P_n$ a donc au moins $(n+1)$ racines distinctes et est de degré n. Par conséquent, $Q_n - P_n = 0$. Pour démontrer l'existence du polynôme d'interpolation de Lagrange, on peut aussi chercher analytiquement un polynôme de la forme

$$P_n(x) = a_n x^n + \cdots + a_1 x + a_0$$

satisfaisant les relations $P_n(x_i) = f(x_i)$. Ce qui revient à résoudre le système linéaire

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

Ce système admet une solution unique, car son déterminant, qui est un déterminant de Vandermonde, est non nul. CQFD.

Pour évaluer l'erreur d'interpolation, considérons P_n le polynôme d'interpolation de f aux points $x_0, x_1, ..., x_n$ et supposons que f soit de classe C^{n+1}. Il existe alors une constante c élément du plus petit intervalle contenant $x_0, x_1, ..., x_n$ et x tel que l'erreur d'interpolation soit égale à

$$R_n(x) = f(x) - P_n(x) = (x - x_0) \cdots (x - x_n) \frac{f^{(n+1)}(c)}{(n+1)!}$$

Cette erreur est majorée par

$$|R_n(x)| \leq \frac{(x_n - x_0)^{n+1}}{(n+1)!} \max_{x \in [a,b]} \left| f^{(n+1)}(x) \right|$$

La vérification de ce majorant est facile. Posons

$$\pi_n(x) = \prod_{j=0}^{n} (x - x_j)$$

et considérons la fonction $t \longmapsto g(t)$ définie pour $x \neq x_i$ par

$$g(t) = f(t) - P_n(t) - (f(x) - P_n(x)) \frac{\pi_n(t)}{\pi_n(x)}$$

g admet $(n+2)$ zéros aux points $x, x_0, x_1, ..., x_n$. D'après le théorème de Rolle, la fonction $g^{(n+1)}(t)$ a un zéro en un certain point c vérifiant

$$g^{(n+1)}(c) = f^{(n+1)}(c) - (f(x) - P_n(x)) \frac{(n+1)!}{\pi_n(x)} = 0$$

d'où

$$R_n(x) = f(x) - P_n(x) = \pi_n(x) \frac{f^{(n+1)}(c)}{(n+1)!}$$

Cette expression qui est démontrée pour $x \neq x_i$ est aussi vraie pour $x = x_i$; ce qui termine la démonstration.

2.2 Interpolation d'Hermite

Charles Hermite (1822-1901) a généralisé l'interpolation de Lagrange en faisant coïncider non seulement f et P_n aux points x_i, mais aussi leurs dérivées d'ordre k_i aux points x_i.

Soit $x_0, x_1, \ldots, x_n$ $(n+1)$ points distincts de l'intervalle $[a,b]$ et f une fonction définie sur $[a,b]$ admettant des dérivées jusqu'à l'ordre k_i aux points x_i. On pose $m = n + k_0 + k_1 + \cdots + k_n$. Il existe un polynôme unique P_m de degré $\leq m$ appelé *polynôme d'interpolation d'Hermite* tel que :

$$P_m^{(j)}(x_i) = f^{(j)}(x_i) \qquad \forall i = 0, \ldots, n \quad \forall j = 0, \ldots, k_i$$

L'interpolation de Lagrange est un cas particulier de l'interpolation d'Hermite ($k_0 = k_1 = \cdots = k_n = 0$). Le polynôme d'Hermite est donné par

$$P_m(x) = \sum_{i=0}^{n} \sum_{j=0}^{k_i} f^{(j)}(x_i) h_{ij}(x)$$

Les polynômes h_{ij} sont donnés par les relations de récurrence définies pour tout $j = 0, 1, \ldots, k_i - 1$

$$h_{ij}(x) = \frac{(x - x_i)^j}{j!} q_i(x) - \sum_{k=j+1}^{k_i} C_k^j q_i^{(k-j)}(x_i) h_{ik}(x)$$

et

$$h_{ik_i}(x) = \frac{(x - x_i)^{k_i}}{k_i!} q_i(x)$$

avec

$$q_i(x) = \prod_{\substack{l=0 \\ l \neq i}}^{n} \left(\frac{x - x_l}{x_i - x_l} \right)^{k_l + 1}$$

Dans le cas $k_0 = k_1 = ... = k_n = 1$, on a les expressions suivantes

$$P_m(x) = \sum_{i=0}^{n} r_i(x) f(x) + s_i(x) f'(x)$$

avec

$$r_i(x) = (1 - 2(x - x_i) l_i'(x)) l_i^2(x)$$

et

$$s_i(x) = (x - x_i) l_i^2(x)$$

où $l_i(x)$ est le polynôme de Lagrange.

Les équations $P_m^{(j)}(x_i) = f^{(j)}(x_i)$ forment un système linéaire à $(n+1)$ inconnues que sont les coefficients de P_m. Il suffit de montrer que le système homogène défini par les relations $P_m^{(j)}(x_i) = 0$ admet une solution unique, le vecteur nul. Ces relations impliquent que chaque x_i est racine d'ordre $k_i + 1$ du polynôme P_m. Par conséquent, P_m se met sous la forme

$$P_m(x) = q(x)\prod_{i=0}^{n}(x - x_i)^{k_i+1}$$

où $q(x)$ est un polynôme. La somme $(k_0 + 1) + \cdots + (k_n + 1) = m + 1$ montre que P_m ne peut être un polynôme de degré $\leq m$ que si q est nul. Par conséquent, P_m est nul. CQFD.

Le résultat suivant permet une évaluation de l'erreur dans le cas de l'interpolation d'Hermite. Soit f une fonction de classe C^{m+1} sur $[a, b]$, $x_0, x_1, ..., x_n$ $(n+1)$ points distincts de l'intervalle $[a, b]$ et $(k+1)$ entiers naturels $k_0, k_1, \ldots, k_n$. On note m l'entier $m = n + k_0 + \cdots + k_n$ et P_m le polynôme d'interpolation d'Hermite de f aux points $x_0, x_1, ..., x_n$. Alors, il existe une constante c (dépendant de x) élément du plus petit intervalle contenant $x_0, x_1, ..., x_n$ et x tel que l'erreur d'interpolation soit égale à

$$R_m(x) = f(x) - P_m(x) = (x - x_0)^{k_0+1} \cdots (x - x_n)^{k_n+1} \frac{f^{(m+1)}(c)}{(m+1)!}$$

Cette erreur est majorée par

$$|R_m(x)| \leq \frac{\left|(x - x_0)^{k_0+1} \cdots (x - x_n)^{k_n+1}\right|}{(m+1)!} \max_{x\in[a,b]} \left|f^{(m+1)}(x)\right|$$

2.3 Interpolation de Tchebychev

Contrairement aux interpolations précédentes dans lesquelles l'utilisateur peut choisir sa subdivision, l'interpolation de Pafnouti Tchebychev (1821-1894) impose une subdivision $x_0, x_1, ..., x_n$ de l'intervalle $[a, b]$ en des points appelés *points de Tchebychev*. L'interpolation utilise les polynômes orthogonaux de Tchebychev, seul cas (parmi les polynômes usuels) où les zéros des polynômes sont connus explicitement. L'interpolation de Tchebychev est encore appelée *interpolation de Lagrange aux points de Tchebychev,* car il s'agit d'une interpolation de Lagrange réalisée en des points particuliers.

Les *points d'interpolation de Tchebychev* d'ordre n sur l'intervalle $[-1, 1]$ sont les racines du polynôme de Tchebychev, qui correspondent aux points

$$u_i = \cos\frac{2(n-i)+1}{2n+2}\pi \qquad 0 \leq i \leq n$$

Les points de Tchebychev sur un intervalle $[a, b]$ quelconque sont définis par un simple changement de variables :

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{2(n-i)+1}{2n+2}\pi\right) \qquad 0 \leq i \leq n$$

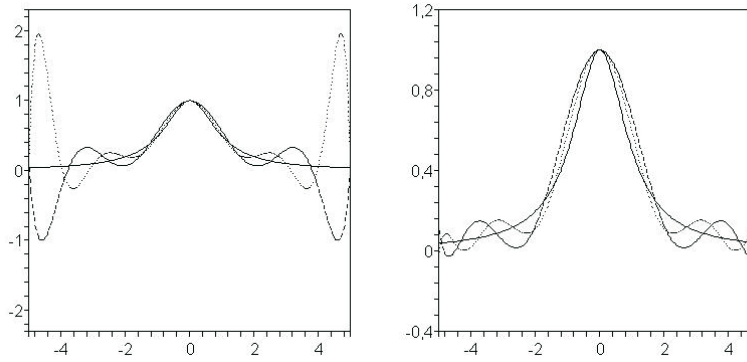
L'interpolation de Tchebychev est l'interpolation de Lagrange de f prise aux points de Tchebychev. L'erreur commise dans une interpolation de Tchebychev vérifie

$$R_n(x) = f(x) - P_n(x) = 2\left(\frac{b-a}{4}\right)^{n+1} \frac{f^{(n+1)}(c)}{(n+1)!}$$

Phénomène de Runge. On pourrait croire que la convergence du polynôme de Lagrange est d'autant meilleure que l'écart entre les points d'interpolation est plus petit. En réalité, il n'en est rien et Carl Runge (1856-1927) a montré en 1901 que quand n croît indéfiniment, le polynôme de Lagrange ne converge pas toujours vers la fonction interpolée en tous points. La divergence s'observe aux bords de l'intervalle : la convergence n'est pas uniforme. En revanche, dans l'interpolation de Tchebychev, il y a convergence uniforme, mais cette méthode impose à l'utilisateur le choix des points d'interpolation. Considérons la fonction

$$f(x) = \frac{1}{1+x^2} \qquad x \in [-5, 5]$$

et les graphes dans les deux interpolations :



Ces graphes montrent un comportement différent : c'est le *phénomène de Runge*. Sur chaque graphe, nous avons tracé la fonction f en trait plein et les polynômes d'interpolation de f pour $n = 8$ et $n = 10$. Dans l'interpolation de Lagrange (graphe de gauche), les points d'interpolation ($x_i = -5 + 10i/n$) sont régulièrement espacés. Lorsque le nombre de points

d'interpolation augmente, les valeurs de l'interpolation de Lagrange se confondent avec la courbe de $f(x)$, sauf au voisinage des bornes de l'intervalle, empêchant la convergence d'être uniforme. Plus on augmente le degré du polynôme de Lagrange, plus la différence au voisinage des bornes de l'intervalle entre la courbe $f(x)$ et le polynôme d'interpolation augmente. Dans l'interpolation de Tchebychev (graphe de droite), la convergence est uniforme. Lorsque le nombre de points d'interpolation augmente, la courbe polynomiale se confond avec la fonction.

2.4 Différences divisées

Au XVI^e siècle, les mathématiciens utilisaient des valeurs numériques tabulées à partir desquelles ils pratiquaient l'interpolation linéaire pour évaluer des valeurs intermédiaires. Cette technique n'étant pas toujours suffisamment précise, les méthodes de calcul à l'aide des différences finies se sont développées. Thomas Harriot (1560-1621), Henry Briggs (1561-1630), James Gregory (1638-1675) et Isaac Newton (1642-1727) ont élaboré la théorie des différences divisées.

Soit f une fonction continue de $[a,b]$ dans R et $x_0 < x_1 < \ldots < x_n$ une subdivision de l'intervalle $[a,b]$, on appelle *différence divisée d'ordre n* de f et on note $f[x_0,\ldots,x_n]$ le coefficient de x^n dans l'unique polynôme d'interpolation de Lagrange P_n de degré inférieur ou égal à n vérifiant $P_n(x_i) = f(x_i)$ pour $0 \leq i \leq n$.

Les propriétés des différences divisées sont les suivantes :
(1) *Formule de Newton*. Le polynôme d'interpolation, appelé dans ce cas *polynôme d'interpolation de Newton*, s'écrit

$$P_n(x) = f(x_0) + \sum_{j=0}^{n} f[x_0,\ldots,x_j](x-x_0)\cdots(x-x_{j-1})$$

(2) Les différences divisées s'expriment comme une différence

$$\forall k \in N^*, \qquad f[x_0,\ldots,x_k] = \frac{f[x_1,\ldots,x_k] - f[x_0,\ldots,x_{k-1}]}{x_k - x_0}$$

(3) *Formule de Leibniz*. Soit f,g,h trois fonctions définies sur l'intervalle $[a,b]$ et telles que $f = gh$, alors

$$f[x_0,\ldots,x_n] = \sum_{j=0}^{n} g[x_0,\ldots,x_j]\, h[xj,\ldots,x_n]$$

(4) Les différences divisées s'expriment comme une somme

$$f\,[x_0,\dots,x_n]=\sum_{j=0}^{n}\frac{f(x_j)}{\pi_n'(x_j)}$$

où $\pi_n'(x)$ est la dérivée de $\pi_n(x)=(x-x_0)(x-x_1)\cdots(x-x_n)$, c'est-à-dire

$$\pi_n'(x_k)=(x_k-x_0)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)$$

La vérification de ces propriétés est facile. La première propriété résulte du fait que P_k-P_{k-1} est un polynôme de degré inférieur ou égal à k admettant $f\,[x_0,\dots,x_k]$ comme différence divisée

$$P_k(x)-P_{k-1}(x)=f\,[x_0,\dots,x_k]\,(x-x_0)(x-x_1)\cdots(x-x_k)$$

En sommant sur l'indice k de 1 à n et en remarquant que $f\,[x_0]=f\,(x_0)$ on en déduit (1).
La deuxième propriété se démontre en considérant P le polynôme d'interpolation de f aux points $x_0,\dots,x_k$, Q le polynôme d'interpolation de f aux points $x_0,\dots,x_{k-1}$ et R le polynôme d'interpolation de f aux points $x_1,\dots,x_k$. Les expressions suivantes

$$P(x)=Q(x)+\frac{x-x_0}{x_k-x_0}\,(R(x)-Q(x))=\frac{(x-x_0)R(x)-(x-x_k)Q(x)}{x_k-x_0}$$

sont vraies, car les deux membres coïncident pour les valeurs $x_0,\dots,x_k$, et par unicité du polynôme sont donc égales pour tout x. L'équation précédente, écrite pour $x=x_k$, conduit à la formule proposée.

Formule d'Hermite-Genocchi. Soit f une fonction de classe C^n, la différence divisée s'écrit sous forme intégrale

$$f\,[x_0,\dots,x_n]=\textstyle\int_0^1 du_1\int_0^{1-u_1}du_2\int_0^{1-u_1-u_2}du_3\dots$$

$$\dots\int_0^{1-u_1\dots-u_{n-1}}f^{(n)}\,(x_0+(x_1-x_0)u_1+\dots+(x_n-x_0)u_n)\,du_n$$

Cette formule se démontre par récurrence sur n. Si $n=1$, on a

$$\int_0^1 f'(x_0+(x_1-x_0)u_1)du_1=\frac{f(x_1)-f(x_0)}{x_1-x_0}=f\,[x_0,x_1]$$

Supposons que la formule soit vraie jusqu'à l'ordre $(n-1)$. En intégrant, on obtient

$$\int_0^1 du_1\int_0^{1-u_1}du_2\dots$$
$$\int_0^{1-u_1\dots-u_{n-1}}f^{(n)}(x_0+(x_1-x_0)u_1+\dots+(x_n-x_0)u_n)\,du_n$$

$$= \int_0^1 du_1 \cdots \int_0^{1-u_1\ldots-u_{n-2}} \frac{1}{(x_n - x_0)} f^{(n-1)}(x_n + (x_1 - x_n)u_1 + \cdots$$

$$\cdots + (x_{n-1} - x_n)u_{n-1}) - f^{(n-1)}(x_0 + (x_1 - x_0)u_1 + \cdots + (x_{n-1} - x_0)u_{n-1})\ du_{n-1}$$

Puis, par changement de variables

$$v_1 = u_1, v_2 = u_3, \ldots, v_{n-2} = u_{n-1}, v_{n-1} = 1 - u_1 - \ldots - u_{n-1}$$

on obtient

$$= \int_0^1 dv_1 \cdots \int_0^{1-v_1\ldots-v_{n-2}} \frac{1}{(x_n - x_0)} f^{(n-1)}(x_1 + (x_2 - x_1)v_1 +$$

$$\cdots + (x_n - x_1)v_{n-1}) - f^{(n-1)}(x_0 + (x_1 - x_0)v_1 + \cdots + (x_{n-1} - x_0)v_{n-1})\ dv_{n-1}$$

En utilisant l'hypothèse de récurrence, on a

$$= \frac{1}{(x_n - x_0)} \left(f\left[x_1, \ldots, x_n\right] - f\left[x_0, \ldots, x_{n-1}\right] \right)$$

La propriété (2) permet alors de conclure.

Le théorème des résidus permet d'évaluer les différences divisées. Soit Ω un domaine simplement connexe du plan complexe dont le bord $\partial\Omega$ est réunion d'arcs de classe C^1 et contenant en son intérieur tous les points $z_0, z_{1,\ldots,}z_n$. Si $f(z)$ est analytique sur Ω et continue sur son bord, alors

$$f\left[z_0, \ldots, z_n\right] = \frac{1}{2i\pi} \int_{\partial\Omega} \frac{f(z)}{(z - z_0) \cdots (z - z_n)} dz$$

Nous supposons maintenant que les points $x_0 < x_1 < \ldots < x_n$ de l'intervalle de $[a, b]$ sont régulièrement espacés $x_{i+1} - x_i = h$ pour $0 \leq i \leq n - 1$. Dans ce cas, l'expression des différences divisées se simplifie.

L'*opérateur des différences progressives*, des *différences à droite* ou encore *opérateur de Bernoulli progressif* est l'opérateur Δ défini par

$$\Delta f(x) = f(x + h) - f(x)$$

Par récurrence, on définit l'opérateur Δ^k par

$$\Delta^k f(x) = \Delta \left(\Delta^{k-1} f \right)(x)$$

On note

$$\Delta_i^k = \Delta^k f(x_i) \qquad f_i = f(x_i) \qquad \text{et} \qquad \Delta^0 f(x_i) = f(x_i)$$

En particulier

$$\begin{aligned}\Delta_i^1 &= f_{i+1} - f_i \\ \Delta_i^2 &= f_{i+2} - 2f_{i+1} + f_i \\ \Delta_i^3 &= f_{i+3} - 3f_{i+2} + 3f_{i+1} - f_i\end{aligned}$$

Ces quantités se calculent facilement en les présentant sous la forme d'un tableau dans lequel chaque élément de celui-ci est la différence des deux éléments voisins situés sur la ligne précédente

$$\begin{array}{ccccccccc} f_0 & & f_1 & & f_2 & & f_3 & & f_4 \\ & \Delta f_0 & & \Delta f_1 & & \Delta f_2 & & \Delta f_3 & \\ & & \Delta^2 f_0 & & \Delta^2 f_1 & & \Delta^2 f_2 & & \\ & & & \Delta^3 f_0 & & \Delta^3 f_1 & & & \\ & & & & \Delta^4 f_0 & & & & \end{array}$$

Propriétés. (1) Les différences progressives satisfont la relation

$$\Delta^k f(x_j) = \Delta_{j+1}^k - \Delta_j^{k-1} = \sum_{j=0}^{k} (-1)^j C_k^j f(x_{i+k-j})$$

(2) Les différences divisées et les différences progressives sont liées par la relation

$$f\left[x_i, \ldots, x_{i+k}\right] = \frac{\Delta^k f(x_i)}{k!\, h^k} \qquad \text{avec} \quad h = x_{i+1} - x_i$$

(3) Si on pose $x = x_0 + \alpha h$, la formule de Newton se simplifie

$$\begin{aligned}&(x - x_0)(x - x_1) \ldots (x - x_{k-1}) f\left[x_0, x_1, \ldots, x_k\right] \\ = \quad &\frac{\alpha(\alpha - 1) \ldots (\alpha - k + 1)}{k!} \Delta^k f(x_0)\end{aligned}$$

De la même façon, on construit l'opérateur des différences régressives. L'*opérateur des différences régressives*, des *différences à gauche* ou encore *opérateur de Bernoulli régressif* est l'opérateur ∇ défini par

$$\nabla f(x) = f(x) - f(x - h)$$

et par récurrence, l'opérateur ∇^k est défini par

$$\nabla^k f(x) = \nabla \left(\nabla^{k-1} f \right)(x)$$

On note

$$\nabla_i^k = \nabla^k f(x_i) \qquad f_i = f(x_i) \qquad \text{et} \qquad \nabla^0 f(x_i) = f(x_i)$$

En particulier

$$\begin{aligned}\nabla_i^1 &= f_i - f_{i-1}\\ \nabla_i^2 &= f_i - 2f_{i-1} + f_{i-2}\\ \nabla_i^3 &= f_i - 3f_{i-1} + 3f_{i-2} - f_{i-3}\end{aligned}$$

Les différences régressives satisfont la relation

$$\nabla_j^k = \sum_{j=0}^{k} (-1)^{j+1} C_k^j f_{i-k+j}$$

De manière analogue, on introduit aussi l'opérateur des différences centrées.

L'*opérateur des différences centrées ou centrales* est l'opérateur noté δ défini par

$$\delta f(x) = f(x + \frac{h}{2}) - f(x - \frac{h}{2})$$

et par récurrence, on définit l'opérateur δ^k par

$$\delta^k f(x) = \delta\left(\delta^{k-1} f\right)(x)$$

Les *différences centrées* ont été introduites en 1899 par le mathématicien William Sheppard (1863-1936). Pour les points x_i, on note

$$\delta_i^k = \delta^k f(x_i) \qquad \text{avec} \quad \delta^0 f(x_i) = f(x_i) = f_i$$

En particulier

$$\begin{aligned}\delta f(x_{i+1/2}) &= f_{i+1} - f_i\\ \delta_i^2 &= f_{i+1} - 2f_i + f_{i-1}\\ \delta_{i+1/2}^3 &= f_{i+2} - 3f_{i+1} + 3f_i - f_{i-1}\end{aligned}$$

Propriétés. (1) L'opérateur de Bernoulli de puissance paire s'écrit

$$\delta_i^{2k} = \sum_{j=0}^{2k} (-1)^j C_{2k}^j f_{i+k-j}$$

(2) L'opérateur de Bernoulli de puissance impaire s'exprime par la relation

$$\delta_{i+\frac{1}{2}}^{2k+1} = \sum_{j=0}^{2k+1} (-1)^j C_{2k+1}^j f_{i+k+1-j}$$

(3) Si n et k ont même parité,

$$\delta_{i/2}^k = \Delta_{(n-k)/2}^k$$

Remarquons que les opérateurs de Bernoulli s'expriment par les opérateurs de translation

$$\tau_h f(x) = f(x-h)$$

et par les formules

$$\Delta = \tau_{-1} - I \qquad \nabla = I - \tau_1 \qquad \delta = \tau_{-1/2} - \tau_{1/2}$$

ce qui permet de développer un calcul symbolique qui a été utilisé par Lagrange à la fin du XVIII^e siècle.

La formule d'interpolation de Newton a été établie indépendamment par James Gregory en 1670 et par Isaac Newton en 1675. Cette formule avait été donnée quelques années auparavant par Thomas Harriot (en 1610). Elle correspond à la formule d'Euler-Mac Laurin tronquée à l'ordre n et dans laquelle dérivées et différences divisées se correspondent. Le polynôme d'interpolation de Newton n'est autre que le polynôme de Lagrange écrit en utilisant les différences divisées.

Soit f une fonction continue d'un intervalle $[a,b]$ dans $\mathbb{R}$ et $x_0 < x_1 < ... < x_n$ $(n+1)$ points distincts de l'intervalle $[a,b]$. On suppose que les points x_i sont régulièrement espacés et on note $h = x_{i+1} - x_i$ la différence entre deux points consécutifs. On pose $x = x_0 + \alpha h$. L'interpolation de Newton, pour laquelle le polynôme d'interpolation de Lagrange s'écrit (*Formule de Newton progressive*)

$$P_n(x) = f(x_0) + \sum_{k=0}^{n} \frac{\alpha(\alpha-1)\dots(\alpha-k+1)}{k!} \Delta^k f(x_0)$$

nécessite (n^2+3n) additions, n multiplications et $(n^2+n)/2$ divisions. L'erreur d'interpolation vaut

$$R_n(x) = f(x) - P_n(x) = \frac{\alpha(\alpha-1)\dots(\alpha-n)}{(n+1)!} h^{n+1} f^{(n+1)}(c)$$

Le nombre c dépend de n et de x et appartient au plus petit intervalle contenant x_0, x_n et x. L'erreur est majorée par

$$|R_n(x)| \leq \frac{h^{n+1}}{n+1} \sup_{i=0..n} f^{(n+1)}(x_i)$$

La vérification s'effectue en utilisant les propriétés des différences divisées. Si on pose $x = x_0 + \alpha h$, la propriété (ii) des différences progressives s'écrit

$$\begin{aligned} & (x - x_0)(x - x_1) \dots (x - x_{k-1}) f[x_0, x_1, \dots, x_k] \\ = \quad & \frac{\alpha(\alpha - 1) \dots (\alpha - k + 1)}{k!} \Delta^k f(x_0) \end{aligned}$$

En substituant dans le polynôme d'interpolation de Lagrange écrit sous la forme

$$P_n(x) = f(x_0) + \sum_{kj=0}^{n} f[x_0, \dots, x_k] (x - x_0) \cdots (x - x_{k-1})$$

on obtient la formule de Newton progressive. Dans l'interpolation de f par $P_n(x)$ le reste, qui est donné par

$$R_n(x) = f(x) - P_n(x) = \int_{x_0}^{x} f^{(n+1)}(t) \frac{(x - t)^n}{n!} dt$$

vérifie la propriété suivante

$$\exists\, \xi \in]x_0, x_n[\qquad R_n(x) = (x - x_0) \cdots (x - x_n) \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

d'où découle la majoration.

En utilisant les différences régressives et en posant $\beta = (x - x_n)/h$, le polynôme d'interpolation s'écrit (*Formule de Newton régressive*)

$$P_n(x) = f(x_n) + \sum_{k=1}^{n} \frac{\beta(\beta + 1) \dots (\beta + k - 1)}{k!} \nabla^k f(x_n) = \sum_{k=0}^{n} C_{\beta+k-1}^{k} \nabla^k f(x_0)$$

L'erreur d'interpolation devient

$$R_n(x) = f(x) - P_n(x) = \frac{\beta(\beta + 1) \dots (\beta + n)}{(n+1)!} h^{n+1} f^{(n+1)}(c)$$

Le nombre c (qui dépend de n) appartient au plus petit intervalle contenant x_0, x_n et x.

Remarquons que les polynômes de Newton

$$N_k(\alpha) = \frac{\alpha(\alpha - 1) \dots (\alpha - k + 1)}{k!} \qquad k = 0, \dots, n$$

forment une base de l'espace des polynômes de degré inférieur ou égal à n. Ils vérifient la relation de récurrence

$$N_{k-1}(\alpha) = N_k(\alpha) - N_k(\alpha - 1)$$

Lorsqu'on utilise les différences centrales, on établit la *formule de Laplace-Everett*. Cette formule était connue de Pierre Simon Laplace (1749-1827) qui l'utilisa dans sa *Théorie analytique des probabilités* publiée en 1812. Elle a été établie par J.D. Everett en 1900.

$$f(x_0+\alpha h) = (1-\alpha)f_0 + \alpha f_1 - \frac{\alpha(\alpha-1)(\alpha-2)}{3!}\delta_0^2 + \frac{(\alpha+1)\alpha(\alpha-1)}{3!}\delta_1^2$$

$$+\cdots - \quad \frac{(\alpha+n-1)(\alpha+n-2)\ldots(\alpha-n+3)}{(2n+1)!}\delta_0^{2n} +$$

$$+ \quad \frac{(\alpha+n)(\alpha+n-1)\ldots(\alpha-n)}{(2n+1)!}^{2n}\delta_1 + R_{2n}(x)$$

Le reste $R_{2n}(x)$ est donné par

$$R_{2n}(x) = h^{2n+2}\frac{(\alpha+n)(\alpha+n-1)\ldots(\alpha-n-1)}{(2n+2)!}f^{(2n+2)}(c)$$

Le nombre c (qui dépend de n) appartient à l'intervalle $x_0 \leq c \leq x_n$.

D'autres formules peuvent être établies à partir des différences centrées. En un point $x = x_j + \alpha h$, on a la *formule de Newton-Stirling*. Cette formule, connue de Newton, a été étudiée par James Stirling (1692-1770) en 1730.

$$\begin{aligned} f(x) \quad \simeq \quad & f(x_j) + \alpha\delta_j^1 + \frac{\alpha^2}{2!}\delta_j^2 + \frac{\alpha(\alpha^2-1)}{3!}\delta_j^3 + \frac{\alpha(\alpha^2-1)}{4!}\delta_j^4 \\ & + \frac{\alpha(\alpha^2-1)(\alpha^2-4)}{5!}\delta_j^5 + \cdots \end{aligned}$$

La *formule de Newton-Bessel,* qui figure dans le *Methodus Differentialis* de Newton sous une forme légèrement différente, a été étudiée par Friedrich Bessel (1784-1846).

$$f(x) \simeq f(x_{j+\frac{1}{2}}) + \alpha\delta_{j+\frac{1}{2}}^1 + \frac{(\alpha^2-\frac{1}{4})}{2!}\delta_{j+\frac{1}{2}}^2 + \frac{\alpha(\alpha^2-\frac{1}{4})}{3!}\delta_{j+\frac{1}{2}}^3 +$$

$$\frac{(\alpha^2-\frac{1}{4})(\alpha^2-\frac{9}{4})}{4!}\delta_{j+\frac{1}{2}}^4 + \cdots$$

2.5 Algorithme de Neville-Aitken

Alexander Craig Aitken (1895-1967) puis Eric Harold Neville (1889-1961) ont proposé un algorithme récurrent de calcul du polynôme d'interpolation de Lagrange sur n points à partir d'une expression portant sur $(n-1)$

points. Pour calculer une interpolation de f en un point x, on peut utiliser les formules des différences divisées. Mais la méthode proposée par Aitken en 1932 évite le calcul des coefficients du polynôme, et ne suppose pas que les points x_i sont uniformément répartis. Elle se fonde sur la proposition suivante :

Soit $f(x\,|x_p,\ldots,x_q)$ l'unique polynôme d'interpolation de degré $(q-p-1)$ qui coïncide avec $f(x)$ aux points $x_p,\ldots,x_q$. On a la relation de récurrence

$$(x_k-x_j)f(x\,|x_p,\ldots,x_q) = \begin{vmatrix} (x_k-x) & f(x\,|x_p,\ldots,x_{j-1},x_{j+1},\ldots,x_q) \\ (x_j-x) & f(x\,|x_p,\ldots,x_{k-1},x_{k+1},\ldots,x_q) \end{vmatrix}$$

Exemple. Supposons connues les valeurs f_0, f_1, f_2, f_3 de f aux points x_0, x_1, x_2, x_3 et calculons pour $P(x)$ un polynôme de degré 3, sa valeur au point x. On calcule d'abord les valeurs $P(x\,|x_0,x_j)$ par la relation

$$(x_j-x_0)P(x\,|x_0,x_j) = \begin{vmatrix} f_0 & x_0-x \\ f_j & x_j-x \end{vmatrix}$$

puis les valeurs $P(x\,|x_0,x_i,x_j)$ par

$$(x_j-x_i)P(x\,|x_0,x_i,x_j) = \begin{vmatrix} P(x\,|x_0,x_i) & x_i-x \\ P(x\,|x_0,x_j) & x_j-x \end{vmatrix}$$

et enfin la valeur cherchée $P(x)=P(x\,|x_0,x_1,x_2,x_3)$ par

$$(x_3-x_2)P(x\,|x_0,x_1,x_2,x_3) = \begin{vmatrix} P(x\,|x_0,x_1,x_2) & x_2-x \\ P(x\,|x_0,x_1,x_3) & x_3-x \end{vmatrix}$$

La méthode de Neville-Aitken nécessite (n^2+2n+1) additions, (n^2+n) multiplications et $\frac{1}{2}(n^2+n)$ divisions. En pratique, on dispose le calcul du polynôme d'interpolation sous la forme d'une table. Notons $P_{i+1,j}$ le polynôme d'interpolation de Lagrange aux points $x_j, x_{j+1},\ldots,x_{j+i+1}$. La formule de la proposition

$$(x_{i+j+1}-x_j)P_{i+1,j} = (x_{i+j+1}-x)P_{i,j} - (x_j-x)P_{i,j+1}$$

Supposons connues les valeurs $f(x_i)=f_i=P_{0,i}$ pour $i=0,1,2,3$.

$$\begin{array}{llll} f_0=P_{0,0} & P_{1,0} & P_{2,0} & P_{3,0} \\ f_1=P_{0,1} & P_{1,1} & P_{2,1} & \\ f_2=P_{0,2} & P_{1,2} & & \\ f_3=P_{0,3} & & & \end{array}$$

On remplit successivement chaque colonne du tableau en utilisant la formule précédente. La valeur $P_{3,0}$ donne la valeur cherchée.

2.6 Meilleure approximation

Déterminer la meilleure approximation n'est pas toujours facile. Elle dépend de la topologie des espaces mis en œuvre. Soit E un espace métrique, A un sous-ensemble de E et f un élément de E. On dit qu'un élément φ de A est une meilleure approximation de f si

$$\|f - \varphi\| = d(f, A) = \inf_{a \in A} \|f - a\|$$

On démontre que si A est compact, alors pour toute fonction f de E il existe au moins une meilleure approximation.

En effet, soit f un élément quelconque de E, notons d la distance de f à A et considérons la suite (a_n) d'éléments de A tels que

$$\lim_{n \to \infty} \|f - a_n\| = d(f, A) = d$$

D'où on déduit l'inégalité

$$\forall \varepsilon > 0, \qquad \exists\ k_1 > 0, \qquad n \geq k_1 \Longrightarrow \qquad \|f - a_n\| \leq d + \varepsilon$$

Comme A est compact, cette suite admet une limite φ, par conséquent

$$\forall \varepsilon > 0, \qquad \exists\ k_2 > 0, \qquad n \geq k_2 \Longrightarrow \qquad \|a_n - \varphi\| \leq \varepsilon$$

En additionnant les deux inégalités, $\forall \varepsilon > 0, \quad \exists\ k > 0,$

$$n \geq k \Longrightarrow \quad \|f - u\| \leq \|f - a_n\| + \|a_n - \varphi\| \leq d + \varepsilon$$

on en déduit que

$$\|f - \varphi\| \leq \|f - a\| \qquad \forall a \in A$$

La limite de la suite est donc une meilleure approximation.

De ce résultat, on déduit que si E est un espace vectoriel de dimension finie et A est un sous-espace vectoriel de E, alors pour toute fonction f de E, il existe au moins une meilleure approximation.

En effet, soit B le sous-ensemble de A

$$B = \{b \in A : \qquad \|b\| \leq 2\,\|f\|\}$$

Cet ensemble est non vide (car il contient au moins 0), fermé, borné : donc B est compact. D'après le résultat précédent, il existe au moins une meilleure approximation de f, notée φ et vérifiant

$$\|f - \varphi\| \leq \|f - b\| \qquad \forall b \in B$$

Considérons maintenant un élément a de A qui n'appartient pas à B, on a

$$\forall a \in A \setminus B, \qquad \|a - f\| \geq \|a\| - \|f\| > \|f\| \geq \|\varphi - f\|$$

ce qui permet d'étendre l'inégalité précédente à tous les éléments de A donc a est une meilleure approximation de f.

Soit E un espace vectoriel normé contenant l'espace $\mathcal{P}_n$ des polynômes de degré inférieur ou égal à n. Pour toute fonction f de E, il existe au moins un polynôme $p_n \in \mathcal{P}_n$ tel que :

$$\|f - p_n\| = \inf_{p \in \mathcal{P}_n} \|f - p\|$$

Le polynôme p_n est appelé polynôme de meilleure approximation pour la norme de E.

Il est facile de vérifier ce résultat. Il existe une suite g_n telle que

$$\lim_n \|f - g_n\| = \inf_{p \in E} \|f - p\|$$

Une telle suite est bornée car $\|g_n\| \leq \|f - g_n\| + \|f\|$ et comme $\mathcal{P}_n$ est de dimension finie, on peut extraire une sous-suite convergeant vers un polynôme p_n qui minimise la distance de f à P_n.

Remarque. L'unicité de la meilleure approximation est un problème difficile. Cette unicité n'est pas toujours vérifiée. Par exemple, si E est l'ensemble des fonctions continues sur $[-1, 1]$ et intégrables sur cet intervalle $E = L^1$ et A est l'ensemble des polynômes de degré inférieur ou égal à 1, $A = P_1$. La fonction $f(x) = 1$ si $x > 0$, et -1 si $x < 0$ admet pour meilleure approximation toutes les fonctions constantes $\varphi(x) = c$, avec $c \in [-1, 1]$ et dans ce cas

$$\|f - \varphi\| = \int_{-1}^{1} |f(x) - \varphi(x)|\, dx = 2$$

On démontre que lorsque E est l'espace des fonctions continues sur un intervalle $[a, b]$, $E = \mathcal{C}[a, b]$ et A est l'espace des polynômes de degré n, $A = \mathcal{P}_n$, le polynôme de meilleure approximation est unique, mais lorsque E est l'espace de Lebesque $E = L^p[a, b]$ et $A = \mathcal{P}_n$, ce polynôme n'est pas toujours unique. Lorsque $E = \mathcal{C}[a, b]$ est muni de la convergence uniforme, on parle de *meilleure approximation uniforme, de Tchebychev, ou minimax.* Lorsque $E = L^2[a, b]$ on parle de *meilleure approximation quadratique* ou *meilleure approximation au sens des moindres carrés.*

On démontre le résultat suivant qui introduit la *constante de Lebesgue*. Soit E un espace vectoriel normé, A un sous-espace vectoriel de E de dimension finie. Soit u l'opérateur linéaire de $E \to A$ défini par $u(f) = \varphi$ où φ est

la meilleure approximation de f et vérifiant $u(a) = a,\ \forall a \in A$. Pour toute fonction f de E, l'erreur d'approximation est majorée par

$$\|f - u(f)\| \leq (1 + \|u\|)d(f, A)$$

Le nombre $\Lambda = \|u\|$ est appelé *constante de Lebesgue*. Il mesure l'amplification de l'erreur. La vérification est immédiate. Soit φ la meilleure approximation de f sur A. Comme $\varphi \in A$, on a $u(\varphi) = \varphi$. Par conséquent

$$\|f - u(f)\| \leq \|f - \varphi\| + \|u(f - \varphi)\| \leq (1 + \|u\|)d(f, A)$$

2.7 Approximation uniforme

Le théorème de Karl Weierstrass (1815-1897), établi en 1885, affirme que toute fonction continue sur un intervalle I peut être approchée uniformément par une suite de polynômes. Le problème est de trouver un moyen de construire cette suite de polynômes. On considère l'espace vectoriel des fonctions continues de $[a, b]$ dans $\mathbb{R}$ muni de la norme uniforme

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

et on note

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

la distance de f à l'ensemble des polynômes de degré inférieur ou égal à n.

Soit f une fonction de $\mathcal{C}([a, b])$ continue sur l'intervalle $[a, b]$ et à valeurs dans $\mathbb{R}$. Pour tout entier n naturel, il existe un et un seul polynôme q_n qui réalise le minimum de

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

Ce polynôme est appelé *polynôme de meilleure approximation uniforme*. Dans le cas complexe, le théorème de Mergelyan établit que si A est un compact de $\mathbb{C}$, f une fonction continue sur A et analytique sur l'intérieur de A, f admet une meilleure approximation polynomiale si et seulement si le complémentaire de A est connexe.

On dit que $f \in \mathcal{C}([a, b])$ équioscille sur $(n+1)$ points de l'intervalle $[a, b]$ s'il existe des points $x_0 < x_1 < ... < x_n$ tels que $|f(x_i)| = \|f\|$ pour $i = 0, ...n$ et $f(x_{i+1}) = -f(x_i)$ pour $i = 0, ..., n - 1$.

On démontre que si $f \in \mathcal{C}([a, b])$ est une fonction continue, le polynôme de meilleure approximation $q_n \in \mathcal{P}_n$ de f est l'unique polynôme de degré n tel que $(f - q_n)$ équioscille sur au moins $(n+2)$ points de l'intervalle $[a, b]$ et que si f est une fonction analytique de série entière de rayon R centrée au point

$(a+b)/2$, alors les polynômes d'interpolation p_n convergent uniformément vers f si

$$R > (\frac{1}{c} + \frac{1}{2})(b-a)$$

la constante c étant donnée par $c = 1$ si les points x_i sont quelconques, $c = e$ si les points x_i sont équidistants, et $c = 4$ si les points x_i sont les points de Tchebychev.

Rappelons qu'une fonction $f : I \to \mathbb{R}$ est *lipschitzienne* de rapport k ou k-lipschitzienne si

$$\forall x, y \in I \qquad |f(x) - f(y)| \leq k\,|x - y|$$

L'ensemble des fonctions *k-lipschitziennes* est un espace vectoriel. Muni de la norme uniforme, l'espace $Lip(I)$ est un espace de Banach (espace vectoriel normé complet).

Une fonction $f : I \to \mathbb{R}$ est *höldérienne* de coefficient α avec $0 < \alpha \leq 1$ et de rapport k si

$$\forall x, y \in I \qquad |f(x) - f(y)| \leq k\,|x - y|^{\alpha}$$

L'ensemble des fonctions höldériennes $Lip_{\alpha,k}(I)$ de coefficient α et de rapport k est un espace vectoriel. Muni de la norme uniforme, c'est un espace de Banach.

Le *module de continuité* d'une fonction f est la fonction $\omega_f : \mathbb{R}^+ \to \mathbb{R}^+$ définie par

$$\omega_f(t) = \sup_{|x-y| \leq t,\ x,y \in [a,b]} \{|f(x) - f(y)|\}$$

Ce module vérifie différentes propriétés.

$$\forall x, y \in [a, b], \qquad |f(x) - f(y)| \leq \omega_f(|x - y|)$$

La fonction $t \longmapsto \omega_f(t)$ est une fonction croissante et la limite vérifie

$$\lim_{t \to 0^+} \omega_f(t) = 0$$

La fonction ω_f est sous-additive

$$\forall s, t \in \mathbb{R}^+, \qquad \omega_f(s + t) \leq \omega_f(s) + \omega_f(t)$$

et vérifie

$$\forall n \in \mathbb{N}, \forall t > 0 \qquad \omega_f(nt) \leq n\omega_f(t)$$

ainsi que la propriété suivante

$$\forall \alpha > 0, \forall t > 0 \qquad \omega_f(\alpha t) \leq (1+\alpha)\omega_f(t)$$

Le module de continuité permet de caractériser différentes classes de fonctions. La fonction f est uniformément continue sur I si et seulement si

$$\lim_{t \to 0} \omega_f(t) = 0$$

La fonction f est lipschitzienne de rapport k si et seulement si

$$\omega_f(t) \leq k\omega_f(t)$$

La fonction f est höldérienne d'ordre α pour $0 < \alpha < 1$ si et seulement si

$$\omega_f(t) \leq kt^{\alpha}$$

Le module de continuité permet d'établir un résultat de Jackson. Soit $f \in \mathcal{C}([a,b])$, il existe des polynômes p_n de degré n appelés *polynômes de Jackson* vérifiant

$$\|f - p_n\| \leq 3\omega_f(\frac{b-a}{n+2})$$

La constante de Lebesgue Λ_n est de l'ordre de

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

si les points (x_i) sont équidistants et

$$\Lambda_n \simeq \frac{2}{\pi}\ln(n)$$

si les points (x_i) sont les points de Tchebychev. Ce résultat permet de montrer que si f est k-lipschitzienne, les polynômes d'interpolation de Tchebychev convergent vers f sur l'intervalle $[a,b]$.

2.8 Polynômes orthogonaux

Soit I un intervalle borné ou non de $\overline{\mathbb{R}}$, $\omega : I \to \overline{\mathbb{R}}$ une fonction numérique continue positive appelée *poids* telle que

$$\forall n \in \mathbb{N}, \qquad \int_I |x|^n \omega(x)dx < \infty$$

On considère l'espace vectoriel E des fonctions continues de I dans K telles que

$$\int_I |f(x)|^2 \omega(x)dx < \infty$$

L'espace E muni du produit scalaire

$$\langle f, g\rangle = \int_I \overline{f(x)} g(x)\omega(x)dx$$

est un espace préhilbertien. Les polynômes orthogonaux relativement à la fonction de poids $\omega(x)$ sont les polynômes Ψ_n de degré n vérifiant la relation d'orthogonalité

$$\langle \Psi_n, \Psi_m\rangle = \int_I \Psi_n(x)\Psi_m(x)\omega(x)dx = 0 \qquad n \neq m, \text{ et } n, m = 0, 1, 2, ...$$

Il existe une suite *unique* de polynômes orthogonaux pour le produit scalaire de E de degré n à coefficients réels et dont le terme de plus haut degré est x^n. La démonstration de ce résultat se fait par récurrence sur l'entier n. On construit Ψ_n par le procédé d'orthogonalisation de Gram-Schmidt. Supposons la propriété établie jusqu'à l'ordre $(n-1)$ et soit $(\Psi_0, \Psi_1, \ldots, \Psi_{n-1})$ une base de l'espace $\mathcal{P}_{n-1}$, construisons un polynôme sous la forme

$$\Phi_n(x) = x^n - \sum_{k=0}^{n-1} a_k \Psi_k$$

Comme le produit scalaire $\langle \Phi_n, \Psi_k\rangle = 0$ s'annule pour les valeurs de $k = 0, ..., n-1$, on en déduit que

$$a_k = \frac{\langle x^n, \Psi_k\rangle}{\|\Psi_k\|_2^2}$$

Le polynôme est ensuite normé

$$\Psi_n = \frac{\Phi_n}{\|\Phi_n\|_2}$$

En pratique, les polynômes orthogonaux ne sont pas toujours de norme 1, on emploie souvent $\lambda_n\Psi_n$ au lieu de Ψ_n. Donnons des exemples de polynômes orthogonaux :

Polynômes de Legendre. Sur l'intervalle $[-1, +1]$, la fonction de pondération vaut $\omega(x) = 1$. Les polynômes sphériques ou *polynômes de Legendre* sont définis par la relation de récurrence

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x)$$

et les conditions initiales

$$P_0(x) = 1 \quad P_1(x) = x$$

Leur norme vérifie la relation

$$\|P_n\|_2^2 = \int_{-1}^{+1} P_n^2(x)dx = \frac{2}{2n+1}$$

Polynômes de Laguerre. Sur l'intervalle $[0, +\infty[$, la fonction de pondération vaut $\omega(x) = e^{-x}$. Les polynômes de Laguerre sont définis par la relation de récurrence

$$(n+1)L_{n+1}(x) = (2n+1-x)L_n(x) - nL_{n-1}(x)$$

et les conditions initiales

$$L_0(x) = 1 \quad L_1(x) = 1 - x$$

Leur norme vérifie la relation

$$\|L_n\|_2^2 = \int_0^{\infty} [L_n(x)]^2 e^{-x}dx = 1$$

Soit $\alpha > -1$. On considère sur l'intervalle $[0, +\infty[$, la fonction de pondération $\omega(x) = x^{\alpha}e^{-x}$. On définit les *polynômes de Laguerre généralisés* par la relation de récurrence

$$(n+1)L_{n+1}^{(\alpha)}(x) = (2n+\alpha+1-x)L_n^{(\alpha)}(x) - (n+\alpha)L_{n-1}^{(\alpha)}(x)$$

et les conditions intiales

$$L_0^{(\alpha)}(x) = 1 \quad L_1^{(\alpha)}(x) = 1 + \alpha - x$$

Le polynôme de Laguerre proprement dit correspond au cas $\alpha = 0$. Leur norme vérifie la relation

$$\left\|L_n^{(\alpha)}\right\|_2^2 = \int_0^{\infty} [L_n^{(\alpha)}(x)]^2 x^{\alpha}e^{-x}dx = \frac{\Gamma(\alpha+n+1)}{n!}$$

Polynôme de Tchebychev. Sur l'intervalle $[-1, 1]$, on considère la fonction de pondération

$$\omega(x) = \frac{1}{\sqrt{1-x^2}}$$

Les *polynômes de Tchebychev de première espèce* sont définis par la relation de récurrence

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

et les conditions initiales

$$T_0(x) = 1 \quad T_1(x) = x$$

Leur norme vérifie la relation

$$\|T_n\|_2^2 = \int_{-1}^{+1} T_n^2(x) \frac{dx}{\sqrt{1-x^2}} = \begin{cases} \frac{\pi}{2} & \text{si } n \neq 0 \\ \pi & \text{si } n = 0 \end{cases}$$

Les *polynômes de Tchebychev de deuxième espèce* sont définis par les relations de récurrence

$$U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x)$$

et les conditions initiales

$$U_0(x) = 1 \quad U_1(x) = 2x$$

Ils sont liés aux polynômes de première espèce par les relations

$$U_n(x) = \frac{1}{1-x^2}(xT_{n+1} - T_{n+2}) \quad U_n(1) = n+1$$

Leur norme vérifie la relation

$$\|U_n\|_2^2 = \int_{-1}^{+1} U_n^2(x) \frac{dx}{\sqrt{1-x^2}} = \frac{\pi}{2}$$

Polynômes d'Hermite. On considère que sur $\mathbb{R}$, la fonction de pondération vaut $\omega(x) = e^{-x^2}$. Les polynômes d'Hermite sont définis par la relation de récurrence

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$$

et les conditions initiales

$$H_0(x) = 1 \quad H_1(x) = 2x$$

Leur norme vérifie la relation

$$\|H_n\|_2^2 = \int_{-\infty}^{+\infty} H_n^2(x) e^{-x^2} dx = \sqrt{\pi} 2^n n!$$

Polynômes de Gegenbauer. Soit $\alpha > -1/2$, on considère sur l'intervalle $]-1, +1[$, la fonction de pondération

$$\omega(x) = (1-x^2)^{\alpha - \frac{1}{2}}$$

Les *polynômes ultrasphériques* ou *polynômes de Gegenbauer* sont définis par la relation de récurrence

$$(n+1)G_{n+1}^{(\alpha)}(x) = 2(n+\alpha)G_n^{(\alpha)}(x) - (n+2\alpha-1)G_{n-1}^{(\alpha)}(x)$$

et les conditions initiales

$$G_0^{(\alpha)}(x) = 1 \quad G_1^{\alpha}(x) = 2\alpha x \quad (\text{si } \alpha \neq 0), \quad G_1^0 = 2x$$

La norme au carré de ces polynômes est donnée par la relation

$$\int_{-1}^{+1} [G_n^{(\alpha)}(x)]^2 (1-x^2)^{\alpha-\frac{1}{2}} dx = \begin{cases} \dfrac{\pi 2^{1-2\alpha}}{n!(n+\alpha)!} \dfrac{\Gamma(n+2\alpha)}{\Gamma(\alpha)^2} & \text{si } \alpha \neq 0 \\ \dfrac{2\pi}{n^2} & \text{si } \alpha = 0 \end{cases}$$

Polynômes de Jacobi. Soit $\alpha > -1$ et $\beta > -1$, on considère sur l'intervalle $]-1, +1[$, la fonction de pondération

$$\omega(x) = (1-x)^\alpha (1+x)^\beta$$

Les polynômes de Jacobi sont définis par la relation de récurrence

$$a_n J_{n+1}(x) = (b_n + x c_n) J_n(x) - d_n J_{n-1}(x)$$

et les conditions initiales

$$J_0(x) = 1 \quad J_1(x) = (\alpha - \beta)/2 + (1 + (\alpha+\beta)/2)x$$

avec

$$\begin{aligned} a_n &= 2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta) \\ b_n &= (2n+\alpha+\beta+1)(\alpha^2-\beta^2) \\ c_n &= (2n+\alpha+\beta) \\ d_n &= 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2) \end{aligned}$$

La norme au carré de ces polynômes est donnée par la relation

$$\int_{-1}^{+1} [J_n(x)]^2 (1-x)^\alpha (1+x)^\beta dx = \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{n!\Gamma(n+\alpha+\beta+1)}$$

Propriétés. Les polynômes orthogonaux ont certaines propriétés communes. (1) Lorsque l'intervalle I est compact, les polynômes Ψ_n forment une base orthogonale de l'espace préhilbertien E. Les polynômes de Laguerre et d'Hermite forment une base orthogonale (bien que I ne soit pas compact). (2) Pour toute valeur de n, $\Psi_n - x\Psi_{n-1}$ est un polynôme de degré strictement inférieur à n et

$$\langle x\Psi_{n-1}, \Psi_n \rangle = \langle \Psi_n, \Psi_n \rangle$$

(3) Tout polynôme Ψ_n a ses n racines réelles, distinctes et intérieures à I. (4) Il existe deux suites de nombres réels λ_n et μ_n avec $\mu_n > 0$ telles que

$$\forall n \geq 1, \qquad \Psi_{n+1} = (x + \lambda_n)\Psi_n - \mu_n \Psi_{n-1}$$

Les constantes sont données par

$$\lambda_n = \frac{\langle x\Psi_n, \Psi_n \rangle}{\|\Psi_n\|_2^2}, \qquad \mu_n = \frac{\|\Psi_n\|_2^2}{\|\Psi_{n-1}\|_2^2}$$

Dans le cas où le polynôme n'est pas normé, on a une relation de la forme

$$\Psi_{n+1} = (a_n x + b_n)\Psi_n - c_n \Psi_{n-1}$$

avec

$$\begin{aligned} \Psi_{n+1} &= u_n x^n + u'_n x^{n-1} + \cdots \\ v_n &= \int_I \Psi_n^2(x)\omega(x)dx \\ a_n &= \frac{u_{n+1}}{u_n}, \quad b_n = a_n\left(\frac{u'_{n+1}}{u_{n+1}} - \frac{u'_n}{u_n}\right), \quad c_n = \frac{u_{n+1}u_{n-1}v_n}{u_n^2 v_{n-1}} \end{aligned}$$

2.9 Approximation quadratique

L'approximation en moyenne quadratique, encore appelée dans le cas discret *approximation des moindres carrés*, a été étudiée au XIX^e siècle par Tchebychev et Hermite. Le polynôme de meilleure approximation s'exprime simplement à l'aide des polynômes orthogonaux. Soit f une fonction de $L^2[a,b]$, on dit que le polynôme p_n de degré n est une meilleure approximation quadratique de f ou une meilleure approximation au sens des moindres carrés si la norme

$$\|f - p_n\|_2^2 = \int_a^b (f(x) - p_n(x))^2 d\mu(x)$$

est minimale. On établit facilement le résultat suivant :

Soit $f \in L^2[a,b]$ une fonction de carré intégrable et p_n une suite de polynômes orthogonaux. Pour un entier naturel n donné, la quantité

$$\left\| f - \sum a_i p_i \right\|_2$$

est minimale si et seulement si $a_i = \dfrac{\langle f, p_i\rangle}{\|p_i\|_2^2}$. Autrement dit, la meilleure approximation quadratique de f sur $[a,b]$ est donnée par

$$q_n(x) = \sum_{i=0}^{n} a_i \frac{p_i(x)}{\|p_i\|_2} \qquad \text{et } a_i = < f, \frac{p_i}{\|p_i\|_2} > = \int_a^b f(x)\frac{p_i(x)}{\|p_i\|_2} d\mu(x)$$

Car l'expression

$$\begin{aligned} \left\| f - \sum a_i p_i \right\|_2^2 &= \|f\|_2^2 + \sum_{i=0}^{n} a_i^2 - 2\sum_{i=0}^{n} a_i \langle f, p_i\rangle \\ &= \left(\|f\|_2^2 - \sum_{i=0}^{n} \langle f, p_i\rangle^2 \right) + \|p_i\|_2^2 \sum_{i=0}^{n} \left(a_i - \frac{\langle f, p_i\rangle}{\|p_i\|_2} \right)^2 \end{aligned}$$

est minimale si

$$a_i = \frac{\langle f, p_i \rangle}{\|p_i\|_2^2}$$

Lissage par les moindres carrés. Considérons un ensemble de points expérimentaux (x_i, y_i). On se propose de déterminer une droite $g(x) = ax + b$ approchant au mieux la fonction f représentée par le nuage de points (x_i, y_i). Cherchons à minimiser la quantité

$$L = \sum_{i=0}^{n} (y_i - ax_i - b)^2$$

Notons $\overline{x}$ la moyenne empirique de l'échantillon définie par

$$\overline{x} = \frac{1}{n} \sum_{i=0}^{n} x_i$$

et

$$\overline{x^2} = \frac{1}{n} \sum_{i=0}^{n} x_i^2 \qquad \text{et} \qquad \overline{xy} = \frac{1}{n} \sum_{i=0}^{n} x_i y_i$$

Les dérivées partielles

$$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial b} = 0$$

conduisent à l'équation matricielle

$$\begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix}$$

d'où on tire

$$a = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - (\overline{x})^2} \qquad \text{et} \qquad b = \frac{\overline{y}\ \overline{x^2} - \overline{x}\ \overline{xy}}{\overline{x^2} - (\overline{x})^2}$$

Notons σ_x^2 la variance des abscisses et σ_y^2 la variance des ordonnées :

$$\sigma_x^2 = \overline{x^2} - (\overline{x})^2 \qquad \text{et} \qquad \sigma_y^2 = \overline{y^2} - (\overline{y})^2$$

et c le coefficient de corrélation

$$c = \overline{xy} - \overline{x}\ \overline{y}$$

La droite $g(x) = \dfrac{c}{\sigma_x}(x - \overline{x}) - \overline{y}$ s'appelle *droite de régression* de y par rapport à x. Elle passe par le point moyen $(\overline{x}, \overline{y})$.

2.10 Polynômes de Bernstein

Les *polynômes de Bernstein* b_n^k sont les polynômes de degré n définis sur $[0,1]$ par les relations

$$b_n^k(x) = C_n^k\,(1-x)^{n-k}x^k \qquad 0 \le k \le n$$

La notation C_n^k désigne le nombre de combinaisons de k objets parmi n $(= n!/k!(n-k)!)$.

Propriétés. (1) Les polynômes de Bernstein sont positifs

$$\forall x \in [0,1] \qquad b_n^k(x) \ge 0$$

(2) Les polynômes de Bernstein forment une base de l'espace des polynômes de degré inférieur ou égal à n.

(3) La somme des polynômes de même degré vaut 1

$$\sum_{k=0}^{n} b_n^k(x) = 1$$

(4) Les polynômes de Bernstein vérifient la relation de symétrie

$$b_n^k(x) = b_n^{n-k}(1-x)$$

(5) Les polynômes de Bernstein vérifient la relation de récurrence

$$b_n^k(x) = x b_{n-1}^k(x) + (1-x) b_{n-1}^{k+1}(x)$$

(6) La dérivée des polynômes de Bernstein vérifie

$$\frac{db_n^k(x)}{dx} = n(b_{n-1}^{k-1}(x) - b_{n-1}^k(x))$$

Exemple. La base de Bernstein des polynômes de degré 3 est formée des fonctions $b_3^0(x) = (1-x)^3$ $b_3^1(x) = 3(1-x^2)x$ $b_3^2(x) = 3x^2(1-x)$ et $b_3^3(x) = x^3$.

Les polynômes de Bernstein permettent de démontrer facilement le théorème de Weierstrass qui affirme que toute fonction continue sur un intervalle $[a,b]$ est limite uniforme d'une suite de fonctions polynomiales. En effet, par un changement de variable, on se ramène à l'intervalle $[0,1]$. Si f est une fonction continue, nous allons montrer qu'on peut trouver une suite de polynômes de degré n qui converge uniformément vers f. Choisissons le polynôme

$$B_n(x) = \sum_{k=0}^{n} C_n^k x^k (1-x)^{n-k} f(\frac{k}{n}) = \sum_{k=0}^{n} b_n^k(x) f(\frac{k}{n})$$

où les b_n^k sont les polynômes de Bernstein. En développant suivant la formule du binôme, on obtient

$$\sum_{k=0}^{n} C_n^k x^k (1-x)^{n-k} = 0$$

et en dérivant deux fois par rapport à x

$$\sum_{k=0}^{n} \frac{k}{n} C_n^k x^k (1-x)^{n-k} = x$$

$$\sum_{k=0}^{n} \left(\frac{k}{n}\right)^2 b_n^k(x) = (1-\frac{1}{n})x^2 + \frac{x}{n}$$

En additionnant ces trois identités, on obtient

$$\sum_{k=0}^{n} \left(\frac{k}{n} - x\right)^2 b_n^k(x) = \frac{1}{n} x(1-x)$$

D'autre part, comme f est une fonction continue sur $[0,1]$, elle est donc uniformément continue et bornée

$$\forall \epsilon > 0, \quad \exists \delta > 0, \quad \forall x, y \in [0,1], \quad |x-y| < \delta \Longrightarrow |f(x) - f(y)| < \epsilon$$

Pour chaque x, notons I l'ensemble des indices k pour lesquels $\left|\frac{k}{n} - x\right| \geq \delta$, on a par continuité uniforme

$$\begin{aligned} |f(x) - B_n(x)| &= \left| \sum_{k=0}^{n} \left(f(x) - f(\frac{k}{n}) \right) b_n^k(x) \right| \\ &\leq \epsilon + \left| \sum_{k \in I} \left(f(x) - f(\frac{k}{n}) \right) b_n^k(x) \right| \end{aligned}$$

La fonction f étant bornée par une constante M, on a

$$\begin{aligned} \left| \sum_{k \in I} \left(f(x) - f(\frac{k}{n}) \right) b_n^k(x) \right| &< 2M \sum_{k \in I} \frac{(\frac{k}{n} - x)^2}{(\frac{k}{n} - x)^2} b_n^k(x) \\ &< \frac{2M}{\delta^2} \frac{x(1-x)}{n} < \frac{M}{2n\delta^2} \end{aligned}$$

Finalement, si $\epsilon > M/2n\delta^2$, on a

$$\forall x \in [0,1], \quad |f(x) - B_n(x)| < \epsilon$$

ce qui démontre le théorème de Weierstrass.

2.11 Fonctions splines

En introduisant les fonctions splines dans les années 40, Schoenberg apporta plus de souplesse dans l'approximation polynomiale. Il permit de diminuer le degré du polynôme approchant la fonction en considérant des fonctions polynomiales par morceaux appelées *splines*, du nom de la tige flexible qu'on fixait sur le papier pour tracer des courbes lisses. Les polynômes de Serge Bernstein (1880-1968) servent dans la construction paramétrique des *B-splines*. Nous envisageons ici le cas des *splines cubiques*. Soit $(x_0, x_1, \ldots, x_n)$ les points d'interpolation d'une fonction f sur un intervalle $[a, b]$. On note $f_i = f(x_i)$. Sur chaque intervalle élémentaire $[x_i, x_{i+1}[$, on cherche un polynôme s_i vérifiant les conditions d'interpolation

$$s_i(x_i) = f_i$$

et les conditions de continuité des dérivées premières et secondes, pour $i = 1, 2, ..., n-1$

$$\begin{cases} s'_{i-1}(x_i) = s'_i(x_i) \\ s''_{i-1}(x_i) = s''_i(x_i) \end{cases}$$

L'ensemble de ces conditions nous incite à chercher un polynôme du troisième degré. Sa dérivée seconde vérifie, en posant $h_i = x_{i+1} - x_i$

$$s''_i(x) = f''_i \frac{x_{i+1} - x}{h_i} - f''_{i+1} \frac{x_i - x}{h_i}$$

La fonction $s_i(x)$ est donc de la forme

$$s_i(x) = f''_i \frac{(x_{i+1} - x)^3}{6h_i} - f''_{i+1} \frac{(x_i - x)^3}{6h_i} + a_i(x_{i+1} - x) - b_i(x_i - x)$$

où les constantes a_i et b_i sont déterminées par les conditions d'interpolation. Comme $s_i(x_i) = f_i$, on en déduit que

$$s_i(x_i) = \frac{f_i}{h_i} - f''_i \frac{h_i}{6}$$

et de la condition $s_i(x_{i+1}) = f_i$, on déduit la valeur

$$s_i(x_i) = \frac{f_{i+1}}{h_i} - f''_{i+1} \frac{h_i}{6}$$

En exprimant la condition de continuité des dérivées premières, on obtient un système de $(n-1)$ équations données par

$$h_i f''_i + 1 + 2(h_i + h_{i-1}) f''_i + h_{i-1} f''_{i-1} = 6\left(\frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_{i-1}}\right)$$

qui détermine complètement la fonction spline. Dans le cas général, la fonction spline est un polynôme par morceaux de degré inférieur ou égal à k.

Sur l'intervalle $[a, b]$, la fonction cherchée $S(t)$ est de classe C^{k-1}. On impose à la fonction spline $(p+1)$ conditions d'interpolation pour $i = 1, 2, ..., p$

$$S_i(x_i) = f_i$$

Les autres conditions sont déterminées par la continuité des dérivées et la condition des splines cubiques

$$S'(a) = f'(a) \quad et \quad S'(b) = f'(b)$$

La fonction polynomiale S est alors déterminée de manière unique. Si la fonction f est de classe C^2,

$$\|f - S\|_\infty = O(1/p^2)$$

et

$$\|f - S'\|_\infty = O(1/p)$$

En général, les splines sont déterminées par leur expression paramétrique. Soit $(t_0, t_1, \ldots, t_m)$ une suite croissante de réels. On appelle *B-splines* de degré k, les courbes définies pour $t \in \mathbb{R}$ et $0 \leq i \leq m-k-1$ par

$$B_{i,0}(t) = 1 \text{ si } t_i \leq t < t_{i+1} \text{ et } 0 \text{ sinon}$$

et si $t_i < t_{i+k}$ et $t_{i+1} < t_{i+k+1}$

$$B_{i,k}(t) = \frac{t - t_i}{t_{i+k} - t_i} B_{i,k-1}(t) + \frac{t_{i+k+1} - t}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(t)$$

Les réels t_i sont appelés *nœuds*. S'il y a r nœuds t_i égaux entre eux, on dit que ce point est un nœud d'ordre r. On pose par convention $0/0 = 0$. Les fonctions splines sont définies paramétriquement. Le paramètre t représente le temps. Les polynômes de Bernstein sont des cas particuliers de *B-splines* à condition de prendre comme nœuds

$$\begin{aligned} t_0 &= \ldots = t_k = 0 \\ t_{k+1} &= \ldots = t_{2k+1} = 1 \end{aligned}$$

Propriétés. (1) $B_{i,k}(t)$ est un polynôme de degré k par morceaux.
(2) $B_{i,k}(t) = 0$ pour $t \notin [t_i, t_{i+k+1}]$
(3) Soit $[a, b]$ un intervalle tel que $t_k \leq a$ et $t_{m-k} \geq b$ alors

$$\sum_{i=0}^{m-k-1} B_{i,k}(t) = 1 \qquad \forall t \in [a, b[$$

(4) Soit $t \in]t_i, t_{i+k+1}[$ alors $B_{i,k}(t) = 1$ si et seulement si $t_{i+1} = \ldots = t_{i+k} = t$

(5) $B_{i,k}(t)$ est continue et indéfiniment dérivable à droite sur $\mathbb{R}$ et sa dérivée vaut

$$B'_{i,k}(t) = k[\frac{B_{i,k-1}(t)}{t_{i+k} - t_i} - \frac{B_{i+1,k-1}(t)}{t_{i+k+1} - t_{i+1}}]$$

avec la convention suivante : on remplace par 0 les expressions dont le dénominateur est nul.
(6) L'intégrale d'une B-spline

$$\int_{-\infty}^{+\infty} B_{i,k}(t)\ dt = \frac{1}{k+1}(t_{i+k+1} - t_i)$$

Soit $(P_0, P_1, \ldots, P_{n-1})$ n points de $\mathbb{R}^s$. On appelle *fonction spline* (ou courbe spline) de degré k associé au polygone $(P_0, P_1, \ldots, P_{n-1})$ la courbe définie par l'expression paramétrique

$$S(t) = \sum_{i=0}^{n-1} P_i B_{i,k}(t)$$

pour t dans l'intervalle $[a, b]$ et aux nœuds $(t_0, t_1, \ldots, t_{n+k})$. Si les fonctions $B_{i,k}(t)$ sont les polynômes de Bernstein, la courbe est appelée *courbe de Bézier*. En général, la courbe spline ne passe pas par les points P_i. Dans le cas où les nœuds t_i sont simples $(k+1 \leq i \leq n+1)$, la courbe spline est de classe C^{k-1} et est formée de n arcs paramétrés polynomiaux de degré $\leq k$. Le choix des nœuds permet de définir facilement des splines

$$\begin{cases} t_0 = t_1 = \ldots = t_k = 0 \\ t_{k+1} = 1 \\ \vdots \\ t_{n-1} = n - k - 1 \\ t_n = t_{n+1} = \ldots = t_{n+k} = n - k \end{cases}$$

La répétition des nœuds garantit que les extrémités de la courbe coïncident avec les extrémités de la ligne polygonale : $S(a) = P_0$ et $S(b) = P_{n-1}$. Pour construire une courbe fermée, il suffit de choisir des nœuds cycliques $[0, 1, ..., n, 0, 1, ...n$, etc.] par exemple

$$\begin{cases} t_0 = t_{k+1} = 0 \\ t_{k+2} = t_1 = 1 \\ t_{k+i+1} = t_i \end{cases}$$

Exemple. Soit quatre points $(P_0, P_1, \ldots, P_4)$. La courbe de Bézier sur ces quatre points aura pour équation

$$S(t) = P_0(1-t)^3 + 3t(1-t)^2 P_1 + 3t^2(1-t)P_2 + t^3 P_3$$

soit encore

$$S(t) = (P_3 - 3P_2 + 3P_1 - P_0)t^3 + 3(P_2 - 2P_1 + P_0)t^2 + 3(P_1 - P_0)t + P_0$$

soit matriciellement

$$S(t) = (1, t, t^2, t^3) \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3 & 3 & 0 & 0 \\ 3 & -6 & 3 & 0 \\ -1 & 3 & -3 & 1 \end{pmatrix} \begin{pmatrix} P_0 \\ P_1 \\ P_2 \\ P_3 \end{pmatrix}$$

Si les points sont des points du plan de coordonnées (x_i, y_i), ce système équivaut à

$$\begin{cases} x(t) = (x_3 - 3x_2 + 3x_1 - x_0)t^3 + 3(x_2 - 2x_1 + x_0)t^2 + 3(x_1 - x_0)t + x_0 \\ y(t) = (y_3 - 3y_2 + 3y_1 - y_0)t^3 + 3(y_2 - 2y_1 + y_0)t^2 + 3(y_1 - y_0)t + y_0 \end{cases}$$

2.12 Approximants de Padé

L'*approximation de Padé* est une approximation locale qui consiste à prendre au voisinage d'un point donné une approximation sous forme de fraction polynomiale $w(x) = P(x)/Q(x)$. En particulier, lorsqu'une fonction f admet un développement en fractions continues, la suite des fractions tronquées à l'ordre n forme une suite f_n qui converge vers f uniformément sur tout compact. Par exemple, la fonction exponentielle :

$$e^x = 1 + \cfrac{x}{1 + \cfrac{x}{-2 + \cfrac{x}{-3 + \cfrac{x}{2 + \cfrac{x}{5 + \cfrac{x}{-2 + \cfrac{x}{-7 + \cdots}}}}}}}$$

La troncature de l'exponentielle conduit à des approximations fractionnaires qui convergent très rapidement. De même, la fonction *cosinus* peut être approchée au voisinage de 0 par l'expression

$$\cos(x) \simeq 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4$$

avec un dénominateur du second degré

$$\cos(x) \simeq \frac{1 - \frac{7}{15}x^2 + \frac{1}{40}x^4}{1 + \frac{1}{30}x^2}$$

ou comme le rapport de deux polynômes du quatrième degré

$$\cos(x) \simeq \frac{1 - \frac{115}{252}x^2 + \frac{313}{15120}x^4}{1 + \frac{11}{252}x^2 + \frac{13}{15120}x^4}$$

2.13 Exercices

1. On considère la fonction
$$f(x) = \frac{4}{1-x}$$
Écrire le polynôme de Lagrange $p(x)$ aux points $x_0 = -1$, $x_1 = 0$, $x_2 = 2$ et $x_3 = 3$. Évaluer l'erreur au point $x = 1 + \sqrt{5}$. Tracer les courbes $f(x)$ et $p(x)$.
2. Écrire pour la fonction de l'exercice précédent, le polynôme d'Hermite vérifiant
$$\begin{aligned} p(0) &= f(0) = 4 \\ p(2) &= f(2) = -4 \\ p'(0) &= f'(0) = 4 \\ p'(2) &= f'(2) = 4 \end{aligned}$$
Tracer les courbes $f(x)$ et $p(x)$. Évaluer l'erreur au point $x = 1 + \sqrt{5}$.
3. Soit la fonction
$$f(x) = 3\exp(x^2 - \frac{3}{4})$$
Écrire l'interpolation de Tchebychev sur trois points.
4. On considère la fonction
$$f(x) = \frac{1}{1+25x^2}$$
sur l'intervalle $[-1, +1]$. Soit $x_j = 1 + jh$, pour $j = 0, 1, 2, ..., n$ une subdivision régulière de pas $h = 2/n$. Déterminer la fonction spline cubique lorsque $n = 4$, et $n = 12$.
5. On définit les polynômes de Tchebychev par la relation de récurrence
$$\begin{cases} T_0(x) = 1 \quad T_1(x) = x \\ T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \end{cases}$$
Montrer que ces polynômes satisfont la relation de récurrence
$$(1-x^2)T_n(x) = -nxT_n(x) + nT_{n-1}(x)$$
Démontrer les relations d'orthogonalité
$$\int_{-1}^{+1} T_n(x)T_m(x)\frac{dx}{\sqrt{1-x^2}} = \frac{\pi}{2}\delta_{n,m} \qquad n \neq 0$$
et pour $n = 0$
$$\int_{-1}^{+1} T_0^2(x)\frac{dx}{\sqrt{1-x^2}} = \pi$$

3
Résolution d'équations

Le problème de la résolution d'équations algébriques ou transcendantes est un problème difficile qui fait intervenir des notions essentielles bien qu'il puisse être posé en des termes simples. La résolution des équations algébriques par radicaux s'ouvre sur la théorie de Galois ; de nombreuses propriétés topologiques (théorèmes de point fixe, indices de fonction) sont à la base des principaux résultats obtenus dans ce domaine. Des structures analogues se retrouvent dans la théorie des équations différentielles. Le problème de la résolution d'équations était connu dès l'Antiquité. Les mathématiciens cherchaient à résoudre par *approximations successives* le problème numérique de l'extraction de racines. Au I^{er} siècle de notre ère, Héron d'Alexandrie proposa un algorithme pour approcher une racine carrée. Au Moyen Âge, al-Tusi étudia les équations cubiques. Au XV^{e} siècle, al-Kashi calcula la valeur approchée de $sin(1^0)$ à partir de $sin(3^0)$ par résolution d'une équation cubique. Vers 1600, François Viète (1504-1603) donne des solutions d'équations algébriques du sixième degré. Paolo Ruffini (1765-1822), François Budan (1761-1840) et William Horner (1786-1837) ont proposé des solutions approchées par transformations des équations polynomiales et approximations successives.

3.1 Équations algébriques

La théorie des équations algébriques repose sur le théorème fondamental de l'algèbre qui assure l'existence de solutions. Ce théorème encore appelé

théorème de d'Alembert affirme que l'équation algébrique $P(x) = 0$ où P est un polynôme de degré n admet exactement n racines distinctes ou non, réelles ou complexes, et dans le cas complexe, deux à deux conjuguées. Depuis l'Antiquité, l'homme a cherché des formules explicites donnant les valeurs des racines en fonction des coefficients du polynôme $P(x)$ à l'image de l'équation du second degré $x^2 - px + q = 0$ qui admet si $p^2 - 4q > 0$ deux racines réelles x_1 et x_2 vérifiant $p = x_1 + x_2$ et $q = x_1 x_2$. Au XVI^e siècle, Niccolo Tartaglia, Scipione del Ferro et Antonio Fior ont cherché à résoudre des équations cubiques, mais c'est Girolamo Cardano (Cardan) qui donna en 1545 dans son *Ars magna* les formules de résolution. L'équation générale du troisième degré exprimée sous forme alternée

$$x^3 - ax^2 + bx - c = 0$$

admet dans certains cas trois racines x_1, x_2, x_3 qui vérifient $a = x_1 + x_2 + x_3$, $b = x_1x_2 + x_2x_3 + x_1x_3$ et $c = x_1x_2x_3$. On démontre que cette équation se ramène par un changement de variable ($X = x - a/3$) à une équation du type

$$x^3 + px + q = 0$$

Elle admet si $\Delta = 4p^3 + 27q^2$ est positif, trois racines distinctes données par les *formules de Cardan*

$$x_1 = \frac{1}{3}(u + v) \qquad x_2 = \frac{1}{3}(ju + j^2 v) \quad et \quad x_3 = \frac{1}{3}(j^2 u + jv)$$

avec $j^3 = 1$ et u et v étant données par

$$u = \sqrt[3]{\frac{-27}{2}q + \frac{3}{2}\sqrt{3}A} \quad v = \sqrt[3]{\frac{-27}{2}q - \frac{3}{2}\sqrt{3}A}$$

Si $\Delta > 0$, alors $A = \sqrt{\Delta}$ u est différent de v. Les racines x_2 et x_3 sont imaginaires conjuguées. x_1 est la seule racine réelle qui s'exprime à l'aide de radicaux réels. En revanche, si $\Delta < 0$, alors $A = i\sqrt{-\Delta}, u = \overline{v}$. Les trois racines sont réelles, mais l'expression qui est sous la racine cubique est complexe. Dans ce cas, il est impossible d'exprimer les racines de l'équation sous forme de radicaux réels.

L'équation du quatrième degré a été résolue par Luigi Ferrari, un disciple de Cardan. Exprimée sous forme alternée, l'équation

$$x^4 - ax^3 + bx^2 - cx + d = 0$$

admet dans certains cas quatre racines x_1, x_2, x_3, x_4 qui vérifient $a = x_1 + x_2 + x_3 + x_4$, $b = x_1x_2 + x_2x_3 + x_1x_3 + x_1x_4 + x_2x_4 + x_3x_4$, $c = x_1x_2x_3 + x_1x_2x_4 + x_1x_3x_4 + x_2x_3x_4$ et $d = x_1x_2x_3x_4$. Par changement de variable, on vérifie que l'équation se ramène à la forme

$$x^4 + px^2 + qx + r = 0$$

Cette équation se résout comme une équation du second degré lorsque $q = 0$. Dans le cas contraire, Ferrari propose de l'exprimer sous la forme

$$(x^2 + \frac{1}{2}y)^2 = (y - p)x^2 - qx + \frac{1}{4}y^2 - r$$

et de déterminer y de telle sorte que le deuxième membre soit un carré du type $(mx + n)^2$. Pour cela on démontre qu'il faut et il suffit que le discriminant $q^2 - 4(\frac{1}{4}y^2 - r)$ soit nul, autrement dit que y vérifie l'équation du troisième degré, appelée résolvante

$$y^3 - py^2 - 4ry + 4pr - q^2 = 0$$

Ainsi, la résolution d'une équation du quatrième degré se ramène à la résolution d'une équation du troisième degré.

L'équation générale de degré supérieur ou égal à 5 n'est pas résoluble par radicaux. Ce résultat a été démontré par N.H. Abel en 1824. Il se déduit des travaux d'Évariste Galois qui a attaché à chaque équation un groupe sur lequel on lit directement les propriétés de l'équation. Il traduit le fait que le groupe symétrique a une structure plus pauvre lorsque $n \geq 5$, car le groupe alterné n'a pas de sous-groupe distingué propre.

3.2 Théorèmes de points fixes

Soit f une application d'un ensemble E dans lui-même. On appelle *point fixe* d'une application f tout élément $u \in E$ tel que $f(u) = u$. On voit que résoudre ce type d'équation est un cas particulier des équations numériques $h(u) = f(u) - u = 0$. De nombreux résultats affirment l'existence de points fixes. Lorsque l'ensemble E est un espace de Banach (c'est-à-dire un espace vectoriel normé complet), toute application contractante (i.e. toute application lipschitzienne de rapport $k < 1$) de E dans lui-même admet un et un seul point fixe u tel que

$$\forall x \in E, \qquad u = \lim_n f^n(x)$$

En effet, soit x_0 un point de E et x_n la suite définie par $x_{n+1} = f(x_n)$. Cette suite est une suite de Cauchy, car f est contractante. Comme $d(x_n, x_{n+1}) \leq kd(x_{n-1}, x_n)$, on a $d(x_n, x_{n+1}) \leq k^n d(x_0, x_1)$, et par suite, l'inégalité

$$d(x_n, x_{n+p}) \leq \sum_{i=n}^{n+p-1} d(x_i, x_{i+1}) \leq d(x_0, x_1) \sum_{i=n}^{\infty} k^i = \frac{k^n}{1-k} d(x_0, x_1)$$

Comme $0 \leq k < 1$, la suite x_n est bien une suite de Cauchy. L'espace E étant complet, la suite x_n converge. Soit u sa limite. La fonction f étant

continue, on déduit de $x_{n+1} = f(x_n)$ que u est un point fixe $u = f(u)$. L'unicité découle de la propriété de f. Supposons que v soit un deuxième point fixe de f, on aurait $d(u, v) = d(f(u), f(v)) \leq kd(u, v)$ d'où $d(u, v) = 0$ et donc $u = v$. Par exemple, la fonction $f(x) = ax + b$, avec $|a| < 1$ conduit à

$$f^n(x) = a^n x + b\frac{1 - a^n}{1 - a}$$

Le point fixe de f est donné par $u = \lim f^n(x) = \dfrac{b}{1 - a}$

Il existe beaucoup de théorèmes de points fixes, qui se fondent sur des propriétés topologiques. Le *théorème de Brouwer* affirme que toute application continue f du disque

$$D^2 = \{(x, y) \in \mathbb{R} \times \mathbb{R} : x^2 + y^2 \leq 1\}$$

sur lui-même admet au moins un point fixe. Car si f n'a pas de point fixe, on démontre qu'alors le cercle S^1 serait contractile, c'est-à-dire homotope à un point, ce qui est faux. Le *théorème de Tychonov*, qui généralise un résultat de Schauder, affirme que si E est un espace séparable localement convexe et f une fonction définie sur un sous-ensemble compact convexe A de E et à valeurs dans A, alors f admet dans A au moins un point fixe.

3.3 Localisation des racines

En pratique, la mise en œuvre d'un algorithme de recherche de solution d'équations suppose que nous connaissons une région dans laquelle se trouve cette solution. La théorie donne quelques critères de localisation lorsque l'équation est une équation polynomiale. Le *théorème de Rolle* (1690) affirme qu'entre deux racines de l'équation $P(x) = 0$ où P est un polynôme, il existe au moins une racine de l'équation dérivée $P'(x) = 0$. La règle de Descartes affirme que le nombre de racines positives d'un polynôme

$$P(x) = a_0 + a_1 x + ... + a_n x^n$$

est inférieur au nombre de changements de signes de la suite $(a_0, a_1, ...a_n)$. Le *théorème de Sturm* (1829) donne un algorithme pour déterminer le nombre de racines d'un polynôme entre deux réels. Soit a et b deux nombres réels $a < b$ et P un polynôme de degré n n'ayant que des racines simples. On note $P_0 = P$, $P_1 = P'$, $P_2 = Q_2 P_1 - P_0$ l'opposé du reste de la division euclidienne de P_0 par P_1, ..., et $P_{i+2} = Q_{i+2} P_{i+1} - P_i$ l'opposé du reste de la division euclidienne de P_i par P_{i+1}. On considère $P_i(a)$ la suite $P_0(a), P_1(a), ..., P_n(a)$ et $P_i(b)$ la suite $P_0(b), P_1(b), ..., P_n(b)$ et on suppose que $P_0(a) \neq 0, P_1(a) \neq 0$. Le nombre de racines réelles de $P(x)$ comprises entre a et b est égal au nombre de changements de signes que présente la

première suite diminué de celui que présente la deuxième suite. Ce nombre est égal à l'indice de P'/P entre a et b

$$-\pi I(\frac{P'}{P}, a, b) = \int_a^b \frac{P''P - P'^2}{P'^2 + P^2} dx - Arctg\left(\frac{P'(b)}{P(b)}\right) + Arctg\left(\frac{P'(a)}{P(a)}\right)$$

Exemple. Soit $P(x) = x^3 - x$ un polynôme de degré 3. P admet trois racines distinctes ($x = -1, 0, 1$). Cherchons le nombre de racines de P dans l'intervalle $[-2, 2]$. La suite de Sturm s'écrit $P_0(x) = x^3 - x$, $P_1(x) = 3x^2 - 1$, $P_2(x) = \frac{2}{3}x$ et $P_3(x) = 1$. Les différentes valeurs de $P_i(x)$ aux points $a = -2$ et $b = 2$ sont résumées dans le tableau suivant :

x	P_0	P_1	P_2	P_3
-2	-6	11	$-4/3$	1
2	6	11	4/3	1

La suite $P_i(-2)$ change trois fois de signe et la suite $P_i(2)$ reste constante. L'indice est donc égal à $3 - 0 = 3$. Le polynôme $P(x)$ admet donc trois racines réelles dans l'intervalle $[-2, 2]$. Ce calcul peut aussi s'effectuer en intégrant l'expression ci-dessus, soit

$$-\pi I(\frac{P'}{P}, a, b) = u(b) - u(a) - Arctg\left(\frac{3b^2 - 1}{b^3 - b}\right) + Arctg\left(\frac{3a^2 - 1}{a^3 - a}\right)$$

avec

$$u(t) = Arctg(11t^3 + \frac{3}{2}t^5 - \frac{7}{2}t) - Arctg(\frac{25}{6}t + \frac{1}{2}t^3) - Arctg(\frac{1}{3}t)$$

En remplaçant a et b par leurs valeurs, on retrouve la valeur précédente $I = 3$.

Notons enfin le *critère de Routh-Hurwitz* qui établit une condition nécessaire et suffisante pour qu'un polynôme $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ à coefficients réels ait toutes ses racines à parties réelles négatives. Pour cela il faut et il suffit que tous les mineurs diagonaux de la matrice de Hurwitz

$$A = \begin{pmatrix} a_1 & a_0 & 0 & \cdots & \cdots & \cdots & 0 \\ a_3 & a_2 & a_1 & a_0 & 0 & \cdots & 0 \\ a_5 & a_4 & a_3 & a_2 & a_1 & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

soient positifs, c'est-à-dire que les quantités

$$\Delta_1 = a_1 \qquad \Delta_2 = \begin{vmatrix} a_1 & a_0 \\ a_3 & a_2 \end{vmatrix} \qquad \Delta_3 = \begin{vmatrix} a_1 & a_0 & 0 \\ a_3 & a_2 & a_1 \\ a_5 & a_4 & a_3 \end{vmatrix} \qquad \Delta_n = \det(A)$$

soient positives. Comme $\Delta_n = a_n.\Delta_{n-1}$, on peut remplacer la condition $\Delta_n > 0$ par $a_n > 0$.

3.4 Approximations successives

Dans les méthodes d'approximations successives, l'équation $f(x) = 0$ est remplacée par l'étude d'une suite numérique convergente

$$x_{n+1} = \varphi(x_n)$$

qui permet d'obtenir en un nombre fini d'itérations une solution approchée de l'équation. En général, on prend $\varphi(x) = x - cf(x)$. Dans la méthode de Lagrange, on remplace la fonction f par le segment de droite passant par les points $(a, f(a))$ et $(b, f(b))$

$$\varphi(x) = a - f(a)\frac{x - a}{f(x) - f(a)}$$

Dans la méthode de Newton, on remplace la fonction f entre les points d'abscisse a et b par la tangente à la courbe en ces points

$$\varphi(x) = x - \frac{f(x)}{f'(x)}$$

3.5 Méthode de la sécante

La *méthode de la sécante*, encore appelée *méthode de la fausse position* ou "*regula falsi*", a été employée au XVI^e siècle par Viète (1540-1603) puis, plus tard par Descartes (1596-1650). C'est une méthode par approximations successives, fondée sur la formule itérative suivante

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$$

Elle correspond à la méthode de Newton dans laquelle la dérivée $f'(x_n)$ est remplacée par le taux d'accroissement selon l'approximation

$$f'(x_n) \simeq \frac{f(x_n) - f(x_{n-1})}{x_n) - x_{n-1}}$$

On arrête l'itération lorsque la différence entre deux pas successifs devient inférieure à une certaine valeur ε. La méthode est d'ordre $(1+\sqrt{5})/2$. (Voir exercices.)

3.6 Méthode de Müller

Connaissant la fonction f en trois points $(x_{n-2},\, x_{n-1},\, x_n)$, on approche $f(x)$ par un polynôme $P(x)$ de Lagrange de degré 2. En résolvant l'équation $P(x) = 0$, on obtient une approximation de la racine de $f(x)$ qui est notée x_{n+1}. On itère l'opération en prenant comme triplet $(x_{n-1},\, x_n,\, x_{n+1})$. On a

$$\begin{aligned} P(x) &= \frac{(x-x_{n-1})(x-x_{n-2})}{(x_n-x_{n-1})(x_n-x_{n-2})} f(x_n) \\ &+ \frac{(x-x_n)(x-x_{n-2})}{(x_{n-1}-x_n)(x_{n-1}-x_{n-2})} f(x_{n-1}) \\ &+ \frac{(x-x_n)(x-x_{n-1})}{(x_{n-2}-x_n)(x_{n-2}-x_{n-1})} f(x_{n-2}) \end{aligned}$$

Ce polynôme est de la forme $a_n x^2 + b_n x + c_n$. On résout $P(x) = 0$ en calculant le discriminant et en prenant la racine x_{n+1} la plus proche de x_n.

3.7 Méthode de la bissection

Dans la *méthode de dichotomie* ou *méthode de la bissection*, l'intervalle de recherche de la solution est coupé en deux à chaque pas d'itération. On détermine progressivement un intervalle de plus en plus fin dans lequel se trouve la solution cherchée. Soit f une fonction numérique strictement monotone sur un intervalle $[a, b]$. On suppose que l'équation $f(x) = 0$ n'a qu'une et une seule solution dans cet intervalle. On se propose de déterminer cette valeur u avec une précision donnée. Soit $[a_0, b_0]$ un intervalle dans lequel $f(a_0)f(b_0) < 0$. On note $c_0 = (a_0 + b_0)/2$ le centre de l'intervalle. Si $f(c_0)f(a_0) < 0$, alors la racine u appartient à l'intervalle $[a_0, c_0]$. On reprend le procédé avec $a_1 = a_0$ et $b_1 = c_0$. Sinon, c'est-à-dire si $f(c_0)f(a_0) > 0$, on pose $a_1 = c_0$ et $b_1 = b_0$. On construit ainsi une suite d'intervalles emboîtés $[a_n, b_n]$ de longueur $(a_0 + b_0)/2^n$. Les suites a_n et b_n sont adjacentes et convergent vers u.

3.8 Méthode de Newton-Raphson

La *méthode de Newton-Raphson*, encore appelée *méthode des tangentes* a été exposée par Newton vers 1669 et complétée par Joseph Raphson (1648-1715) en 1690. C'est une méthode par approximations successives fondée sur le théorème suivant :

Soit E et F deux espaces de Banach, U un ouvert de E, f une application de U dans F de classe C^1 telle que la différentielle Df de f soit lipschitzienne

sur U

$$\exists\ M > 0, \qquad \|Df(x) - Df(y)\| \leq M\,\|x - y\| \qquad \forall x, y \in U$$

et telle qu'il existe un point x_0 de U au voisinage duquel $Df(x_0)$ soit un isomorphisme local. Alors la suite (x_n) définie par la relation de récurrence

$$x_{n+1} = x_n - (Df(x_n))^{-1}\, f(x_n)$$

converge vers l'unique solution de l'équation $f(x) = 0$.

La formule de récurrence offre une formule itérative qu'on initialise à partir d'un point arbitraire suffisamment voisin de la racine que l'on cherche à déterminer. Lorsque f est une fonction à variable réelle, la formule donnant x_{n+1} est l'intersection de la tangente passant par le point $(x_n, f(x_n))$ avec l'axe des abscisses

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

Lorsque la racine est double, on se ramène à une racine simple en remplaçant la fonction f par la fonction $f(x)/f'(x)$. On arrête l'itération lorsque la différence entre deux pas consécutifs est inférieure à la précision souhaitée $|x_{n+1} - x_n| < \varepsilon$. Cette méthode qui converge très rapidement sert aussi à la résolution de systèmes non linéaires.

Exemple. Considérons le système

$$\left\{ \begin{array}{l} x - y^2 + xe^y = 2 \\ ye^y + x^3 = 1 \end{array} \right.$$

La fonction $f(x, y) = (x - y^2 + xe^y - 2\ ,\ ye^y + x^3 - 1)$ est de classe C^1 sur $\mathbb{R}^2$. Elle admet pour dérivée la matrice

$$f'(x, y) = \left(\begin{array}{cc} 1 + e^y & -2y + xe^y \\ 3x^2 & (y+1)e^y \end{array} \right)$$

La suite (x_{n+1}, y_{n+1}) est donc définie par :

$$\left(\begin{array}{c} x_{n+1} \\ y_{n+1} \end{array} \right) = \left(\begin{array}{c} x_n \\ y_n \end{array} \right) - \frac{1}{\Delta} A.B$$

avec

$$A = \left(\begin{array}{cc} (y_n+1)e^{y_n} & 2y_n - x_n e^{y_n} \\ -3x_n^2 & 1 + e^{y_n} \end{array} \right) \quad \text{et} \quad B = \left(\begin{array}{c} x_n - y_n^2 + x_n e^{y_n} - 2 \\ y_n e^{y_n} + x_n^3 - 1 \end{array} \right)$$

$\Delta = \det f'(x_n, y_n)$ est le déterminant de la dérivée. En partant du point $(x_0 = 2, y_0 = -2)$, on obtient successivement $(x_1 = 1.45,\ y_1 = -0.89)$, $(x_2 = 1,18,\ y_2 = -0,35)$ $(x_3 = 1,06,\ y_3 = -0,11)$ $(x_4 = 1,015,\ y_4 = -0,024)$, etc. qui converge très rapidement vers la solution exacte $x = 1$, $y = 0$. Noter qu'en partant du point $(x_0 = 1, y_0 = 1)$, le système converge vers une autre solution $x \approx 0.9297..., y \approx 0.1662...$

3.9 Méthode de Steffensen

Pour améliorer la convergence de la méthode des approximations successives, Aitken a démontré que la suite y_n donnée par l'expression

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

converge plus rapidement que la suite x_n. Steffensen a donc proposé de remplacer le calcul de x_n par celui de y_n. Ce qui revient à considérer la fonction

$$\varphi(x) = \frac{xu\,(u(x)) - u^2(x)}{u\,(u(x)) - 2u(x) + x}$$

avec $u(x) = x - f(x)$.

3.10 Méthode de Brent

La *méthode de Brent* est une amélioration d'une méthode inventée dans les années 60 par Van Wijgaarden et Dekker et qui consiste à combiner la méthode de la bissection et l'approximation quadratique. Comme dans la méthode de la bissection, on construit trois suites de points a_n, b_n et c_n et à chaque itération, on évalue l'interpolation

$$\begin{aligned} x_n \quad = \quad & \frac{(y - f(a_n))(y - f(c_n))b_n}{(f(b_n) - f(a_n))(f(b_n) - f(c_n))} + \frac{(y - f(c_n))(y - f(b_n))a_n}{(f(a_n) - f(c_n))(f(a_n) - f(b_n))} \\ & + \frac{(y - f(b_n))(y - f(a_n))c_n}{(f(c_n) - f(b_n))(f(c_n) - f(a_n))} \end{aligned}$$

Cette expression peut encore s'écrire pour $y = 0$ sous la forme

$$x_n = c_n + \frac{P_n}{Q_n}$$

Les quantités P_n et Q_n sont déterminées par les relations

$$P_n = (A_n - B_n)B_nC_n(b_n - c_n) - (1 - A_n)B_n(c_n - a_n)$$

et

$$Q_n = (A_n - 1)(B_n - 1)(C_n - 1)$$

avec

$$A_n = \frac{f(c_n)}{f(b_n)} \quad B_n = \frac{f(c_n)}{f(a_n)} \quad C_n = \frac{f(a_n)}{f(b_n)}$$

L'expression P_n/Q_n est un résidu qui diminue à chaque pas. La suite x_n converge vers la solution c_n.

3.11 Méthode de Frobenius

La *méthode de Frobenius*, encore appelée *méthode de la matrice associée* ou *de la matrice compagnon*, permet de déterminer les zéros d'un polynôme en résolvant un problème de détermination de valeurs propres. Soit $P(x)$ le polynôme

$$P(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

et A la matrice compagne (ou compagnon)

$$A = \begin{pmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & & \ddots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}$$

Comme

$$P(x) = (-1)^n \, det(A - xI)$$

les zéros de $P(x)$ sont les valeurs propres de A. Il suffit donc de savoir déterminer les valeurs propres d'une matrice pour résoudre une équation polynomiale. Si le polynôme est de la forme

$$Q(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$$

avec $b_n \neq 0$, il suffit de diviser chaque coefficient par b_n, et d'appliquer la méthode de Frobenius pour trouver les racines de Q qui coïncident avec les racines de Q/b_n.

3.12 Méthode de Bairstow

La *méthode de Bairstow* permet de déterminer les zéros d'un polynôme. C'est une application de la méthode de Newton-Raphson, qui consiste à factoriser à chaque étape un trinôme du second degré dont les racines sont les racines du polynôme initial. Soit

$$P(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n$$

on écrit P sous la forme

$$P(x) = (x^2 + px + q)P_{n-2}(x) + Rx + S$$

avec

$$P_{n-2}(x) = b_0 x^{n-2} + \cdots + b_{n-3}x + b_{n-2}$$

et on cherche à déterminer p et q de façon à annuler R et S (ce qui n'est pas toujours possible). On pose $b_{n-1} = R$ et $S = pb_{n-1} + b_n$. L'algorithme

est alors le suivant : On se donne deux constantes arbitraires p_0 et q_0 et on calcule les coefficients b_n définis par

$$\begin{cases} b_0 = a_0 \\ b_1 = a_1 - p_0 b_0 \\ b_2 = a_2 - p_0 b_1 - q_0 b_0 \\ \cdots \\ b_{n-1} = a_{n-1} - p_0 b_{n-2} - q_0 b_{n-3} \\ b_n = a_n - p_0 b_{n-1} - q_0 b_{n-2} \end{cases}$$

On calcule ensuite les coefficients c_n définis par

$$\begin{cases} c_0 = b_0 \\ c_1 = b_1 - p_0 c_0 \\ c_2 = b_2 - p_0 c_1 - q_0 c_0 \\ \cdots \\ c_{n-1} = b_{n-1} - p_0 c_{n-2} - q_0 c_{n-3} \\ c_n = b_n - p_0 c_{n-1} - q_0 c_{n-2} \end{cases}$$

En posant

$$p_1 = \frac{b_{n-1}c_{n-2} - b_n c_{n-3}}{c_{n-2}^2 - c_{n-1}c_{n-3}}$$

$$q_1 = \frac{b_n c_{n-2} - b_{n-1}c_{n-1}}{c_{n-2}^2 - c_{n-1}c_{n-3}}$$

On reprend les mêmes opérations avec les valeurs de p_1, q_1 à la place de p_0, q_0, et ainsi de suite. On obtient de cette manière un couple (p_j, q_j). On arrête l'itération lorsque le test d'arrêt suivant pour ε donné, est vérifié

$$\frac{|p_j - p_{j-1}| + |q_j - q_{j-1}|}{|p_j + q_j|} < \varepsilon$$

La résolution du trinôme $x^2 + p_j x + q_j = 0$ donne deux racines de $P_n(x)$. On recommence les mêmes opérations avec le polynôme $P_{n-2}(x)$ jusqu'à ce que le polynôme résiduel soit de degré inférieur à 2.

3.13 Méthode d'Aitken

Soit $P(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n$ un polynôme de degré n. Notons $x_1, x_2, ..., x_n$ ses racines et considérons les sommes

$$S_p = x_1^p + x_2^p + \cdots + x_n^p = \sum_{i=0}^{n} x_i^p$$

Comme $P(x)$ se factorise

$$P(x) = a_0(x - x_1)(x - x_2)...(x - x_n)$$

sa dérivée s'écrit

$$P'(x) = P(x)\sum_{i=0}^{n}\frac{1}{(x-x_i)} = P(x)(nx^{-1} + S_1x^{-2} + S_2x^{-3} + ...)$$

en identifiant avec l'expression de P'

$$P'(x) = na_0x^{n-1} + (n-1)a_1x^{n-2} + ... + a_{n-1}$$

on obtient les formules de calcul des S_j

$$\begin{cases} a_0S_1 + a_1 = 0 \\ a_0S_2 + a_1S_1 + 2a_2 = 0 \\ \cdots \\ a_0S_p + a_1S_{p-1} + \cdots + a_{p-1}S_1 + pa_p = 0 \\ \cdots \\ a_0S_n + a_1S_{n-1} + \cdots + a_{n-1}S_1 + na_n = 0 \end{cases}$$

Les déterminants

$$D_i(k) = \begin{vmatrix} S_k & S_{k+1} & \cdots & S_{k+i-1} \\ S_{k-1} & S_k & \cdots & S_{k+i-2} \\ \vdots & \vdots & & \vdots \\ S_{k-i+1} & S_{k-i+2} & \cdots & S_k \end{vmatrix}$$

vérifient la formule de récurrence suivante

$$D_{i+1}(k) = \frac{1}{D_{i-1}(k)}\begin{vmatrix} D_i(k) & D_i(k+1) \\ D_i(k-1) & D_i(k) \end{vmatrix}$$

ce qui permet de démontrer la formule donnant le produit de j racines

$$P_j = \lim_{k\longrightarrow\infty}\frac{D_j(k+1)}{D_j(k)} = x_1x_2x_3...x_j$$

La méthode d'Aitken consiste à calculer à partir des sommes S_j les déterminants D_j de façon à déterminer le produit des racines.

Exemple. Considérons l'équation

$$x^3 - 6x^2 + 11x - 6 = 0$$

Cette équation admet trois racines distinctes 1, 2 et 3. Calculons les sommes $S_0 = 3$, $S_1 = -a_1/a_0 = 6$, $S_2 = 14$, etc. À partir de ces valeurs, on calcule les déterminants $D_1(k)$, $D_2(k)$ et $D_3(k)$. D'où les valeurs

$$x_1 = \frac{D_1(12)}{D_1(11)} = \frac{535538}{179196} = 2.988...$$

puis

$$x_1 x_2 = \frac{D_2(11)}{D_2(10)} = \frac{-60703396}{-10156940} = 5.976...$$

et

$$x_1 x_2 x_3 = \frac{D_3(10)}{D_3(9)} = \frac{6718464}{1119744} = 6$$

On détermine ainsi de proche en proche l'ensemble des valeurs x_j.

3.14 Exercices

1. Déterminer le nombre de racines du polynôme

   $$P(x) = x^5 - 5x^3 + 4x$$

   dans l'intervalle $[-3, +3]$ sans calculer les racines.
2. Montrer que la fonction

   $$f(x) = \cos(x) - xe^x$$

   admet une racine unique dans l'intervalle $[0, \pi/2]$. Expliciter l'algorithme de Newton-Raphson sur cet exemple.
3. On considère la méthode de la sécante pour évaluer la solution de l'équation $f(x) = 0$

   $$x_{n+1} = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$$

   Montrer que l'erreur e_n commise à chaque pas est de la forme

   $$e_{n+1} \sim K e_n . e_{n-1}$$

   En déduire que l'ordre de la méthode est le nombre d'or $(1 + \sqrt{5})/2$.
4. Déterminer les racines du polynôme

   $$P(x) = x^4 - 5x^2 + 4$$

   par la méthode d'Aitken.
5. Déterminer les racines du polynôme

   $$x^3 - 6x^2 + 11x - 6 = 0$$

   par la méthode de Bairstow.

4
Intégration numérique

Dans les méthodes d'intégration, l'intégrale d'une fonction continue sur un intervalle borné $[a, b]$ est remplacée par une somme finie. Le choix de la subdivision de l'intervalle d'intégration et celui des coefficients qui interviennent dans la somme approchant l'intégrale sont des critères essentiels pour minimiser l'erreur. Ces méthodes se répartissent en deux grandes catégories : les *méthodes composées* dans lesquelles la fonction f est remplacée par un polynôme d'interpolation sur chaque intervalle élémentaire $[x_i, x_{i+1}]$ de la subdivision et les *méthodes de Gauss* fondées sur les polynômes orthogonaux pour lesquelles les points de la subdivision sont imposés.

4.1 Principes généraux

Soit f une fonction continue de $[a, b]$ dans $\mathbb{R}$. On se propose d'évaluer l'intégrale $\int_a^b f(x)d\mu(x)$ en subdivisant l'intervalle d'intégration

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$$

et en approchant f sur chaque intervalle par une somme finie de la forme

$$\int_a^b f(x)\, d\mu(x) \simeq \sum_{i=0}^{n-1} a_i\, f(x_i)$$

Une *méthode d'intégration* est dite *d'ordre* k si l'erreur commise en approchant l'intégrale par une somme discrète

$$e(f) = \int_a^b f(x)\, d\mu(x) - \sum_{i=0}^{n-1} a_i\, f(x_i)$$

est nulle lorsque f est un polynôme de degré inférieur ou égal à k et non nulle pour au moins un polynôme de degré supérieur ou égal à $k+1$. On rappelle le *théorème de Rolle* (1690). Soit P un polynôme. Entre deux racines de l'équation $P(x) = 0$, il existe au moins une racine de l'équation dérivée $P'(x) = 0$. Lorsque f est une fonction numérique intégrable sur un intervalle $[a, b]$, la *première formule de la moyenne* affirme que si f est continue alors il existe un nombre $c \in]a, b[$ tel que

$$\int_a^b f(x)dx = f(c)(b-a)$$

et la *deuxième formule de la moyenne* assure, lorsque ω est une fonction positive intégrable sur $]a, b[$ telle que l'intégrale $\int_a^b \omega(x)dx$ converge et pour toute fonction continue f sur $[a, b]$, l'existence d'un point $c \in]a, b[$ tel que

$$\int_a^b f(x)\omega(x)dx = f(c)\int_a^b \omega(x)dx$$

Soit u un opérateur linéaire borné, on appelle *noyau de Peano*, la fonction $K(t) = u(s_t)$ où s_t est la fonction

$$\begin{cases} s_t(x) = (x-t)_+^k = sup(0, (x-t)^k) & \text{si } k \neq 0 \\ s_t(x) = 1_{[t,\infty[}(x) & \text{si } k = 0 \end{cases}$$

Lorsque u est l'erreur d'une méthode d'intégration d'ordre k, la fonction $K(t) = e(s_t)$ est le noyau de Peano associé à cette méthode

$$K(t) = e(s_t) = \int_a^b (x-t)_+^k\, d\mu(x) - \sum_{i=0}^{n-1} a_i\, (x_i - t)_+^k$$

Pour une méthode numérique d'ordre k et pour une fonction f de classe C^{k+1} sur un intervalle $[a, b]$, l'erreur d'intégration est donnée par

$$e(f) = \int_a^b f(x)\, dx - \sum_{i=0}^{n-1} a_i\, f(x_i) = \frac{1}{k!}\int_a^b K(t)\, f^{(k+1)}(x)\, dx$$

$K(t)$ désigne le noyau de Peano associé à la méthode numérique. Si K garde un signe constant sur $[a, b]$, alors il existe $c \in [a, b]$ tel que

$$e(f) = \frac{f^{(k+1)}(x)}{(k+1)!} e(r_k)$$

la fonction r_k est définie par $r_k(x) = x^{k+1}$. Pour démontrer ce résultat, appliquons la formule de Taylor au point a. On a, puisque $x \geq t$

$$f(x) = p_k(x) + \frac{1}{k!} \int_a^x (x-t)_+^k \; f^{(k+1)}(t) \; dt$$

où p_k est un polynôme de degré inférieur ou égal à k. Comme la méthode est d'ordre k, l'erreur $e(p_k) = 0$ est nulle. Par conséquent

$$e(f) = \frac{1}{k!} \left[\int_a^b \int_a^b (x-t)_+^k \; f^{(k+1)}(t) \; dtdx - \sum_{i=0}^{n-1} a_i \int_a^b (x_i - t)_+^k \; f^{(k+1)}(t) \; dt \right]$$

d'où

$$e(f) = \frac{1}{k!} \int_a^b \left\{ \int_a^b (x-t)_+^k dx - \sum_{i=0}^{n-1} a_i (x_i - t)_+^k \right\} \; f^{(k+1)}(t) \; dt$$

soit finalement

$$e(f) = \frac{1}{k!} \int_a^b K(t) \; f^{(k+1)}(x) \; dx$$

Supposons que K soit de signe constant, appliquons la deuxième formule de la moyenne

$$\exists c \in [a,b], \qquad e(f) = \frac{f^{(k+1)}(c)}{k\,!} \int_a^b K(t) \; dt$$

en introduisant la fonction r_k pour laquelle

$$e(r_k) = (k+1) \int_a^b K(t) \; dt$$

on trouve la formule proposée.

4.2 Méthode des rectangles

Dans la méthode des rectangles, on remplace la fonction à intégrer f par une fonction constante par morceaux $h(x)$ sur chaque intervalle élémentaire $[x_i, x_{i+1}]$, soit par les rectangles à gauche : $h(x) = f(x_i)$ pour $x \in [x_i, x_{i+1}]$

$$\int_a^b f(x)dx \simeq \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i)$$

soit par les rectangles à droite : $h(x) = f(x_{i+1})$ pour $x \in [x_i, x_{i+1}]$

$$\int_a^b f(x)dx \simeq \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_{i+1})$$

On considère une fonction f continue sur $[a,b]$, dérivable sur l'intervalle ouvert $]a,b[$ et on se donne $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$ une subdivision régulière de l'intervalle $[a,b]$. On note h le pas de cette subdivision. Lorsque la subdivision se réduit à sa plus simple expression, $x_0 = a$, $x_1 = b$ on a

$$\int_a^b f(x)dx \simeq (b-a)f(a)$$

La méthode des rectangles est une méthode d'ordre 0. Lorsque la dérivée première de f est bornée par une constante M, l'erreur dans la méthode des rectangles est donnée par l'expression

$$\left| \int_a^b f(x)\ dx - h\sum_{i=0}^{n-1} f(a+ih) \right| \leq \frac{1}{2}\frac{(b-a)^2}{n} \sup_{x\in[a,b]} |f'(x)|$$

En effet, posons

$$F(h) = \int_\alpha^{\alpha+h} f(x)\ dx$$

On a $F'(h) = f(\alpha+h)$ et $F''(h) = f'(\alpha+h)$. En appliquant la formule de Taylor au deuxième ordre

$$\exists c \in]0,h[, \qquad F(h) = F(0) + hF'(0) + \frac{h^2}{2}F''(c)$$

Soit encore

$$\exists c \in]0,h[, \qquad \int_\alpha^{\alpha+h} f(x)\ dx = hf(\alpha) + \frac{h^2}{2}f'(\alpha+c)$$

Posons

$$S = h\sum_{i=0}^{n-1} f(a+ih)$$

En appliquant la formule précédente, on obtient la majoration cherchée

$$\begin{aligned} \left| \int_a^b f(x)\ dx - S \right| &\leq \sum_{i=0}^{n-1} \left| \int_{a+ih}^{a+(i+1)h} f(x)\ dx - hf(x_i) \right| \\ &\leq \frac{h^2}{2}\sum_{i=0}^{n-1} |f'(a+ih+c)| \\ &\leq \frac{1}{2}\frac{(b-a)^2}{n} \sup_{x\in[a,b]} |f'(x)| \end{aligned}$$

puisque $h = (b-a)/n$.

4.3 Méthode des trapèzes

Soit f une fonction continue sur $[a,b]$, dérivable sur $]a,b[$ et $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$ une subdivision régulière de l'intervalle $[a,b]$. On note h le pas de cette subdivision. Dans la méthode des trapèzes, la fonction f est remplacée sur chaque intervalle $[x_i, x_{i+1}]$ par la droite joignant les points $(x_i, f(x_i))$ et $(x_{i+1}, f(x_{i+1}))$, soit

$$h(x) = \frac{(x - x_i)f(x_{i+1}) - (x - x_{i+1})f(x_i)}{x_{i+1} - x_i} \qquad x \in [x_i, x_{i+1}]$$

La méthode s'écrit

$$\int_a^b f(x)dx \simeq \sum_{i=0}^{n-1}(x_{i+1} - x_i)\frac{f(x_i) + f(x_{i+1})}{2}$$

Lorsque la subdivision se réduit à sa plus simple expression, $x_0 = a$, $x_1 = b$ on a

$$\int_a^b f(x)dx \simeq \frac{1}{2}(b-a)(f(a) + f(b))$$

La méthode des trapèzes est une méthode d'ordre 1. L'erreur dans la méthode des trapèzes est donnée par l'expression

$$\left| \int_a^b f(x)\, dx - S \right| \leq \frac{1}{12}\frac{(b-a)^3}{n^2} \sup_{x\in[a,b]} |f''(x)|$$

La somme S s'exprime par

$$S = \frac{h}{2}\left(f(a) + f(b) + \sum_{i=1}^{n-1} f(x_i) \right)$$

Pour améliorer la précision, on considère parfois la *formule des trapèzes corrigée* suivante

$$\int_a^b f(x)\, dx \simeq \frac{h}{2}\left(f(a) + f(b) + \sum_{i=1}^{n-1} f(x_i) \right) - \frac{h^2}{12}\left(f'(b) - f'(a)\right)$$

4.4 Méthode de Simpson

Dans la méthode de Thomas Simpson (1710-1761), la fonction f est remplacée par un polynôme du second degré définissant un arc de parabole passant par les points d'ordonnées $f(x_i)$, $f(x_{i+1})$ et $f(x_{i+2})$. La méthode s'écrit

$$\int_a^b f(x)dx \simeq \sum_{i=0}^{n-1}\frac{1}{6}(x_{i+1} - x_i)\left(f(x_{i+1}) + f(x_i) + 4f(\frac{x_{i+1} + x_i}{2}) \right)$$

Lorsque la subdivision se réduit à sa plus simple expression, $x_0 = a$, $x_1 = (a+b)/2$, $x_2 = b$ la formule précédente devient

$$\int_a^b f(x)dx \simeq \frac{1}{3}(b-a)\left(f(a) + 4f(\frac{a+b}{2}) + f(b)\right)$$

La méthode de Simpson est une méthode d'ordre 4. L'erreur dans la méthode de Simpson est donnée par

$$\left|\int_a^b f(x)\,dx - S\right| \leq \frac{1}{2880}\frac{(b-a)^5}{n^4}\sup_{x\in[a,b]}\left|f^{(5)}(x)\right|$$

La somme S qui approche l'intégrale s'exprime par

$$S = \frac{h}{2}\sum_{i=0}^{n-1}(f(a+ih) + f(a+(i+1)h) + 4f(a+ih+\frac{h}{2}))$$

4.5 Méthode de Newton-Côtes

La méthode de Roger Côtes (1682-1716) publiée en 1707, généralise la méthode des trapèzes et la méthode de Simpson : la fonction f est approchée par un polynôme de degré n. L'intégrale est évaluée selon l'expression

$$\int_a^b f(x)\,dx \simeq a_0 f(x_0) + a_1 f(x_1) + \cdots + a_n f(x_n)$$

Pour déterminer les coefficients a_j, il suffit d'écrire que la relation précédente est exacte lorsque f est un polynôme de degré inférieur ou égal à n. En prenant successivement $f(x) = x^k$ pour $k = 0, 1, ..., n$, on obtient le système linéaire suivant

$$\begin{cases} a_0 + a_1 + \cdots + a_n = b - a \\ a_0x_0 + a_1x_1 + \cdots + a_nx_n = \dfrac{b^2-a^2}{2} \\ \cdots \\ a_0x_0^n + a_1x_1^n + \cdots + a_nx_n^n = \dfrac{b^n-a^n}{n+1} \end{cases}$$

Le déterminant de ce système est un déterminant de Vandermonde, qui vaut $(x_0 - x_1)(x_1 - x_2)...(x_n - x_0)$. Lorsque les points sont régulièrement espacés, on obtient les formules de Newton-Côtes. Pour $n = 1$ (méthode des trapèzes)

$$\int_{x_0}^{x_1} f(x)\,dx \simeq \frac{h}{2}\left(f(x_0) + f(x_1)\right)$$

Pour $n = 2$ (méthode de Simpson)

$$\int_{x_0}^{x_1} f(x)\,dx \simeq \frac{h}{2}\left(f(x_0) + 4f(x_1) + f(x_2)\right)$$

Pour $n = 3$

$$\int_{x_0}^{x_1} f(x)\,dx \simeq \frac{3h}{8}\left(f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right)$$

Pour $n = 4$ (méthode de Villarceau)

$$\int_{x_0}^{x_1} f(x)\,dx \simeq \frac{2h}{45}\left(7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)\right)$$

Pour $n = 6$ (méthode de Hardy)

$$\begin{aligned}\int_{x_0}^{x_1} f(x)\,dx \quad \simeq \quad & \frac{h}{140}\left(41f(x_0) + 216f(x_1) + 27f(x_2) + 272f(x_3) + \right. \\ & \left. 27f(x_4) + 216f(x_5) + 41f(x_6)\right)\end{aligned}$$

4.6 Méthode de Poncelet

La méthode de Poncelet est une amélioration de la méthode des trapèzes. L'intervalle de base $[a, b]$ est partagé en $2n$ parties égales $x_0 = a, x_1, .., x_{2n-1}, x_{2n} = b$. On note $h = (b - a)/2n$ le pas de la subdivision. Une première valeur approchée de l'intégrale est calculée par la méthode des trapèzes. Ensuite, sur chaque intervalle $[x_{2i-2}, x_{2i}]$, la fonction f est approchée par la tangente de f au point x_{2i-1}. Une deuxième valeur approchée de l'intégrale est alors calculée. L'intégrale est remplacée par la moyenne des deux valeurs calculées :

$$\int_a^b f(x)dx \simeq \frac{h}{4}\left(f(x_0) + f(x_{2n}) + 7\left(f(x_1) + f(x_{2n-1})\right) + 8\sum_{i=1}^{n-2} f(x_{2i+1})\right)$$

Notons M_i le point de coordonnées (x_i, f_i) et $f_i = f(x_i)$. Calculons la valeur de l'intégrale par la méthode des trapèzes. En remplaçant la courbe par la ligne polygonale $(M_0, M_1, M_3, ..., M_{2n-3}, M_{2n-1}, M_{2n})$, on obtient

$$I_1 = \frac{h}{2}f_0 + \frac{3}{2}hf_1 + 2hf_3 + 2hf_5 + ... + \frac{3}{2}hf_{2n-1} + \frac{h}{2}f_{2n}$$

La deuxième valeur est obtenue en remplaçant la courbe entre x_{2k} et x_{2k+1} par la tangente au point M_{2k+1}. En approchant la pente par l'expression

$$f'(x_{2i+1}) = \frac{f_{2i+1} - f_{2i}}{h}$$

on obtient

$$I_2 = 2hf_1 + ... + 2hf_{2n-1}$$

On en déduit l'estimation suivante en prenant la moyenne des deux valeurs précédentes

$$I = \frac{I_1 + I_2}{2}$$

4.7 Méthode de Romberg

La méthode de W. Romberg (1955) utilise l'extrapolation de Richardson à partir de 2^n applications de la méthode des trapèzes. Soit $A_{n,0}$ les évaluations de l'intégrale par la méthode des trapèzes

$$\begin{cases} A_{0,0} = \frac{b-a}{2}(f(a) + f(b)) \\ A_{1,0} = \frac{b-a}{4}(f(a) + f(b) + 2f(a + \frac{b-a}{2})) \\ \cdots \\ A_{n,0} = \frac{1}{2}A_{n-1,0} + \frac{(b-a)}{2^n}\sum_{k=0}^{2^{n-1}-1} f(a + (2k+1)\frac{b-a}{2^n}) \end{cases}$$

Si la dérivée seconde de f est continue bornée sur $[a, b]$, la suite $A_{n,0}$ converge vers la valeur exacte de l'intégrale. Pour accélérer la vitesse de la convergence, on applique l'extrapolation de Richardson, au couple $A_{n,0}$, $A_{n-1,0}$ pour définir $A_{n,1}$ qui converge vers la valeur de l'intégrale si la dérivée quatrième de f est continue bornée.

$$A_{n,1} = \frac{4A_{n,0} - A_{n-1,0}}{3}$$

De proche en proche, on définit ainsi les valeurs extrapolées

$$A_{n,l} = \frac{4^l A_{n,l-1} - A_{n-1,l-1}}{4^l - 1}$$

Lorsque n tend vers l'infini, on a alors

$$A_{n,l} = \int_a^b f(x)dx + O(4^{-n(l+1)})$$

4.8 Méthodes de Gauss

Les méthodes de Carl Friedrich Gauss (1777-1855) utilisent une subdivision particulière où les points x_j sont les racines d'une famille de polynômes orthogonaux, qui ne sont pas régulièrement espacés, contrairement aux méthodes composées. La fonction à intégrer est approchée par une interpolation de Lagrange sur les points x_j. Les méthodes de Gauss sont les

méthodes les plus répandues et les plus précises, car l'intégration est exacte pour tout polynôme de degré inférieur ou égal à $2n+1$ (au lieu de n ou $n+1$ dans les méthodes composées). Soit (Ψ_n) une famille de polynômes orthogonaux pour la fonction $\omega(x)$ sur l'intervalle $[u, v]$. Cherchons à exprimer l'intégrale $\int_u^v f(x)\omega(x)dx$. Écrivons la fonction f en utilisant la formule de Lagrange

$$f(x) = \sum_{i=0}^{n} L_i(x) f(x_i) + \prod_{i=0}^{n}(x - x_i)\frac{f^{(n+1)}(c)}{(n+1)\,!}$$

avec $c \in [u, v]$ et

$$L_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \left(\frac{x - x_j}{x_i - x_j}\right)$$

Si (Ψ_n) est une base de polynômes orthogonaux pour la fonction de poids $w(x)$, on a

$$\int_u^v \Psi_m(x)\Psi_n(x)w(x)dx = 0 \quad \text{si } n \neq m$$

Développons sur cette base le produit

$$\prod_{i=0}^{n}(x - x_i) = \sum_{i=0}^{n+1} a_i \Psi_i(x)$$

et si f est un polynôme de degré $(2n+1)$, notons

$$Q_n(x) = \frac{f^{(n+1)}(x)}{(n+1)\,!} = \sum_{i=0}^{n} b_i \Psi_n(x)$$

Le reste s'exprime par

$$\begin{aligned} R_n(x) &= \prod_{i=0}^{n}(x - x_i)\frac{f^{(n+1)}(x)}{(n+1)\,!} \\ &= \sum_{i=0}^{n}\sum_{j=0}^{n} a_i b_i \Psi_i(x)\Psi_j(x) + a_{n+1}\sum_{i=0}^{n} b_i \Psi_i(x)\Psi_{n+1}(x) \end{aligned}$$

d'où en intégrant

$$\int_u^v f(x)w(x)dx = \int_u^v \sum_{i=0}^{n} L_i(x) f(x_i) w(x)dx + \int_u^v R_n(x)w(x)dx + \varepsilon$$

soit en vertu de l'orthogonalité des polynômes

$$\int_u^v R_n(x)w(x)dx = \sum_{i=0}^{n} a_i b_i \int_u^v \Psi_i^2(x)w(x)dx$$

En choisissant les points (x_j) de la subdivision comme les $(n+1)$ racines du polynôme de degré $n+1$, on impose $a_i = 0$, pour $i = 0, 1, ..., n$ et $a_{n+1} \neq 0$, c'est-à-dire

$$\prod_{i=0}^{n}(x - x_i) = \sum_{i=0}^{n+1} a_i \Psi_i(x) = a_{n+1}\Psi_{n+1}(x)$$

d'où

$$\int_u^v R_n(x)w(x)dx = 0$$

Par conséquent, la méthode de Gauss appliquée à une fonction f conduit à une approximation de la forme

$$\int_u^v f(x)w(x)dx = \sum_{i=0}^{n} w_i f(x_i) + \varepsilon$$

avec

$$w_i = \int_u^v L_i(x)w(x)dx$$

L'erreur est de la forme $\varepsilon = \varepsilon_n f^{(2n+2)}(c)$ où ε_n dépend du choix des polynômes orthogonaux (Ψ_n).

4.9 Intégration de Gauss-Legendre

Lorsque la famille de polynômes orthogonaux est la famille des polynômes de Legendre relative à la fonction de pondération $w(x) = 1$ sur l'intervalle $[-1, 1]$, l'intégrale est approchée par la formule

$$\int_{-1}^{+1} f(x)dx = \sum_{i=0}^{n} w_i f(x_i) + \varepsilon$$

où les nombres w_i sont donnés par

$$w_i = \int_{-1}^{+1} \prod_{\substack{j=0 \\ j\neq i}}^{n} \left(\frac{x - x_j}{x_i - x_j}\right)dx$$

et les x_i sont les racines du polynôme de Legendre P_{n+1}. L'erreur s'exprime par

$$\varepsilon = \frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3} f^{(2n+2)}(c) \qquad \text{avec } c \in [-1, +1]$$

Exemple. Pour $n = 1$, la relation de récurrence définissant les polynômes de Legendre donne $P_2(x) = (3x^2 - 1)/2$. Ce polynôme admet deux racines

$x_0 = -1/\sqrt{3}$ et $x_1 = 1/\sqrt{3}$ définissant la subdivision de l'intervalle de base. Les valeurs w_i s'en déduisent. La première valeur se calcule par

$$w_0 = \int_{-1}^{+1} \frac{x - x_1}{x_0 - x_1} dx = \int_{-1}^{+1} \frac{-x + 1/\sqrt{3}}{2/\sqrt{3}} dx = 1$$

et de la même manière, on montre que $w_1 = 1$. L'intégrale se réduit à

$$\int_{-1}^{+1} f(x)dx \simeq f(\frac{-1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}})$$

Le changement de variables $y = (b + a)/2 + (b - a)x/2$ conduit à une approximation de l'intégrale

$$\int_a^b f(x)dx \simeq \frac{b-a}{2}(f(\frac{b+a}{2} - \frac{b-a}{2\sqrt{3}}) + f(\frac{b+a}{2} + \frac{b-a}{2\sqrt{3}}))$$

Exemple. Pour $n = 2$, le polynôme $P_3(x)$ admet trois racines, $x_0 = -\sqrt{3/5}$, $x_1 = 0$ et $x_2 = \sqrt{3/5}$. Le calcul des valeurs $w_0 = w_2 = 5/9$ et $w_1 = 8/9$ conduit à l'approximation de l'intégrale

$$\int_{-1}^{+1} f(x)dx \simeq \frac{5}{9} f(-\sqrt{\frac{3}{5}}) + \frac{8}{9} f(0) + \frac{5}{9} f(\sqrt{\frac{3}{5}})$$

4.10 Intégration de Gauss-Laguerre

Les polynômes de Laguerre sont orthogonaux sur l'intervalle $[0, \infty[$ relativement à la fonction de pondération $w(x) = e^{-x}$. Ils permettent de calculer une approximation de l'intégrale

$$\int_0^\infty f(x)e^{-x}dx \simeq \sum_{i=1}^{n} w_i f(x_i)$$

L'erreur est donnée par

$$\varepsilon = \frac{[(n+1)!]^2}{(2n+2)!} f^{(2n+2)}(c)$$

Exemple. Pour $n = 1$, le polynôme $P_2(x) = x^2 - 4x + 2$ admet deux racines $x_0 = 2 - \sqrt{2}$ et $x_1 = 2 + \sqrt{2}$. Les valeurs de w_i sont $w_0 = (2 + \sqrt{2})/4$ et $w_1 = (2 - \sqrt{2})/4$. D'où l'approximation

$$\int_0^\infty f(x)e^{-x}dx \simeq \frac{2+\sqrt{2}}{4} f(2 - \sqrt{2}) + \frac{2-\sqrt{2}}{4} f(2 + \sqrt{2})$$

4.11 Intégration de Gauss-Tchebychev

Les polynômes de Tchebychev forment une base orthogonale sur $[-1,+1]$ par rapport à la fonction de pondération $w(x) = 1/\sqrt{1-x^2}$. Les racines du polynôme T_{n+1} de degré $n+1$ sont données par

$$x_i = \cos\left(\frac{(2i+1)\pi}{2n+2}\right)$$

Les valeurs w_i ont, dans ce cas, une expression analytique générale $w_i = \pi/(n+1)$. Les polynômes de Tchebychev permettent de calculer une approximation de l'intégrale

$$\int_{-1}^{+1} f(x)\frac{1}{\sqrt{1-x^2}}dx \simeq \sum_{i=1}^{n} w_i f(x_i)$$

L'erreur est donnée par

$$\varepsilon = \frac{2\pi}{2^{2n+2}(2n+2)!} f^{(2n+2)}(c)$$

Exemple. Pour $n=1$, le polynôme du second degré $T_2(x) = 2x^2 - 1$ admet deux racines $x_0 = 1/\sqrt{2}$ et $x_1 = -1/\sqrt{2}$. Les valeurs $w_0 = w_1 = \pi/2$ conduisent à l'approximation

$$\int_{-1}^{+1} f(x)\frac{1}{\sqrt{1-x^2}}dx \simeq \frac{\pi}{2}(f(-1/\sqrt{2}) + f(1/\sqrt{2}))$$

4.12 Intégration de Gauss-Hermite

Les polynômes d'Hermite forment une base orthogonale sur l'intervalle $]-\infty,+\infty[$ par rapport à la fonction de pondération $w(x) = e^{-x^2}$. Ils permettent de calculer une approximation de l'intégrale

$$\int_{-1}^{+1} f(x)e^{-x^2}dx \simeq \sum_{i=1}^{n} w_i f(x_i)$$

L'erreur est donnée par

$$\varepsilon = \frac{(n+1)!\sqrt{\pi}}{2^{n+1}(2n+2)!} f^{(2n+2)}(c)$$

Exemple. Pour $n=1$, le polynôme d'Hermite $H_2(x) = 4x^2 - 2$ admet deux racines $x_0 = -1/\sqrt{2}$ et $x_1 = 1/\sqrt{2}$. Les valeurs $w_0 = w_1 = (\sqrt{\pi})/2$ conduisent à l'approximation suivante

$$\int_{-1}^{+1} f(x)e^{-x^2}dx \simeq \frac{\sqrt{\pi}}{2}(f(-1/\sqrt{2}) + f(1/\sqrt{2}))$$

4.13 Exercices

1. Calculer par les méthodes de Newton, l'intégrale

$$\int_0^1 \frac{1}{1+x^2}\, dx$$

Évaluer l'erreur commise.

2. Calculer par la méthode de Romberg l'intégrale suivante

$$\int_0^1 5(1-x^4)\, dx$$

et évaluer l'erreur commise.

3. Pour l'intégrale

$$\int_{-4}^{+4} e^{-x^2}\, dx$$

comparer les méthodes d'intégration des trapèzes, de Simpson, de Romberg et de Gauss.

4. Construire une méthode pour évaluer l'intégrale double

$$\int_0^2 \int_1^4 xy^2\, dxdy$$

Évaluer l'erreur commise.

5. On considère l'intégrale

$$\int_a^b f(x)\, dx$$

On se donne une subdivision $a = x_0$, x_1,..., $x_n = b$ de l'intervalle $[a, b]$ et on pose $x_j = a + jh$ et $h = (b-a)/n$. Dans la méthode des rectangles, on remplace la fonction f par une fonction constante par morceaux. Soit h la fonction définie par

$$h(x) = f(x_j) \quad si \quad x \in [x_j, x_{j+1}]$$

1) Montrer que si on pose

$$S = \int_a^b h(x)\, dx$$

on a

$$S = h\sum_{j=0}^{n-1} f(a+jh)$$

2) On suppose que la fonction f est continue sur l'intervalle $[a, b]$ et dérivable sur l'intervalle ouvert $]a, b[$. On pose

$$\varphi(x) = \int_{\alpha}^{\alpha+h} f(x)\ dx - hf(\alpha)$$

où $a \leq \alpha \leq \alpha + h \leq b$. Montrer qu'il existe un nombre $c \in]0, h[$ tel que

$$\varphi(h) = \frac{h^2}{2} f^{'}(c + \alpha)$$

3) On suppose que le module de la dérivée première de f est borné par un nombre k, montrer que

$$\left| \int_{a}^{b} f(x)\ dx - S \right| \leq \frac{1}{2} k \frac{(b-a)^2}{n}$$

4) Dans la *méthode des trapèzes*, on remplace la fonction f par une fonction affine par morceaux. Soit h la fonction définie sur l'intervalle $[x_j, x_{j+1}]$ par

$$h(x) = \frac{f(x_j + 1) - f(x_j)}{h}(x - x_j) + f(x_j)$$

Montrer que

$$S = \frac{h}{2} \sum_{j=0}^{n-1} (f(a + jh) + f(a + (j + 1)h))$$

5) Soit f et g deux fonctions continues sur l'intervalle $[a, b]$ et dérivables sur $]a, b[$. On pose

$$\varphi(x) = f(b) - f(x) - \frac{f(b) - f(a)}{g(b) - g(a)}(g(b) - g(x))$$

Montrer qu'il existe une constante c de $]a, b[$ vérifiant

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(c)}{g'(c)}$$

6) On suppose que f est deux fois dérivable sur $]a, b[$. Montrer que si on pose

$$\varphi(x) = \int_{\alpha}^{\alpha+h} f(x)\ dx - \frac{h}{2}(f(\alpha) + f(\alpha + h))$$

il existe une constante c sur $]0, h[$ telle que

$$\varphi(h) = h^3 \frac{\varphi''(c)}{6c}$$

7) En déduire qu'il existe une constante c de $]0, h[$ telle que

$$\varphi(h) = -\frac{h^2}{12} f''(\alpha + c)$$

Montrer que si le module de la dérivée seconde de f est borné par un nombre k, on a

$$\left| \int_a^b f(x)\, dx - S \right| \leq \frac{1}{12} k \frac{(b-a)^3}{n^2}$$

5
Systèmes linéaires

L'analyse matricielle étudie deux problèmes fondamentaux : l'inversion de matrices ou la résolution de systèmes linéaires qui fait l'objet du présent chapitre et le calcul des valeurs et des vecteurs propres d'une matrice qui sera traité dans le chapitre suivant. Les algorithmes de résolution des systèmes linéaires se classent en trois grandes catégories : les méthodes directes (méthodes de Gauss, Cholesky, Householder), les méthodes itératives (méthodes de Jacobi, Gauss-Seidel, relaxation) et les méthodes projectives (méthode de la plus profonde descente et méthodes du gradient conjugué). Les algorithmes et leurs implantations en machine mettent en jeu des techniques spéciales lorsque les matrices ont des formes particulières (matrices bandes, tridiagonales, creuses, diagonales par blocs, etc.).

5.1 Généralités sur les matrices

L'ensemble des matrices à m lignes n colonnes à coefficients dans un corps $\mathbb{K}$ ($\mathbb{R}$ ou $\mathbb{C}$) est un espace vectoriel noté $M_{m,n}(\mathbb{K})$ de dimension $m \times n$. L'ensemble des matrices carrées $n \times n$ est simplement noté $M(n, \mathbb{K})$. Les théorèmes obtenus dans le cas complexe s'appliquent au cas réel en remplaçant les termes : adjoint (par transposée), complexe (par réel), unitaire (par orthogonal) et hermitien (par symétrique). La matrice *adjointe* de A est notée A^*, elle est définie par $A^* = \overline{A}^t$. Une matrice A est dite *hermitienne* si $A^* = A$, autrement dit si ses coefficients vérifient $a_{ji} = \overline{a_{ij}}$ (en particulier ses coefficients diagonaux sont réels). Une matrice A est dite

inversible s'il existe une matrice notée A^{-1} telle que $AA^{-1} = A^{-1}A = I$ et *singulière* ou *non inversible* dans le cas contraire. La matrice A est dite *symétrique* si $A^t = A$ (autrement dit si $a_{ij} = a_{ji}$) et *antisymétrique* si $A^t = -A$. On dit que A est *orthogonale* si A est une matrice telle que $A^{-1} = A^t$. Une matrice A est dite *unitaire* si $A^{-1} = A^*$ et A est dite *normale* si $AA^* = A^*A$. Une matrice L est dite *triangulaire inférieure* si $a_{ij} = 0$ pour $j > i$ et triangulaire inférieure stricte si les termes diagonaux sont nuls. Une matrice U est dite *triangulaire supérieure* si la transposée de U est triangulaire inférieure $a_{ij} = 0$ si $i > j$

$$L = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ \vdots & & & 0 \\ a_{n1} & \cdots & \cdots & a_{nn} \end{pmatrix} \qquad U = \begin{pmatrix} a_{11} & \cdots & \cdots & a_{1n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}$$

Une matrice T est dite *tridiagonale* si $a_{ij} = 0$ si $i \geq j + 2$ ou si $j \geq i + 2$. Une matrice est appelée *matrice bande* si $a_{ij} = 0$ si $i \geq j + k + 1$ ou si $j \geq i + k + 1$, la bande est dite de largeur $(2k + 1)$. Une matrice H est une *matrice de Hessenberg* si $a_{ij} = 0$ si $i \geq j + 2$

$$T = \begin{pmatrix} \times & \times & 0 & & 0 \\ \times & \times & \times & & \\ 0 & \times & \times & \ddots & 0 \\ & & \ddots & \ddots & \times \\ 0 & & 0 & \times & \times \end{pmatrix} \qquad H = \begin{pmatrix} \times & \times & \cdots & \times & \times \\ \times & \times & \times & \cdots & \times \\ 0 & \times & \times & \ddots & \vdots \\ & & \ddots & \ddots & \times \\ 0 & & 0 & \times & \times \end{pmatrix}$$

Une *matrice de Toeplitz* est une matrice dont les éléments sont identiques sur chaque diagonale $a_{ij} = a_{i+1,j+1}$. Une *matrice de Hankel* est une matrice dont les éléments sont identiques sur les antidiagonales $a_{ij} = a_{i+1,j-1}$.

Changement de base. Si A est la matrice d'une application linéaire u exprimée dans une base (e_i) et si (f_i) est une autre base, on note P la matrice de passage de la base (e_i) à la base (f_i) dont le jème vecteur colonne est formé des composantes du vecteur f_j dans la base (e_i). La matrice de u dans la base (f_j) est donnée par

$$B = P^{-1}AP$$

Le polynôme caractéristique d'une matrice carrée A de $M(n, \mathbb{C})$ s'écrit

$$\begin{aligned} P(\lambda) &= \det(A - \lambda I) \\ &= (-1)^n\lambda^n + (-1)^{n-1}Tr(A)\lambda^{n-1} + (-1)^{n-2}s_2\lambda^{n-2} + \cdots \\ &\quad \cdots - s_{n-1}\lambda + \det(A) \end{aligned}$$

où s_k est égal à $(-1)^k$ fois la somme des $r = C_n^k$ mineurs principaux d'ordre k. En particulier, le polynôme caractéristique d'une matrice carrée d'ordre 2 vaut $P(\lambda) = \det(A - \lambda I) = \lambda^2 - Tr(A)\lambda + \det(A)$. Par exemple pour la matrice

$$A = \begin{pmatrix} 1 & -4 & -1 \\ 2 & 0 & 5 \\ -1 & 1 & -2 \end{pmatrix}$$

les quantités s_k valent : $s_1 = Tr(A) = -1$, $s_n = det(A) = -3$ et

$$s_2 = \begin{vmatrix} 1 & -4 \\ 2 & 0 \end{vmatrix} + \begin{vmatrix} 1 & -1 \\ -1 & -2 \end{vmatrix} + \begin{vmatrix} 0 & 5 \\ 1 & -2 \end{vmatrix} = 0$$

Diagonalisation. Soit E un espace vectoriel de dimension finie sur $\mathbb{C}$, u un endomorphisme de E de matrice A. La matrice A est semblable à une matrice de la forme

$$\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & A_p \end{pmatrix}$$

où A_p est une matrice carrée de polynôme caractéristique $(\lambda_i - \lambda)^{h_i}$. Si $\lambda_1, \lambda_2, ...\lambda_p$ sont les valeurs propres distinctes de A, le polynôme caractéristique s'écrit

$$P(\lambda) = \det(A - \lambda I) = (\lambda_1 - \lambda)^{h_1}(\lambda_2 - \lambda)^{h_2} \dots (\lambda_p - \lambda)^{h_p}$$

Le noyau $E_i = Ker(A - \lambda_i I)^{h_i}$ est stable par u et E est somme directe des sous-espaces E_i

$$E = E_1 \oplus E_2 \oplus \cdots \oplus E_p$$

Une matrice A est diagonalisable s'il existe une base dans laquelle A est semblable à une matrice diagonale D de la forme

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

Soit A une matrice carrée d'ordre n. Si son polynôme caractéristique n'a que des racines simples $\lambda_1, \lambda_2, ...\lambda_n$, alors A est diagonalisable.

$$P(\lambda) = \det(A - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \dots (\lambda_n - \lambda)$$

Une matrice N est dite nilpotente d'indice p si $N^p = 0$ et $N^k \neq 0$, $1 \leq k \leq p - 1$. Par exemple, la matrice suivante est nilpotente.

$$\begin{pmatrix} 0 & 3 & 4 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

Jordanisation. Soit A une matrice carrée d'ordre n. Il existe une base de E telle que la matrice A dans cette base soit de la forme

$$J = \begin{pmatrix} \lambda_1 & \upsilon_1 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & & \upsilon_{n-1} \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

c'est-à-dire formée des valeurs propres sur sa diagonale et des valeurs $\upsilon_i = 0$ ou 1 sur la diagonale supérieure. La matrice J est appelée *forme de Jordan*. Il existe une matrice diagonale D et une matrice nilpotente N telles que $A = D + N$ et $DN = ND$. Cette décomposition est unique. On rappelle également le *théorème de Cayley-Hamilton,* qui affirme que si P est le polynôme caractéristique d'un endomorphisme u, alors $P(u) = 0$. Si A est la matrice de u dans une base de E, on a $P(A) = 0$. Rappelons que les valeurs propres d'une matrice hermitienne (resp. symétrique, resp. unitaire) sont réelles. Une matrice hermitienne A est positive si $x^*Ax \geq 0$, $\forall x \in E$. Cette matrice A est dite définie positive si $x^*Ax > 0$, $\forall x \in E \setminus \{0\}$. Une matrice hermitienne est *définie positive* (resp. positive) si et seulement si toutes ses valeurs propres sont > 0 (resp. ≥ 0). Un résultat important pour les algorithmes numériques est le *théorème de Schur* qui affirme que si A est une matrice carrée à coefficients complexes, il existe une matrice unitaire U telle que U^*AU soit triangulaire supérieure de la forme

$$\begin{pmatrix} \lambda_1 & b_{12} & \cdots & b_{1n} \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & b_{n-1,n} \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

où les λ_i sont les valeurs propres de A. Le corollaire de ce théorème affirme que si A est une matrice hermitienne (ou symétrique réelle), il existe une matrice unitaire U telle que $U^*AU = D$ où D est une matrice diagonale dont les éléments diagonaux sont les valeurs propres de A. Ces valeurs propres sont réelles. En particulier, une matrice hermitienne (ou symétrique réelle) a ses vecteurs propres orthogonaux. Notons que pour qu'une matrice A soit *normale* ($AA^* = A^*A$), il faut et il suffit qu'elle soit de la forme $A = U^*DU$ où U est une matrice unitaire et D une matrice diagonale. Une matrice triangulaire supérieure est normale si et seulement si elle est diagonale. Les matrices normales, et en particulier les matrices hermitiennes, sont diagonalisables.

Normes. Sur l'espace vectoriel $M_n(\mathbb{C})$ on emploie traditionnellement plusieurs normes. Soit A une matrice carrée d'ordre n, l'application définie

par

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_{l^p}}{\|x\|_{l^p}} \qquad \text{avec } \|x\|_{l^p} = \left(\sum_i |x_i|^p\right)^{1/p}$$

est une norme sur l'espace des matrices à coefficients complexes $M(n,\mathbb{C})$. Les normes les plus usuelles sont définies par les relations $(p = 1, 2, \infty)$

$$\begin{aligned} \|A\|_1 &= \max_j \sum_i |a_{ij}| \\ \|A\|_2 &= \|A^*\|_2 = \sqrt{\rho(AA^*)} = \sqrt{\rho(A^*A)} \\ \|A\|_\infty &= \max_i \sum_j |a_{ij}| \end{aligned}$$

on emploie aussi la norme euclidienne définie par

$$\|A\|_e = \sqrt{tr(A^*A)} = \left(\sum_{i,j} |a_{i,j}|^2\right)^{1/2}$$

elle vérifie l'inégalité

$$\|A\|_2 \leq \|A\|_e \leq \sqrt{n}\,\|A\|_2 \qquad \forall A \in M_n(\mathbb{C})$$

Toute transformation unitaire $U \in M_n(\mathbb{C})$ laisse invariantes les normes $\|A\|_2$ et $\|A\|_e$

$$\begin{aligned} \|A\|_e &= \|UA\|_e = \|AU\|_e \\ \|A\|_2 &= \|UA\|_2 = \|AU\|_2 \end{aligned}$$

Soit u un endomorphisme de l'espace vectoriel E, on appelle *rayon spectral* de u et on note $\rho(u)$ la borne inférieure des nombres réels $\|u^n\|^{1/n}$ pour tous les entiers n non nuls.

$$\rho(u) = \inf_{n>0} \|u^n\|^{1/n}$$

Le rayon spectral est indépendant de la norme choisie. Ce nombre vérifie les propriétés suivantes

$$\rho(u) = \lim \|u^n\|^{1/n}$$

Le rayon spectral est le plus grand module des valeurs propres de u

$$\rho(u) = \max_{\lambda \in Sp(u)} (|\lambda|)$$

Le rayon spectral vérifie

$$\forall \alpha \in \mathbb{C} \qquad \rho(\alpha u) = |\alpha|\,\rho(u)$$

Le rayon spectral d'une matrice à la puissance n est la puissance n-ième du rayon spectral de la matrice d'origine

$$\forall n \in \mathbb{N}^* \qquad \rho(u^n) = \rho(u)^n$$

Si le rayon spectral est nul, alors toutes les valeurs propres de u sont nulles (u est nilpotent). Soit $A \in M_n(\mathbb{C})$ une matrice à coefficients complexes et une norme quelconque, alors le rayon spectral de A est inférieur à la norme de A.

$$\rho(A) \leq \|A\|$$

5.2 Méthodes directes

5.2.1 Méthode de remontée

On se propose de résoudre l'équation matricielle $Ax = b$. On suppose que la matrice A est inversible. Lorsque A est une matrice triangulaire supérieure (ou inférieure), la résolution du système est immédiate

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \cdots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \\ a_{n,n}x_n = b_n \end{cases}$$

On calcule successivement x_n à partir de la dernière équation, puis x_{n-1} à partir de l'avant-dernière et ainsi de suite. Ce qui donne

$$\begin{cases} x_n = b_n/a_{nn} \\ x_{n-1} = (b_{n-1} - a_{n-1,n}x_n)/a_{n-1,n} \\ \cdots \\ x_1 = (b_1 - a_{12}x_2 - \ldots - a_{1n}x_n)/a_{11} \end{cases}$$

La méthode de remontée s'étend aux matrices triangulaires par blocs. Elle nécessite $n(n-1)/2$ additions, $n(n-1)/2$ multiplications et n divisions. Étant donné la simplicité de la résolution d'un système triangulaire, de nombreuses méthodes se ramènent à la résolution d'un système triangulaire. Le problème est alors de construire par un changement de base une matrice triangulaire.

5.2.2 Élimination de Gauss

La *méthode de triangularisation de Gauss*, encore appelée *méthode du pivot de Gauss* ou *élimination de Gauss*, est fondée sur le théorème suivant qui affirme que pour une matrice carrée A d'ordre n, il existe au moins une matrice inversible P telle que PA soit une matrice triangulaire supérieure.

L'algorithme consiste alors à remplacer à chaque étape la matrice A par une matrice $A^{(k)}$ dont les k-ièmes premiers vecteurs colonnes correspondent au début d'une matrice triangulaire. À la $(k+1)$-ième étape, on conserve les k premières lignes et les $(k-1)$ premières colonnes de $A^{(k)}$

$$\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} && i = k+1, ..., n \text{ et } j = k+1, ..., n \\ a_{ik}^{(k+1)} &= 0 && i = k+1, ..., n \\ b_i^{(k+1)} &= b_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} b_k^{(k)} \end{aligned}$$

En pratique, si le *pivot*, c'est-à-dire l'élément $a_{kk}^{(k)}$ situé à la k-ième ligne et à la k-ième colonne, est petit ou nul, l'algorithme n'est plus valable. On emploie dans ce cas des permutations de lignes et de colonnes appelées *stratégies de pivot* (voir le paragraphe *Problème des pivots*).

Exemple. Considérons le système linéaire

$$\begin{pmatrix} 2 & 8 & 4 \\ 2 & 10 & 6 \\ 1 & 8 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

À la première étape, on fait apparaître le vecteur $(1, 0, ..., 0)$ à la première colonne. Pour cela, on divise la première ligne par 2 (le terme $a_{11} = 2$ est pris comme pivot) et on retranche la première ligne aux autres lignes de la matrice, soit

$$\begin{pmatrix} 1 & 4 & 2 \\ 0 & 2 & 2 \\ 0 & 4 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \end{pmatrix}$$

À la deuxième étape, on poursuit la triangulation en annulant les termes situés sous la diagonale. On divise la deuxième ligne par 2 (a_{22} est pris comme pivot) et on retranche la deuxième ligne multipliée par 4 à la troisième, de façon à faire apparaître un zéro en troisième ligne et deuxième colonne

$$\begin{pmatrix} 1 & 4 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \end{pmatrix}$$

À la troisième étape, on divise la troisième ligne par -4 (a_{33}) afin d'avoir une matrice triangulaire n'ayant que des 1 sur la diagonale

$$\begin{pmatrix} 1 & 4 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ -1/8 \end{pmatrix}$$

On résout le système en remontant les équations : $z = -1/8$, donc $y = 1/8$ et $x = 1/4$. Pour une matrice d'ordre n, la méthode de Gauss nécessite $n(n-1)(2n+5)/6$ additions, $n(n-1)(2n+5)/6$ multiplications et $n(n+1)/2$ divisions, soit au total $(4n^3 + 9n^2 - 7n)/6$ opérations élémentaires. En utilisant les formules de Cramer, on aurait $(n+1)(n!-1)$ additions, $(n+1)(n-1)n!$ multiplications et n divisions. Pour $n = 10$, la méthode de Gauss nécessite 805 opérations contre 399 168 000 opérations pour la résolution par les formules de Gabriel Cramer (1704-1752), $x_i = \det B_i / \det A$ où B_i est la matrice formée des éléments a_{ij} sauf sur la colonne i où on place les éléments du vecteur b.

5.2.3 Méthode de Gauss-Jordan

Dans la méthode de Gauss-Jordan, on cherche non pas à trianguler A comme dans la méthode de Gauss, mais à remplacer A par l'identité.

Exemple. Reprenons l'exemple proposé

$$\begin{pmatrix} 2 & 8 & 4 \\ 2 & 10 & 6 \\ 1 & 8 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

À la première étape, on fait apparaître le vecteur $(1, 0, ..., 0)$ en première colonne. On divise la première ligne par 2 (a_{11} est appelé *le pivot*) et on retranche la première ligne aux autres lignes de la matrice

$$\begin{pmatrix} 1 & 4 & 2 \\ 0 & 2 & 2 \\ 0 & 4 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \end{pmatrix}$$

À la deuxième étape, on fait apparaître le vecteur $(0, 1, 0)$ dans la deuxième colonne. Pour cela, on divise la deuxième ligne par 2. On retranche à la première et troisième lignes 4 fois la deuxième ligne, soit

$$\begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \end{pmatrix}$$

À la troisième étape, on fait apparaître le vecteur (0,0,1) dans la dernière colonne. On divise la troisième ligne par (-4). Puis, on retranche (-2) fois la troisième ligne à la première et une fois la troisième ligne à la deuxième. On obtient ainsi directement la solution du système.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/4 \\ 1/8 \\ -1/8 \end{pmatrix}$$

Le même algorithme est utilisé pour calculer l'inverse d'une matrice. On écrit A sous la forme AI et on applique à l'identité I toutes les manipulations que subit A.

Exemple. Reprenons le même exemple, écrivons

$$A = \begin{pmatrix} 2 & 8 & 4 \\ 2 & 10 & 6 \\ 1 & 8 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

À la première étape, on divise la première ligne de A et de I par 2. On retranche à la deuxième ligne (de A et de I) les éléments de la première ligne multipliés par 2 et à la troisième ligne les éléments de la première

$$\begin{pmatrix} 1 & 4 & 2 \\ 0 & 2 & 2 \\ 0 & 4 & 0 \end{pmatrix} \begin{pmatrix} 1/2 & 0 & 0 \\ -1 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix}$$

À la deuxième étape, on divise la deuxième ligne par 2. On fait apparaître le vecteur (0,1,0) dans la deuxième colonne

$$\begin{pmatrix} 1 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & -4 \end{pmatrix} \begin{pmatrix} 5/2 & -2 & 0 \\ -1/2 & 1/2 & 0 \\ 3/2 & -2 & 1 \end{pmatrix}$$

À la troisième étape, on fait apparaître le vecteur $(0, 0, 1)$ dans la dernière colonne. On obtient ainsi l'inverse de la matrice A.

$$A^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 7/4 & -3 & -1/2 \\ -1/8 & 1 & 1/4 \\ -3/8 & -1/2 & -1/4 \end{pmatrix}$$

Si A est une matrice réelle, la méthode de Gauss-Jordan nécessite $n(n^2 - 1)/2$ multiplications, $n(n^2 - 1)/2$ additions et $n(n + 1)/2$ divisions.

5.2.4 Problème des pivots

Lorsqu'un pivot est nul, la méthode de Gauss ou de Jordan n'est plus applicable. Si le pivot est très petit, l'algorithme conduit à des erreurs d'arrondi importantes. C'est pourquoi des algorithmes qui échangent les éléments de façon à avoir le pivot le plus grand possible ont été développés. Les programmes optimisés intervertissent les lignes à chaque étape de façon à placer en pivot le terme de coefficient le plus élevé de la ligne : c'est la méthode du *pivot partiel*, à la k-ième étape le pivot est l'élément

$$a_{ik}^{(k)} = \max_{p=k,..,n} \left| a_{pk}^{(k)} \right|$$

D'autres programmes intervertissent les lignes et les colonnes de façon à placer en pivot le terme de coefficient le plus élevé de la matrice : c'est la méthode du *pivot total*. À la k-ième étape, le pivot est l'élément

$$a_{ij}^{(k)} = \max_{p,q=k,\ldots,n} \left| a_{pq}^{(k)} \right|$$

Exemple. Considérons le système

$$\begin{cases} 10^{-4}x + y = 1 \\ x + y = 2 \end{cases}$$

La solution de ce système est $x = 1,0001$ et $y = 0,99990$. Supposons que notre calculateur travaille avec une mantisse de trois chiffres significatifs. Comme $a_{11} = 10^{-4}$ est très petit, l'élimination conduit au système suivant obtenu en multipliant la première équation par (-10^4) et en ajoutant la seconde :

$$\begin{cases} 10^{-4}x + y = 1 \\ -9990\, y = -9990 \end{cases}$$

si les chiffres $(-10^4+1 = -9999)$ et $(-10^4+2 = -9998)$ sont approchés par le même nombre -9990. La solution devient alors $x = 1$ et $y = 0$. L'erreur est importante pour le nombre y. En revanche, si on échange les équations

$$\begin{cases} x + y = 2 \\ 10^{-4}x + y = 1 \end{cases}$$

et on prend pour pivot l'élément $a_{11} = 1$, on obtient

$$\begin{cases} x + y = 2 \\ 0.999y = 0.999 \end{cases}$$

car les chiffres $(-10^{-4}+1 = 0.9999)$ et $(-2.10^{-4}+1 = 0.9998)$ sont arrondis au même nombre 0.999. Dans ce cas, la solution est correcte $(x = y = 1)$.

Exemple. Considérons un calculateur travaillant sur quatre chiffres significatifs. Soit à résoudre le système suivant

$$\begin{pmatrix} 0,001 & 1 & 1 \\ -2 & 4,732 & 2,736 \\ -1 & 3,643 & 1,821 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2,001 \\ 5,468 \\ 4,464 \end{pmatrix}$$

Ce système admet comme solution le triplet (1, 1, 1). En prenant comme pivot le terme 0,001 et ne retenant pour chaque calcul que quatre chiffres significatifs, on a

$$\begin{pmatrix} 0,001 & 1 & 1 \\ 0 & 2005 & 2003 \\ 0 & 1004 & 1002 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2,001 \\ 4007 \\ 2005 \end{pmatrix}$$

car $5,468+2000\times 2,001 \simeq 4007$ et $4,464+1000\times 2,001 \simeq 2005$. L'élimination, compte tenu du fait que $1004/2005 \simeq 0,501$ et $1002-0,501\times 2003 \simeq -2,000$ et $2005-0,501\times 4007 \simeq -3,000$ conduit à

$$\begin{pmatrix} 0,001 & 1 & 1 \\ 0 & 2005 & 2003 \\ 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2,001 \\ 4007 \\ -3 \end{pmatrix}$$

D'où les solutions par la méthode de remontée $z = 1,5$ puis $y = 0,4998$ et enfin $x = 1,000$. Dans ce cas, l'erreur relative sur y et z est de 50 %.

5.2.5 Méthode de Crout. Factorisation LU

La *méthode de Crout* est fondée sur la *factorisation LU* qui affirme que pour une matrice carrée $A = (a_{ij})$ d'ordre n telle que les n sous-matrices

$$\Delta_k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix}$$

soient inversibles, il existe une matrice triangulaire inférieure $L = (l_{ij})$ avec $l_{ii} = 1$ $(1 \leq i \leq n)$ et une matrice triangulaire supérieure U telles que $A = LU$. Cette décomposition est unique. En particulier, toute matrice inversible admet une factorisation LU. L'algorithme est alors le suivant : On calcule les matrices L et U telles que $A = LU$ par les formules

$$\begin{aligned} &u_{1j} = a_{1j} \text{ si } j = 1, ..., n \\ &l_{i1} = a_{i1}/u_{11} \text{ si } i = 1, ..., n \\ &l_{ii} = 1 \\ &l_{ij} = 0 \text{ si } j = i+1, .., n \\ &u_{ij} = 0 \text{ si } j = 1, ..., i-1 \\ &u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj} \\ &lij = (a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj})/u_{jj} \end{aligned}$$

Puis, on résout le système $Ax = b$ par la méthode de remontée en remarquant que si on pose $y = Ux$, le système s'écrit $Ly = b$. Le système est alors résolu par la méthode de remontée en y. La même méthode, appliquée en sens inverse, donne les valeurs de x.

Exemple. Cas d'une matrice tridiagonale. Soit A la matrice

$$A = \begin{pmatrix} b_1 & c_1 & 0 & 0 & \cdots & 0 \\ a_2 & b_2 & c_2 & 0 & & 0 \\ 0 & a_3 & b_3 & c_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & \cdots & \cdots & 0 & a_n & b_n \end{pmatrix}$$

Mise sous la forme LU, la matrice A s'écrit

$$A = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ l_1 & 1 & 0 & & \vdots \\ 0 & l_2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & l_{n-1} & 1 \end{pmatrix} \begin{pmatrix} u_1 & c_1 & 0 & \cdots & 0 \\ 0 & u_2 & c_2 & \ddots & \vdots \\ \vdots & 0 & u_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & \cdots & 0 & 0 & u_n \end{pmatrix}$$

avec

$$\begin{aligned} &l_i = a_{i+1}/u_i \text{ si } i = 1, ..., n-1 \\ &u_1 = b_1 \\ &u_i = b_i - \frac{a_i c_{i-1}}{u_{i-1}} \text{ si } i = 2, ..., n \end{aligned}$$

Posons $Z = UX$. L'expression $AX = LUX = Y$ devient $LZ = Y$ et l'équation $LZ = Y$ s'écrit

$$\begin{cases} z_1 = y_1 \\ z_i = y_i - l_{i-1} z_{i-1} \text{ si } i = 2, ..., n \end{cases}$$

L'équation $UX = Z$ permet alors de calculer x_i.

$$\begin{cases} x_n = z_n/u_n \\ x_i = (z_i - c_i x_{i+1})/u_i \text{ si } i = 1, ..., n-1 \end{cases}$$

La méthode est aussi une méthode pour calculer le déterminant d'une matrice tridiagonale. Soit D_n le déterminant d'ordre n, on vérifie par récurrence que

$$\begin{cases} D_0 = 1 \\ D_1 = b_1 \\ D_n = b_n D_{n-1} - a_n c_{n-1} D_{n-2} \end{cases}$$

Comme $A = LU$, on a

$$det(A) = det(L) det(U) = \prod_{i=1}^{n} u_i$$

d'où, en posant $c_0 = 0$ et $u_0 = 1$

$$D_n = \prod_{i=1}^{n} u_i = \prod_{i=1}^{n} (b_i - a_i \frac{c_{i-1}}{u_{i-1}})$$

5.2.6 Méthode de Cholesky

La méthode d'André-Louis Cholesky (1875-1918) s'applique aux matrices hermitiennes ou symétriques. Elle est fondée sur le théorème suivant qui affirme que pour une matrice hermitienne (resp. symétrique) définie positive A, il existe (au moins) une matrice triangulaire supérieure inversible U telle que $A = U^*U$ (resp. $A = U^tU$). Si les éléments diagonaux de U sont strictement positifs, la matrice U est unique. Ce théorème est valable pour une matrice triangulaire inférieure. Soit A une matrice symétrique définie positive à coefficients réels, mettons A sous la forme $A = L^tL$ en appliquant l'algorithme à la i-ième étape, on calcule la matrice L par

$$l_{ij} = (a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{kj})/l_{jj} \qquad j = 1, ..., i-1$$
$$l_{ii} = (a_{ij} - \sum_{k=1}^{i-1} l_{ik}^2)^{1/2}$$

On résout ensuite le système $Ly = b$ puis $L^tx = y$ par un double balayage

$$y_i = (b_i - \sum_{k=1}^{i-1} l_{ik} y_k)/l_{ii} \qquad i = 1, ..., n$$
$$x_i = (y_j - \sum_{k=i+1}^{n} l_{ki} x_k)/l_{ii}$$

5.2.7 Méthode de Householder. Factorisation QR

La *méthode de triangularisation de Householder* pour la résolution du système $Ax = b$ consiste à trouver $(n-1)$ matrices de Householder $H_1, ..., H_n$ telles que la matrice produit $H_{n-1}... H_2H_1A$ soit triangulaire supérieure. On résout alors le système par la méthode de remontée. La méthode repose sur la *factorisation QR* qui affirme que pour une matrice carrée d'ordre n à coefficients complexes, il existe une matrice unitaire unique Q et une matrice triangulaire supérieure, à éléments diagonaux positifs ou nuls, telles que $A = QR$. Si A est inversible, la décomposition est unique. Si A est une *matrice de Hessenberg supérieure*, Q est aussi une matrice de Hessenberg supérieure. Soit x un vecteur colonne non nul, $x \in \mathbb{C}^n \setminus \{0\}$, une matrice de Householder est une matrice de la forme

$$H(x) = I - \frac{2}{x^*.x} x.x^*$$

où $x.x^*$ désigne le produit scalaire de x par $x^* = \overline{x}^t$. Les matrices de Householder sont hermitiennes, unitaires, conservent la 2-norme et réalisent une symétrie par rapport à l'hyperplan passant par P et orthogonal à x. Dans le cas réel, les matrices de Householder sont de la forme

$$H(x) = I - \frac{2}{x^t.x} x.x^t = I - \frac{2}{\|x\|_2^2} \begin{pmatrix} x_1^2 & x_1x_2 & ... & x_1x_n \\ x_2x_1 & x_2^2 & ... & x_2x_n \\ & & \ddots & \\ x_nx_1 & x_nx_2 & ... & x_n^2 \end{pmatrix}$$

Les matrices de Householder H sont orthogonales et symétriques, d'où $H^{-1} = H$. L'algorithme de la méthode de Householder est le suivant : À la k-ième étape, pour $k = 1, 2, ..., n-1$, on construit un vecteur a orthonormé

$$\|a\| = \left(\sum_{i=k}^{n} a_{ik} \overline{a_{ik}} \right)^{1/2}$$

Si $\|a\| = 0$ ou si $|a_{kk}| = 0$, on passe à la valeur suivante de k, sinon on pose

$$\begin{aligned} &\alpha = -\|a\| \frac{a_{kk}}{|a_{kk}|} \qquad \text{si } a_{kk} \neq 0 \\ &\alpha = \|a\| \qquad \text{si } a_{kk} = 0 \end{aligned}$$

Le calcul des quantités

$$\begin{aligned} &\beta = \sqrt{2\|a\|(\|a\| + |a_{kk}|)} \\ &u_i = 0 \qquad \text{si } i = 1, ..., k-1 \\ &u_k = (a_{kk} - \alpha)/\beta \\ &u_i = a_{ik}/\beta \qquad \text{si } i = k+1, ..., n \end{aligned}$$

conduit à la matrice de Householder

$$H_k = I - 2u^t u$$

et à la forme triangulaire du système $Ax = b$

$$H_{n-1}.....H_2H_1Ax = H_{n-1}.....H_2H_1b$$

Exemple. Soit A la matrice

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix}$$

Si $k = 1$, on a

$$\begin{aligned} &\alpha = -\sqrt{3} \\ &\beta^2 = 2\sqrt{3}(1 + \sqrt{3}) \\ &u = ((1 + \sqrt{3})/\beta, 1/\beta, 1/\beta) \end{aligned}$$

d'où la matrice H_1

$$H_1 = \frac{-1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & -\frac{2+\sqrt{3}}{1+\sqrt{3}} & \frac{1}{1+\sqrt{3}} \\ 1 & \frac{1}{1+\sqrt{3}} & -\frac{2+\sqrt{3}}{1+\sqrt{3}} \end{pmatrix}$$

et la matrice $H_1 A$

$$H_1 A = \frac{-1}{\sqrt{3}} \begin{pmatrix} 3 & -1 & 1 \\ 0 & 2 & \frac{-2}{1+\sqrt{3}} \\ 0 & 2 & \frac{2(2+\sqrt{3})}{1+\sqrt{3}} \end{pmatrix}$$

si $k = 2$, les valeurs

$$\begin{aligned} &\alpha = \sqrt{2} \\ &\beta^2 = 2\sqrt{2}(1+\sqrt{2}) \\ &u = (0, -(1+\sqrt{2})/\beta, -1/\beta) \end{aligned}$$

conduisent à la matrice H_2

$$H_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

La matrice $H_2 H_1 A$ est triangulaire supérieure

$$H_2 H_1 A = \frac{-1}{\sqrt{6}} \begin{pmatrix} 3 & -1 & 1 \\ 0 & -4 & -2 \\ 0 & 0 & 2\sqrt{3} \end{pmatrix}$$

5.3 Méthodes itératives

Dans les méthodes itératives, le système $Ax = b$ est mis sous la forme $Mx = Nx + b$. Lorsque la matrice M est inversible, $x = M^{-1}Nx + M^{-1}b$. Remarquer que cette équation est une équation de la forme $x = f(x)$. Par conséquent, les méthodes itératives sont des méthodes de point fixe. La détermination du point fixe repose sur l'itération de l'équation

$$x_{k+1} = M^{-1}Nx_k + M^{-1}b$$

en notant x_k le vecteur de composantes $x_k = (x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)})$. L'algorithme est initialisé par un vecteur arbitraire $x_0 = (x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)})$ et s'arrête quand $\forall i \in \mathbb{N}$, $\left| x_i^{(k)} - x_i^{(k-1)} \right| < \varepsilon$ pour un ε donné. Lorsque la suite x_k converge, i.e. $\lim_{k \to \infty} x_k = x$, on dit que la méthode converge. On

démontre que la convergence de la méthode ne dépend pas du choix de x_0 et le résultat suivant : la méthode itérative $x_{k+1} = M^{-1}Nx_k + M^{-1}b$ converge si et seulement si le rayon spectral de la matrice $M^{-1}N$ est strictement inférieur à 1, $\rho(M^{-1}N) < 1$. Selon les choix des matrices M et N on a différentes méthodes itératives. On note D la matrice formée des seuls éléments diagonaux de A, $-E$ la matrice formée des a_{ij} si $i > j$ et $-F$ la matrice formée des a_{ij} si $i < j$, de sorte que $A = D - (E + F)$.

$$D = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}$$

$$-E = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n,1} & \cdots & a_{n,n-1} & 0 \end{pmatrix} \quad -F = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & 0 & 0 \end{pmatrix}$$

5.3.1 Méthode de Jacobi

Dans la méthode de Jacobi, encore appelée *méthode des déplacements simultanés*, la matrice A du système $Ax = b$ est décomposée en $A = M - N$. La matrice M correspond à la diagonale de A (et des zéros en dehors de la diagonale) $M = D = a_{ij}\delta_{ij}$ et la matrice N est la matrice A dans laquelle on a remplacé les éléments de la diagonale par des zéros $N = E + F$. La matrice $J = M^{-1}N = D^{-1}(E + F) = I - D^{-1}A$ est appelée matrice de Jacobi. À chaque pas, on calcule

$$x_i^{(k+1)} = (b_i - \sum_{j \neq i, j=1}^{n} a_{ij} x_j^{(k)})/a_{ii}$$

À chaque itération, on effectue $(n - 1)$ multiplications, n additions et une division. Pour stocker A et les vecteurs b, x_k et x_{k+1} on utilise $(n^2 + 3n)$ mémoires. La méthode ne converge pas toujours. On démontre que si A est une matrice définie positive, la méthode itérative converge. De même, si A est une matrice diagonalement dominante, c'est-à-dire si

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|$$

alors la méthode de Jacobi converge. Par conséquent, on peut avoir intérêt à réarranger les termes de A de façon à mettre A sous la forme d'une matrice dont les éléments diagonaux sont les plus grands possibles. On démontre que si A est une matrice tridiagonale par blocs, la méthode converge.

Exemple. Considérons le système

$$\begin{pmatrix} 4 & 2 & 1 \\ -1 & 2 & 0 \\ 2 & 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 9 \end{pmatrix}$$

mis sous la forme

$$\begin{cases} x = 1 - y/2 - z/4 \\ y = 1 + x/2 \\ z = 9/4 - x/2 - y/4 \end{cases}$$

Soit $x_0 = (0, 0, 0)$ le vecteur initial, en calculant les itérées on trouve

$$\begin{aligned} x_1 &= (1, 1, 9/4) \\ x_2 &= (-1/16, 3/2, 3/2) \\ x_3 &= (-1/8, -1/32, 61/32) \\ x_4 &= (5/128, 15/16, 265/128) \\ x_5 &= (7/512, 261/256, 511/256) \end{aligned}$$

La suite x_k converge vers la solution du système $(0, 1, 2)$.

5.3.2 Méthode de Gauss-Seidel

Dans la méthode de Gauss-Seidel, publiée en 1874 par Ludwig Seidel (1821-1896), on choisit $M = D - E$ et $N = F$, ce qui conduit à considérer la relation de récurrence

$$x_{k+1} = (D - E)^{-1} F x_k + (D - E)^{-1} b$$

C'est une amélioration de la méthode de Jacobi dans laquelle les valeurs calculées sont utilisées au fur et à mesure du calcul et non à l'issue d'une itération comme dans la méthode de Jacobi. On améliore ainsi la vitesse de convergence. Considérons un système à trois équations

$$\begin{cases} x = (b_1 - a_{12}y - a_{13}z)/a_{11} \\ y = (b_2 - a_{21}x - a_{23}z)/a_{22} \\ z = (b_3 - a_{31}x - a_{32}y)/a_{33} \end{cases}$$

À la première itération, on calcule à partir du vecteur initial

$$x_0 = (x^{(0)}, y^{(0)}, z^{(0)})$$

la valeur $x^{(1)}$

$$x^{(1)} = (b_1 - a_{12}y^{(0)} - a_{13}z^{(0)})/a_{11}$$

Cette valeur est réintroduite immédiatement dans le calcul de la deuxième composante (ce qui différencie cette méthode de la méthode de Jacobi, car on utilise ici la valeur $x^{(1)}$ et non $x^{(0)}$)

$$y^{(1)} = (b_2 - a_{21}x^{(1)} - a_{23}z^{(0)})/a_{22}$$

De même, on porte $x^{(1)}$ et $y^{(1)}$ dans le calcul de $z^{(1)}$

$$z^{(1)} = (b_3 - a_{31}x^{(1)} - a_{32}y^{(1)})/a_{33}$$

À chaque itération, on effectue $(n-1)$ multiplications, n additions et une division. Pour stocker A et les vecteurs b, x_k et x_{k+1}, on utilise (n^2+2n) mémoires. Si A et b sont calculés, on emploie n mémoires. La méthode ne converge pas toujours. On démontre que si A est une matrice définie positive, la méthode itérative converge. De même, si A est une matrice diagonalement dominante, c'est-à-dire si

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|$$

alors la méthode de Gauss-Seidel converge.

Exemple. Considérons le système

$$\begin{cases} x = 1 - y/2 - z/4 \\ y = 1 + x/2 \\ z = 9/4 - x/2 - y/4 \end{cases}$$

Partant du point $x_0 = (0,0,0)$, on calcule successivement

$$\begin{aligned} x_1 &= (1, 3/2, 11/8) \\ x_2 &= (-3/32, 61/64, 527/256) \\ x_3 &= (9/1024, 2047/2048, 16349/8192) \end{aligned}$$

Cet ensemble de points converge vers la solution exacte $(0,1,2)$. La méthode de Gauss-Seidel est aussi utilisée pour résoudre des systèmes non linéaires.

Exemple. Soit à résoudre le système

$$\begin{cases} x = \sin(xy) - y/2\pi \\ y = 2\pi x - (\pi - 1/4)(e^{2x-1} - 1) \end{cases}$$

Partant du point $(2/5, 3)$, on calcule successivement

$$\begin{aligned} x_1 &= (0.455, 3.03) \\ x_2 &= (0.499, 3.11) \\ x_3 &= (0.505, 3.14) \end{aligned}$$

qui converge vers la solution $x = 1/2$, $y = \pi$.

5.3.3 Méthodes de relaxation

La convergence d'une méthode itérative ne dépend pas du choix du vecteur initial x_0, mais la rapidité de convergence en dépend. D'où l'idée d'introduire un facteur de relaxation ω non nul. Les matrices M et N sont choisies comme dans la méthode de Gauss mais pondérées par le facteur de relaxation $M = (\frac{1}{\omega}D - E)$ et $N = \frac{1-\omega}{\omega}D + F$. La matrice

$$L = M^{-1}N = (\frac{1}{\omega}D - E)^{-1}(\frac{1-\omega}{\omega}D + F)$$

est appelée *matrice de relaxation*. L'algorithme est fondé sur le calcul des itérées

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij}x_j^{(k)})$$

On démontre que si le facteur de relaxation dépasse 2, la méthode diverge. Pour $\omega = 1$, on retrouve la méthode de Gauss-Seidel. Lorsque $0 < \omega < 1$, on parle de *sous-relaxation* et lorsque $1 < \omega < 2$, on parle de *surrelaxation* (*SOR, Successive Over Relaxation*). Le théorème d'Ostrowski-Reich affirme que si A est une matrice définie positive et si le facteur de relaxation $0 < \omega < 2$, alors la méthode converge. Lorsque A est une matrice tridiagonale par blocs dont les blocs diagonaux sont inversibles, si on note J la matrice $J = D^{-1}M = D^{-1}(E + F)$, et $\rho(J)$ son rayon spectral (c'est-à-dire le plus grand module des valeurs propres de J), alors la valeur optimale du facteur de relaxation est donnée par

$$\omega_0 = \frac{2}{1 + \sqrt{1 - \rho(J)^2}}$$

Dans certains cas, on utilise différents facteurs ω pour différents blocs de A : c'est la méthode de relaxation par blocs.

Exemple. Pour un système de trois équations à trois inconnues, l'itération conduit à calculer

$$\begin{cases} x^{(k+1)} = x^{(k)} + \omega\ (b_1 - a_{11}x_k - a_{12}y^{(k)} - a_{13}z^{(k)})/a_{11} \\ y^{(k+1)} = y^{(k)} + \omega\ (b_2 - a_{21}x^{(k+1)} - a_{22}y^{(k)} - a_{23}z^{(k)})/a_{22} \\ z^{(k+1)} = z^{(k)} + \omega\ (b_3 - a_{31}x^{(k+1)} - a_{32}y^{(k+1)} - a_{33}z^{(k)})/a_{33} \end{cases}$$

La *surrelaxation successive symétrique* (*SSOR, Symetric Successive Over Relaxation*) consiste à faire jouer le même rôle aux matrices E et F, en introduisant un vecteur intermédiaire y d'itérée $y^{(k)}$:

$$\begin{cases} (\frac{1}{\omega}D - E)y^{(k)} = (\frac{1-\omega}{\omega}D + F)x^{(k)} \\ (\frac{1}{\omega}D - F)x^{(k+1)} = (\frac{1-\omega}{\omega}D + E)y^{(k)} \end{cases}$$

5.3.4 Méthode d'Uzawa

La méthode d'Uzawa (1958) est un cas particulier des méthodes de relaxation. Soit A une matrice carré d'ordre n, symétrique définie positive, B une matrice $p \times n$ et $b \in \mathbb{R}^n$, $v \in \mathbb{R}^p$. On considère le problème

$$\min_{Bx=v} \frac{1}{2}(Ax, x) - (b, x)$$

On démontre que x est un minimum de cette équation si et seulement si il existe un réel $y \in \mathbb{R}^p$ vérifiant

$$\begin{pmatrix} A & B^t \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b \\ v \end{pmatrix}$$

L'algorithme d'Uzawa de paramètre ω consiste alors à choisir une condition initiale $x^{(0)}$, $y^{(0)}$ et à itérer, pour une précision donnée $\epsilon > 0$, tant que $\left\|Bx^{(k)} - b\right\| > \epsilon$, le calcul des quantités

$$\begin{cases} Ax^{(k+1)} = b - B^t y^{(k+1)} \\ y^{(k+1)} = y^{(k)} + \omega(Bx^{(k)} - v) \end{cases}$$

Ces équations s'écrivent sous la forme

$$y^{(k+1)} = (1 - \omega BA^{-1}B^t)y^{(k)} + \omega(BA^{-1}b - v)$$

qui correspond au cas des méthodes de relaxation avec $J = 1 - \omega BA^{-1}B^t$. Si on note $\lambda_1, \lambda_2, .., \lambda_p$, les valeurs propres de $BA^{-1}B^t$,

$$\rho(J) < 1 \quad \Leftrightarrow \quad \forall i, \quad |1 - \omega\lambda_i| < 1$$

Par conséquent, comme $BA^{-1}B^t$ est une matrice symétrique, définie positive, ses valeurs propres sont strictement positives. Il faut donc choisir

$$0 < \omega < \frac{2}{\max \lambda_i}$$

pour que la méthode d'Uzawa converge.

5.4 Méthodes projectives

Les méthodes de projection partent du constat que pour résoudre l'équation matricielle $Ax = b$ il suffit de déterminer le minimum de la forme quadratique

$$J(x) = \frac{1}{2}(x, Ax) - (b, x) = \frac{1}{2}x^t Ax - b^t x$$

La dérivée de cette forme quadratique

$$J'(x) = \frac{1}{2}A^t x + \frac{1}{2}Ax - b$$

se réduit à $J'(x) = Ax - b$ lorsque A est symétrique. Par conséquent, lorsque A est *symétrique, définie positive*, $J(x)$ a pour minimum $Ax = b$. À chaque pas, on détermine une valeur α_k qui minimise la quantité $J(x_k + \alpha_k r_k)$. Le succès de la méthode du gradient conjugué a incité de nombreux auteurs à proposer des méthodes plus générales dans le cas où A n'est pas une matrice symétrique, définie positive. La méthode la plus simple, lorsque A n'est pas symétrique, consiste à remplacer l'équation $Ax = b$ par $A^tAx = A^tb$ dans laquelle A^tA est symétrique, mais cette méthode a l'inconvénient d'effectuer des produits supplémentaires et d'amplifier le mauvais conditionnement éventuel de A puisque $cond(A^tA) = cond(A)^2$. D'autres méthodes plus efficaces ont été proposées pour A non symétrique comme la méthode *CGS* (*Conjugate Gradient Square*), la méthode *BiCGStab* (*Bi-Conjugate Gradient Stabilized*) ou la méthode *GMRES* (*Generalized Minimum Residual Method*).

5.4.1 Méthode de la plus profonde descente

La *méthode de la plus profonde descente* cherche à minimiser le résidu r_k. Elle se fonde sur l'algorithme suivant : On se donne un vecteur x_0, puis on calcule successivement les quantités

$$\begin{cases} r_k = b - Ax_k \\ \alpha_k = \dfrac{(r_k, r_k)}{(r_k, Ar_k)} \\ x_{k+1} = x_k + \alpha_k r_k \end{cases}$$

où (r_k, r_k) est le produit scalaire de r_k par lui-même. L'efficacité de la méthode dépend du conditionnement de la matrice A. Notons

$$\mathcal{E}(x_k) =< A(x_k - x), (x_k - x) >^{1/2}$$

la norme "*énergétique*". Soit A une matrice symétrique définie positive, et $\kappa = \lambda_{\max}/\lambda_{\min}$ le conditionnement de la matrice A, κ est le rapport de la plus grande valeur propre de A sur la plus petite. La convergence de la méthode de la plus profonde descente est donnée par

$$\mathcal{E}(x_k) \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \mathcal{E}(x_0)$$

La méthode n'est pas toujours efficace : si la matrice A a un grand conditionnement, on voit dans l'expression précédente que si κ est élevé, la norme énergétique n'évolue presque pas. Par conséquent, le vecteur résiduel r_k ne change pas beaucoup d'une itération à l'autre : la convergence est très lente. Pour éviter ce problème, Fox, Husky et Wilkinson ont proposé en 1949 de remplacer la minimisation le long du vecteur résiduel par une minimisation le long de la direction orthogonale : c'est la méthode des directions conjuguées. Hestenes et Stiefel ont montré qu'on pouvait choisir ces directions à chaque pas : c'est la méthode du gradient conjugué.

5.4.2 Méthode du gradient conjugué

La *méthode du gradient conjugué* est une amélioration de la méthode de la plus profonde descente, dans laquelle le calcul de $x_{k+1} = x_k + \alpha_{k+1} p_k$ se fait le long de nouvelles directions $(p_0, ..., p_k)$. On suppose que A est une matrice symétrique, définie positive. Les trois premières équations de l'algorithme correspondent à la minimisation de J sur l'espace $x_0 + Vect(p_0, ..., p_k)$, où x_0 est un point arbitraire choisi comme point initial de l'algorithme. Les deux dernières équations correspondent au calcul de la nouvelle direction. Elle se fonde sur l'algorithme suivant. On se donne un point x_0 de $\mathbb{R}^n$, $r_0 = b - Ax_0$ et $p_0 = r_0$ et on calcule les quantités suivantes

$$\begin{cases} \alpha_{k+1} = \dfrac{(r_k, r_k)}{(p_k, Ap_k)} \\ x_{k+1} = x_k + \alpha_{k+1} p_k \\ r_{k+1} = r_k - \alpha_{k+1} A p_k \\ \beta_{k+1} = \dfrac{(r_{k+1}, r_{k+1})}{(r_k, r_k)} \\ p_{k+1} = r_{k+1} + \beta_{k+1} p_k \end{cases}$$

On démontre qu'il existe un polynôme de degré k noté P_k vérifiant $x_k = P_k(A)x_0$ et $P_k(0) = 1$. Le polynôme qui minimise l'expression

$$\mathcal{E}^2(x_k) \leq \min_{P_k} \max_{\lambda} P_k(\lambda)^2 \sum_j a_j^2 \lambda_j$$

est donné par

$$P_k(\lambda) = \frac{T_k\left(\dfrac{\lambda_{\max} + \lambda_{\min} - 2\lambda}{\lambda_{\max} - \lambda_{\min}}\right)}{T_k\left(\dfrac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right)}$$

où T_k est le polynôme de Tchebychev. Si $\kappa = \lambda_{\max}/\lambda_{\min}$ est le conditionnement de la matrice A, on a

$$\mathcal{E}(x_k) \leq 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \mathcal{E}(x_0)$$

La méthode du gradient conjugué nécessite $n^3 + 5n^2 - 3n$ additions, $n^3 + 6n^2$ multiplications et $2n$ divisions. Si la matrice A est mal conditionnée, la convergence de l'algorithme du gradient conjugué est lente. Dans ce cas, on cherchera à améliorer la vitesse de convergence : c'est la méthode du gradient conjugué préconditionné.

5.4.3 Méthode du gradient conjugué préconditionné

Si les valeurs propres de la matrice A sont dispersées, il faut procéder à un préconditionnement et remplacer le système $Ax = b$ par $MAx = Mb$ où M

est une matrice symétrique définie positive telle que $cond(MA) \ll cond(A)$. On modifie l'algorithme précédent par les formules

$$\begin{cases} r_0 = b - Ax_0 \\ p_0 = Mr_0 \\ \alpha_k = \dfrac{(r_k, Mr_k)}{(p_k, Ap_k)} \\ x_{k+1} = x_k + \alpha_k p_k \\ r_{k+1} = r_k - \alpha_k Ap_k \\ \beta_{k+1} = \dfrac{(r_{k+1}, Mr_{k+1})}{(r_k, Mr_k)} \\ p_{k+1} = Mr_{k+1} + \beta_{k+1} p_k \end{cases}$$

5.4.4 Méthode du gradient conjugué pour les moindres carrés

La méthode du gradient conjugué pour les moindres carrés est une méthode issue des recherches d'adaptation de l'algorithme conjugué lorsque la matrice A n'est pas symétrique. Elle s'appuie sur la remarque suivante : Si A est une matrice carrée, inversible, les solutions de $A^tAx = A^tb$ sont les points critiques de $\|Ax - b\|^2$. Le problème revient alors à minimiser $\min_x \|Ax - b\|^2$. L'algorithme est le suivant : On choisit x_0, et on pose

$$s_0 = b - Ax_0, r_0 = p_0 = A^t(b - Ax_0) = A^t s_0$$

et $q_0 = Ap_0$. Pour $k = 0, 1, 2, ...$ on calcule successivement les quantités

$$\begin{cases} \alpha_{k+1} = \dfrac{(r_k, r_k)}{(q_k, q_k)} \\ x_{k+1} = x_k + \alpha_{k+1} p_k \\ s_{k+1} = s_k - \alpha_{k+1} q_k \\ r_{k+1} = A^t s_{k+1} \\ \beta_{k+1} = \dfrac{(r_{k+1}, r_{k+1})}{(r_k, r_k)} \\ p_{k+1} = r_{k+1} + \beta_{k+1} p_k \\ q_{k+1} = Ap_{k+1} \end{cases}$$

Si la matrice A est mal conditionnée, on procédera à un *préconditionnement.*

5.4.5 Méthode du gradient biconjugué

La méthode du gradient biconjugué s'applique à une matrice non nécessairement symétrique. L'algorithme repose sur un double traitement de l'équation $Ax = b$ et de l'équation $A^t\widetilde{x} = \widetilde{b}$, sous la forme du système

$$\begin{pmatrix} A & 0 \\ 0 & A^t \end{pmatrix} \begin{pmatrix} x \\ \widetilde{x} \end{pmatrix} = \begin{pmatrix} b \\ \widetilde{b} \end{pmatrix}$$

On choisit x_0, $\widetilde{b}$ et $\widetilde{x}_0$, et on pose

$$\begin{aligned} r_0 &= b - Ax_0, p_0 = r_0 \\ \widetilde{r}_0 &= \widetilde{b} - A\widetilde{x}_0, \widetilde{p}_0 = \widetilde{r}_0 \end{aligned}$$

Pour $k = 0, 1, 2, ...$ on calcule successivement les quantités

$$\begin{cases} \alpha_{k+1} = \dfrac{(\widetilde{r}_k, r_k)}{(\widetilde{p}_k, Ap_k)} \\ x_{k+1} = x_k + \alpha_{k+1} p_k \\ r_{k+1} = r_k - \alpha_{k+1} Ap_k \\ \widetilde{r}_{k+1} = \widetilde{r}_k - \alpha_{k+1} A^t p_k \\ \beta_{k+1} = \dfrac{(\widetilde{r}_{k+1}, r_{k+1})}{(\widetilde{r}_k, r_k)} \\ p_{k+1} = r_{k+1} + \beta_{k+1} p_k \\ \widetilde{p}_{k+1} = \widetilde{r}_{k+1} + \beta_{k+1} \widetilde{p}_k \end{cases}$$

Les résidus et les directions de descente vérifient les relations d'orthogonalité

$$(r_k, \widetilde{r}_k) = (Ap_{k+1}, \widetilde{p}_k) = 0$$

5.4.6 Méthode d'Arnoldi

La méthode d'Arnoldi est une méthode de projection orthogonale sur un sous-espace de Krylov permettant de construire, pour toute matrice A, une base orthonormée de ce sous-espace et une matrice apparentée à une matrice de Hessenberg. Soit A une matrice carrée d'ordre n, un degré $m \in \mathbb{N}^*$ et v un vecteur de $\mathbb{R}^n$, l'espace

$$\mathcal{K}_m(A, v) = vect\{v, Av, A^2v, ..., A^{m-1}v\}$$

est appelé espace de Krylov. Cet espace est donc celui des images des polynômes de A de degré inférieur ou égal à $m-1$ appliqués à v. À partir d'un vecteur normé $v_1 = v/\|v\|$, la méthode d'Arnoldi construit de proche en proche les vecteurs v_j, pour $j = 1, ..., m$ par multiplications sucessives et orthonormalisation du vecteur v_k par rapport à tous les vecteurs v_j déjà calculés. L'orthonormalisation est exécutée par une procédure de Gram-Schmidt et en même temps les éléments apparentés à la matrice de Hessenberg sont calculés. Pour résoudre $Ax = b$, on choisit un vecteur x_0 arbitrairement et on pose $v = b - Ax_0$. On note V_k la matrice formée des vecteurs colonnes $(v_1, v_2, ..., v_k)$, $\widetilde{H}_k$ la matrice carrée constituée des k premières lignes de la matrice de Hessenberg supérieure H_k, complétée d'une ligne dont le seul élément non nul est $h_{k+1,k}$ et e_k le k-ième vecteur de la

base canonique de $\mathbb{R}^k$. La matrice $\widetilde{H}_k$ est de la forme

$$\widetilde{H}_k = \begin{pmatrix} h_{1,1} & h_{1,2} & ... & h_{1,k-1} & h_{1,k} \\ h_{2,1} & h_{2,2} & ... & h_{2,k-1} & h_{2,k} \\ 0 & \ddots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & h_{k,k-1} & h_{k,k} \\ 0 & 0 & ... & 0 & h_{k+1,k} \end{pmatrix}$$

L'algorithme d'Arnoldi conduit à la relation

$$\begin{aligned} AV_k &= V_{k+1}\widetilde{H}_{k-1} \\ &= V_k H_k + h_{k+1,k} v_k e_k^t \end{aligned}$$

D'où on déduit la relation

$$V_k^t A V_k = H_k$$

Partant du vecteur $v_1 = v/\|v\|$, l'algorithme d'Arnoldi calcule pour $k = 1, ..., m-1$, les quantités

$$\begin{cases} w = Av_k \\ \text{Pour } i = 1 \text{ à } k, \text{ Faire } h_{i,k} = v_i^t w,\ w = w - h_{i,k} v_i \\ h_{k+1,k} = \|w\| \\ \text{Si } h_{k+1,k} = 0 \text{ alors Stop Sinon } v_{k+1} = w/h_{k+1,k} \end{cases}$$

La solution de l'équation $Ax = b$ est obtenue en déterminant le vecteur y_m solution du système

$$H_m y_m = \beta e_1$$

où β est la norme du résidu initial $\beta = \|r_0\| = \|b - Ax_0\|$ et $e_1 = (1, 0, ..., 0)$. La solution approchée x_m est alors

$$x_m = x_0 + V_m y_m$$

L'espace mémoire utilisé est de l'ordre de mn, ce qui correspond au stockage de m vecteurs de taille n. Pour réduire la taille de l'espace de Krylov $\mathcal{K}_m(A, v)$, on procède à des redémarrages. Ceci correspond à appliquer la méthode d'Arnoldi pour un nombre m de pas, puis à utiliser la solution obtenue comme vecteur initial d'un nouvel ensemble de m pas.

5.4.7 Méthode GMRES

La méthode *GMRES* (*Generalized Minimum Residual Method*) diffère peu de la méthode d'Arnoldi. Dans la méthode d'Arnoldi, la solution choisie dépend d'un vecteur y_m, qui est construit de sorte que le résidu $V_m y_m$ reste orthogonal à tous les vecteurs de l'espace de Krylov $\mathcal{K}_m(A, r_0)$. Au contraire, la méthode GMRES cherche à construire un vecteur y_m qui minimise le résidu $V_m y_m$. On remplace donc la résolution du système $H_m y_m = \beta e_1$ par une procédure qui détermine le vecteur y_m qui minimise

$$\left\| \beta e_1 - \widetilde{H}_m y_m \right\|$$

Une solution pour ce problème consiste à transformer la matrice $\widetilde{H}_m$ en une matrice triangulaire supérieure à l'aide des rotations de Givens. Ces rotations correspondent aux matrices

$$G_i = \begin{pmatrix} 1 & 0 & \dots & & \dots & & \dots & 0 \\ 0 & \ddots & & & & & & 0 \\ 0 & & 1 & & & & & \\ & & & c_i & s_i & & & \vdots \\ \vdots & & & -s_i & c_i & & & \\ & & & & & 1 & & 0 \\ & & & & & & \ddots & 0 \\ 0 & \dots & & & & \dots & 0 & 1 \end{pmatrix}$$

Les valeurs c_i et s_i sont choisies de façon à éliminer l'élément $h_{i+1,i}$. Par exemple, en choisissant

$$c_1 = \frac{h_{1,1}}{\sqrt{h_{11}^2 + h_{21}^2}} \text{ et } s_1 = \frac{h_{2,1}}{\sqrt{h_{11}^2 + h_{21}^2}}$$

le produit $G_1 \widetilde{H}_m$ conduit à une matrice dont l'élément $h_{2,1}$ a été éliminé. En répétant ce procédé, on construit une matrice triangulaire supérieure

$$G_m G_{m-1} \dots G_1 \widetilde{H}_m = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1,m-1} & h_{1,m} \\ 0 & h_{22} & & & h_{2,m} \\ \vdots & & \ddots & & \vdots \\ 0 & & & h_{m-1,m-1} & h_{m-1,m} \\ 0 & 0 & \dots & 0 & h_{m,m} \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

En posant $B_m = G_m G_{m-1} \dots G_1$, le vecteur y_m cherché est solution de l'équation

$$\left\| B_m \beta e_1 - B_m \widetilde{H}_m y_m \right\| = 0$$

Ce système se résout par la méthode de remontée.

5.5 Exercices

1. On considère la matrice

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 20 & 26 \\ 3 & 26 & 70 \end{pmatrix}$$

Écrire la décomposition LU de cette matrice et résoudre le système linéaire $Ax = b$ selon la méthode de Cholesky. On prendra $b = (7, 50, 102)$.

2. *Méthode du gradient conjugué.* Soit A la matrice carrée à coefficients réels suivante

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$

et b le vecteur $(1, 0)$.
1) Calculer le nombre de conditionnement $\kappa(A)$ de la matrice A
2) Appliquer la méthode du gradient conjugué pour résoudre le système

$$Ax = b$$

On choisira l'origine $(0, 0)$ comme point de départ de l'algorithme.

3. *Méthode de la plus profonde descente.* Soit A une matrice symétrique définie positive d'ordre n de valeurs propres $0 < \lambda_1 \leq ... \leq \lambda_n$ et b un vecteur de composantes réelles de dimension n. La résolution du système linéaire $Ax = b$ est approchée par la *méthode de la plus profonde descente.*

$$\begin{cases} r_i = b - Ax_i \\ \alpha_i = \dfrac{r_i^t \, r_i}{r_i^t A r_i} \\ x_{i+1} = x_i + \alpha_i (b - Ax_i) \end{cases}$$

x_i désigne la i-ième itérée de x_0 et r_i est le résidu de l'itération. On note $e_i = x_i - x$ l'erreur commise à chaque pas et on définit la norme d'un vecteur u par

$$\|u\| = \sqrt{u^t . Au}$$

1) Montrer que la norme de l'erreur d'indice $i+1$ est proportionnelle à la norme d'erreur d'indice i, c'est-à-dire que l'on a la relation

$$\|e_{i+1}\| = w \, \|e_i\|$$

et que w vérifie

$$w^2 = 1 - \frac{({}^t r_i r_i)^2}{({}^t r_i A r_i)({}^t e_i A e_i)}$$

2) Soit p_i un système de vecteurs propres orthonormés de la matrice A associés aux valeurs propres $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$. Montrer que w^2 s'écrit

$$w^2 = 1 - \frac{(\sum_{j=1}^{n} v_j^2 \lambda_j^2)^2}{(\sum_{j=1}^{n} v_j^2 \lambda_j^3)(\sum_{j=1}^{n} v_j^2 \lambda_j)}$$

où les coefficients v_j désignent les composantes de l'erreur e_i dans la base des vecteurs propres p_j.

$$e_i = \sum_{j=1}^{n} v_j p_j$$

3) Soit $\kappa = cond_2(A) = \lambda_n/\lambda_1$ le conditionnement de la matrice A pour la 2-norme, montrer que w est majoré par

$$w \leq \frac{\kappa - 1}{\kappa + 1}$$

4) En déduire l'inégalité

$$\|e_k\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|e_0\|$$

Que peut-on dire du cas $\kappa = 1$?

4. *Accélération de convergence pour une méthode itérative.* On se propose d'étudier une méthode d'accélération de la convergence qui utilise les polynômes de Tchebychev. On désigne par A une matrice symétrique définie positive d'ordre n de valeurs propres $0 < \lambda_1 \leq ... \leq \lambda_n$. La résolution du système linéaire $Ax = b$ est approchée par une méthode itérative de la forme

$$x_{k+1} = x_k + \alpha_k(b - Ax_k)$$

x_k désigne la k-ième itérée de x_0 et α_k est une suite de nombres strictement positifs à choisir de manière optimale.

1) En exprimant l'erreur de la k-ième itération $e_k = x_k - x$, montrer qu'il existe un polynôme q_k d'ordre n à coefficients réels tel que $e_k = q_k(A)e_0$. Préciser les racines de ce polynôme.

2) Sachant que la 2-norme vérifie $\|A\|_2 = \rho(A)$ pour une matrice symétrique, démontrer l'inégalité

$$\|q_k(A)\|_2 \leq \max_{\lambda_1 \leq t \leq \lambda_n} |q_k(t)|$$

3) Sachant qu'une méthode itérative quelconque $x_{k+1} = Bx_k + c$ converge si et seulement si $\rho(B) < 1$, montrer que, dans le cas où les nombres α_k sont indépendants de l'indice k, ($\alpha_k = \alpha$), la méthode itérative converge si et seulement si

$$0 < \alpha < \frac{2}{\lambda_n}$$

Déterminer la valeur optimale de α.

4) On définit les polynômes de Tchebychev par

$$T_k(x) = \cos(n \arccos x) \qquad \text{si } |x| \leq 1$$

$$T_k(x) = \frac{1}{2}\{\left(x - \sqrt{x^2-1}\right)^k + \left(x + \sqrt{x^2-1}\right)^k\} \qquad \text{si } |x| > 1$$

Montrer que T_k est un polynôme de degré k vérifiant les relations de récurrence, pour tout x réel :

$$\begin{cases} T_0(x) = 1, & T_1(x) = x \\ T_{k+2}(x) = 2xT_{k+1}(x) - T_k(x) \end{cases}$$

Vérifier que le polynôme $T_k(x)$ admet dans l'intervalle $[-1, 1]$, $(k+1)$ extremums aux points $x_i = cos(\frac{i\pi}{k})$ $0 \leq i \leq k$ pour lesquels il prend alternativement les valeurs $+1$ et -1.

5) On note $\mathcal{P}_k$ l'ensemble des polynômes de degré inférieur ou égal à k et on désigne par a un réel quelconque vérifiant $|a| > 1$. Démontrer que le polynôme de Tchebychev T_k est solution du problème suivant : trouver un polynôme $p(x)$ de $\mathcal{P}_k$ tel que

$$\max_{-1 \leq x \leq 1} |p(x)| \leq \inf_{q \in E_k} \max_{-1 \leq x \leq 1} |q(x)|$$

où E_k est l'ensemble des polynômes de $\mathcal{P}_k$ qui coïncident avec le polynôme de Tchebychev au point a

$$E_k = \{q \in \mathcal{P}_k : \quad q(a) = T_k(a)\}$$

6) Montrer que la solution de ce problème de minimisation est unique.

7) On désigne par u et v deux réels tels que $0 < u \leq v < 1$. Montrer que le problème : trouver un polynôme $p \in \mathcal{P}_k$ tel que

$$\max_{u \leq x \leq v} |p(x)| \leq \inf_{q \in B_k} \max_{u \leq x \leq v} |q(x)|$$

et

$$B_k = \{q \in \mathcal{P}_k : \quad q(0) = 1\}$$

a une solution et une seule que l'on explicitera.

8) Déterminer, pour k fixé, les nombres α_j qui minimisent la 2-norme de $q_k(A)$.

9) On définit la vitesse asymptotique de convergence par la quantité

$$\lim_{k\to\infty} \|q_k(A)\|_2^{1/k}$$

Comparer cette quantité pour la méthode de convergence à α constant de la question 3 avec celle de la méthode optimale définie à la question 8.

6
Valeurs et vecteurs propres

Après la présentation des algorithmes de résolution des systèmes linéaires, nous étudions dans ce chapitre les problèmes liés au calcul des valeurs et des vecteurs propres d'une matrice ainsi que les techniques de calcul du polynôme caractéristique.

6.1 Méthode des puissances

La *méthode des puissances*, encore appelée *méthode de la puissance itérée,* repose sur l'idée qu'en appliquant un grand nombre de fois la matrice sur un vecteur quelconque, les vecteurs successifs obtenus prennent une direction qui se rapproche de la direction du vecteur propre associé à la plus grande valeur propre en valeur absolue. Supposons que la matrice A possède n valeurs propres simples distinctes et qu'il n'y en ait qu'une de module maximum. Notons $|\lambda_1| > |\lambda_2| \geq, ..., \geq |\lambda_n|$ les valeurs propres supposées rangées par ordre décroissant et $v_1, v_2, ..., v_n$ les vecteurs propres associés. L'algorithme consiste à calculer la suite des itérées $y^{(k+1)} = Ay^{(k)} / \left\|Ay^{(k)}\right\|$. Pour cela, on se donne $x^{(0)} = (x_1^{(0)}, ..., x_n^{(0)})$ un vecteur arbitraire et on pose $y^{(0)} = x^{(0)}$. À la k-ième étape, on calcule le vecteur $x^{(k)} = (x_1^{(k)}, ..., x_n^{(k)})$

$$x^{(k)} = Ay^{(k-1)}$$

puis le vecteur

$$y^{(k)} = \left(\frac{x_1^{(k)}}{x_p^{(k)}}, ..., \frac{x_n^{(k)}}{x_p^{(k)}} \right)$$

où $x_p^{(k)}$ est la composante de plus grand module du vecteur $x^{(k)}$ telle que

$$\left| x_p^{(k)} \right| = \sup_i \left| x_i^{(k)} \right|$$

La p-ième composante de $y^{(k)}$ vaut alors 1. La valeur propre estimée à la k-ième itération est la p-ième composante du vecteur $x^{(k)}$.

$$\lambda_1^{(k)} = x_p^{(k)}$$

L'itération s'arrête dès que la différence entre deux estimations de la valeur propre est suffisamment petite

$$\left| \lambda_1^{(k)} - \lambda_1^{(k-1)} \right| \leq \varepsilon$$

Le vecteur $y^{(k)}$ converge vers un vecteur propre v_1 associé à la valeur propre λ_1.

Exemple. Soit A une matrice de valeurs propres $\lambda_1 = 3$, $\lambda_2 = 2$, $\lambda_3 = 1$ et de vecteurs propres associés $v_1 = (0, 0, 1)$, $v_2 = (1, 1, 1)$ et $v_3 = (1, 0, 0)$.

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 3 \end{pmatrix}$$

et $x_0 = (0, 1, 0)$ un vecteur arbitraire. Calculons $x_1 = Ax_0 = (1, 2, -1)$. Dans ce cas $p = 2$ et $y_1 = (1/2, 1, -1/2)$. Le calcul de $x_2 = Ay_1 = (3/2, 2, -5/2)$ donne une estimation de la valeur propre $\lambda_1^{(1)} = 2$. Comme $y_2 = (-3/5, -4/5, 1)$, la valeur de $p = 3$ conduit à $x_3 = Ay_2$ soit $x_3 = (-7/5, -8/5, 19/5)$ et à une estimation de la valeur propre égale à $\lambda_1^{(2)} = 19/5$. L'algorithme se poursuit. On calcule successivement les quantités $y_3 = (-7/19, -8/19, 1)$, $x_4 = Ay_4 = (-15/19, -16/19, 65/19)$ d'où $p = 3$ et $\lambda_1^{(3)} = 65/19$. Ensuite $y_4 = (-15/65, -16/65, 1)$ permet le calcul $x_5 = (-31/65, -352/65, 211/65)$, d'où $\lambda_1^{(4)} = 211/65$ puis $y_5 = (-31/211, -32/211, 1)$, qui permet le calcul de $x_6 = Ay_5$ soit encore $x_6 = (-63/211, -64/211, 665/211)$ d'où $\lambda_1^{(5)} = 665/211$. La suite $\lambda_1^{(k)}$ converge vers $\lambda_1 = 3$ et y_k converge vers le vecteur propre $v_1 = (0, 0, 1)$. 3 est la plus grande valeur propre de A et $(0, 0, 1)$ son vecteur propre associé.

Remarquons que si on choisit un vecteur propre de la matrice A comme vecteur initial x_0 de la méthode, on risque de ne pas avoir la plus grande

des valeurs propres (prendre par exemple le vecteur $(1, 0, 0)$ dans l'exemple précédent). La précision de la méthode est mal contrôlée : on ne peut obtenir une estimation de la valeur propre à ε près lorsque l'itération s'arrête. Enfin, notons que lorsque l'algorithme est appliqué à la matrice A^{-1}, il détermine la plus petite valeur propre de A : c'est la *méthode des itérées inverses*, encore appelée *méthode de la puissance itérée inverse*.

6.2 Déflation de Wielandt

La *méthode de la déflation de Wielandt* permet le calcul des autres valeurs propres. Ayant obtenu la valeur propre λ_1 et un vecteur propre associé, on construit une matrice A_1 admettant comme valeurs propres $0, \lambda_2, ..., \lambda_n$ et comme vecteurs propres $v_1, v_2, ..., v_n$. En appliquant la méthode des puissances à la matrice A_1, on obtient la valeur propre de A de plus grand module après λ_1 et un vecteur propre associé. On procède de la façon suivante : On cherche un vecteur propre w_1 de la matrice transposée A^t associé à la valeur propre λ_1 en résolvant le système

$$(A^t - \lambda_1 I)w_1 = 0$$

et on calcule la matrice A_1 par la formule

$$A_1 = A - \lambda_1 \frac{v_1\, w_1^t}{w_1^t v_1}$$

Exemple. Poursuivons l'exemple précédent. L'équation $\det(A^t - \lambda I) = 0$ conduit au vecteur propre $w_1 = (0, -1, 1)$ d'où

$$v_1\, w_1^t = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

et la matrice A_1

$$A_1 = A - 3\frac{v_1\, w_1^t}{w_1^t v_1} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & -4 & 0 \end{pmatrix}$$

6.3 Méthode de Jacobi

La *méthode de Jacobi* est une méthode itérative applicable à une matrice A symétrique. Elle consiste à faire opérer le groupe des rotations planes sur A, c'est-à-dire à multiplier A par des transformations orthogonales afin de la mettre sous forme diagonale, les éléments diagonaux étant les valeurs

propres de la matrice A. Étudions le principe de la méthode. Considérons la matrice H dont les éléments sont égaux à ceux de la matrice identité sauf pour les quatre valeurs suivantes $h_{pp} = \cos(\alpha)$, $h_{pq} = \sin(\alpha)$, $h_{qq} = \cos(\alpha)$ et $h_{qp} = -\sin(\alpha)$, avec $p < q$. La matrice H est une matrice orthogonale $H^t H = I$. À la première étape, on calcule la matrice $A_1 = H_1^{-1} A H = H^t A H$, en remarquant que seules les lignes et les colonnes p et q sont modifiées, pour $j = p$ ou q, on a :

$$\begin{cases} a_{pj}^{(1)} = a_{jp}^{(1)} = a_{jp}\cos(\alpha) - a_{jq}\sin(\alpha) \\ a_{qj}^{(1)} = a_{jq}^{(1)} = a_{jp}\cos(\alpha) + a_{jq}\sin(\alpha) \\ a_{pp}^{(1)} = a_{pp}\cos^2(\alpha) + a_{qq}\sin^2(\alpha) - 2a_{pq}\sin(\alpha)\cos(\alpha) \\ a_{qq}^{(1)} = a_{qq}\cos^2(\alpha) + a_{pp}\sin^2(\alpha) + 2a_{pq}\sin(\alpha)\cos(\alpha) \\ a_{pq}^{(1)} = a_{qp}^{(1)} = a_{pq}(\cos^2(\alpha) - \sin^2(\alpha)) + (a_{pp} - a_{qq})\sin(\alpha)\cos(\alpha) \end{cases}$$

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

$$tg(2\alpha) = \frac{2a_{pq}}{a_{pp} - a_{qq}} \equiv \theta \qquad \text{avec } |\alpha| \leq \frac{\pi}{4}$$

ou encore

$$\cos(\alpha) = \sqrt{\tfrac{1}{2}(1 + \tfrac{1}{\sqrt{1+\theta^2}})}$$
$$\sin(\alpha) = sgn(\theta)\sqrt{\tfrac{1}{2}(1 - \tfrac{1}{\sqrt{1+\theta^2}})}$$

Si $a_{pp} = a_{qq}$, on choisira

$$\cos(\alpha) = \tfrac{1}{\sqrt{2}}$$
$$\sin(\alpha) = \frac{sgn(a_{pq})}{\sqrt{2}}$$

On a alors

$$(a_{pp}^{(1)})^2 + (a_{qq}^{(1)})^2 = a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2$$

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

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

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

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

La suite des matrices A_k converge vers une matrice diagonale dont les éléments diagonaux sont les valeurs propres de la matrice initiale A. La suite des matrices $P_k = H_1 H_2 ... H_k$ converge vers la matrice dont les colonnes

sont constituées de vecteurs propres. Au cours des itérations un terme peut redevenir nul, mais on démontre que

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

On arrête l'itération quand

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

En pratique, on a le choix à chaque pas d'itération du couple (p, q). On définit différentes stratégies. Dans la *méthode de Jacobi classique*, on choisit (p, q) tels que

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

Dans la *méthode de Jacobi cyclique*, on effectue un balayage systématique en prenant pour (p, q) les couples $(1, 2)$, $(1, 3)$, ...,$(1, n)$ puis $(2, 3)$,...,$(2, n)$, etc., jusqu'à $(n - 1, n)$. Dans la *méthode de Jacobi cyclique avec seuil*, on effectue comme précédemment un balayage sur les éléments triangulaires supérieurs, chaque élément a_{ij} étant pris comme élément à annuler a_{pq}, mais on ne retient le couple (p, q) que si $|a_{ij}|$ est supérieur à un certain seuil qui peut être réajusté à chaque itération. La méthode de Jacobi est stable, mais sa convergence est lente, ce qui en fait une méthode très peu utilisée.

6.4 Méthode de Givens-Householder

Proposée en 1958, la *méthode de Givens-Householder* est la réunion de deux algorithmes. La méthode de Householder met la matrice initiale A sous la forme tridiagonale symétrique (cet algorithme a été étudié dans le chapitre précédent). L'algorithme de Givens calcule les valeurs propres d'une matrice tridiagonale symétrique. Supposons que A soit mis sous la forme

$$B = \begin{pmatrix} b_1 & c_1 & 0 & \cdots & 0 \\ c_1 & b_2 & c_2 & \ddots & \vdots \\ 0 & c_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & \cdots & 0 & c_{n-1} & b_n \end{pmatrix}$$

et notons B_k la sous-matrice

$$B_k = \begin{pmatrix} b_1 & c_1 & 0 & \cdots & 0 \\ c_1 & b_2 & c_2 & \ddots & \vdots \\ 0 & c_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{k-1} \\ 0 & \cdots & 0 & c_{k-1} & b_k \end{pmatrix}$$

Les polynômes caractéristiques des matrices B_k vérifient les relations de récurrence

$$\begin{array}{l} p_0(\lambda) = 1 \quad p_1(\lambda) = b_1 - \lambda \\ p_k(\lambda) = (b_k - \lambda)p_{k-1}(\lambda) - c_{k-1}^2 p_{k-2}(\lambda) \quad \text{pour } k = 2, ..., n \end{array}$$

Ils vérifient les propriétés suivantes

$$\lim_{\lambda \to -\infty} p_k(\lambda) = +\infty$$

Si $p_k(\lambda_0) = 0$, alors $p_{k-1}(\lambda_0)p_{k+1}(\lambda_0) < 0$ pour $k = 1, ..., n-1$. Le polynôme p_k a k racines réelles distinctes qui séparent les $(k+1)$ racines du polynôme p_{k+1} (i.e. $x < y < z$ avec $p_{k+1}(x) = p_{k+1}(z) = 0$ et $p_k(y) = 0$). Soit a un réel quelconque, si on pose

$$Sgn(p_k(a)) = \begin{cases} sgn(p_k(a)) & \text{si} \quad p_k(a) \neq 0 \\ sgn(p_{k-1}(a)) & \text{si} \quad p_k(a) = 0 \end{cases}$$

alors on démontre que le nombre $N(k, a)$ de changements de signes entre éléments de l'ensemble ordonné $\{+, Sgn(p_1(a)), ..., Sgn(p_k(a))\}$ est égal au nombre de racines du polynôme p_k qui sont strictement inférieures à a.

Algorithme de Givens. Pour déterminer une valeur propre de la matrice B, on se donne un intervalle arbitraire $[a_0, b_0]$ contenant λ_i. On prendra par exemple $a_0 = -b_0 = \|B\|$. Soit c_0 le milieu de l'intervalle $[a_0, b_0]$

$$\begin{array}{lll} \text{si } N(n, c_0) \geq i & & \lambda_i \in [a_0, c_0[\\ \text{si } N(n, c_0) < i & & \lambda_i \in [c_0, b_0] \end{array}$$

On restreint alors l'intervalle de recherche à $[a_1, b_1]$ dans lequel on peut trouver λ_i. On détermine ainsi une suite d'intervalles emboîtés $[a_k, b_k]$ contenant λ_i et de longueur $(b_0 - a_0)/2^k$.

6.5 Méthode de Rutishauser

La *méthode de Rutishauser* est fondée sur la décomposition LU où L est une matrice triangulaire inférieure dont les éléments diagonaux sont égaux

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

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

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

6.6 Méthode de Francis

La *méthode de Francis* est identique à la méthode de Rutishauser à ceci près qu'elle utilise la décomposition QR (au lieu de la décomposition LU). À chaque étape, la matrice B_k est mise sous la forme d'un produit Q_kR_k où Q_k est une matrice unitaire et R_k une matrice triangulaire supérieure. Ces matrices sont réutilisées pour former la matrice $B_{k+1} = R_kQ_k$ qui est à son tour décomposée. L'algorithme est le suivant. On décompose la matrice A en $A = Q_1R_1$. Connaissant les matrices R_1 et Q_1, on forme le produit $B_1 = R_1Q_1$. Puis, on décompose B_1 en $B_1 = Q_2R_2$. À chaque étape, on décompose B_k en

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

les matrices sont réutilisées pour calculer

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

La matrice B_k

$$B_{k+1} = P_k^*AP_k \qquad \text{avec } P_k = Q_1Q_2...Q_k$$

est une matrice triangulaire supérieure ayant sur sa diagonale les valeurs propres de A. La matrice P_k est la matrice des vecteurs propres de A, i.e. dont les colonnes sont les vecteurs propres associés. Pour obtenir une

décomposition QR, on introduit les matrices de Jacobi $H_{p,q}$ en choisissant α de façon à annuler les coefficients triangulaires inférieurs $(a_{2,1}, ..., a_{n,1})$, puis $a_{3,2}, ...a_{n,2}$, etc. Les matrices

$$R_1 = H^t_{n,n-1}...H^t_{n,1}\, H^t_{n-1,n-2}... H^t_{n-1,1} \,...\, H^t_{3,2}\, H^t_{3,1}\, H^t_{2,1} A$$

et

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

satisfont la décomposition QR.

6.7 Méthode de Lanczòs

Lorsque la matrice A est symétrique, la *méthode de Lanczòs* est un cas particulier de l'algorithme d'Arnoldi. Nous présentons ici le cas A symétrique. Le procédé de Lanczòs consiste à construire une base orthonormée V_k de l'espace de Krylov $\mathcal{K}_k(A, v)$, où v est un vecteur arbitraire. On pose $v_1 = v/\|v\|$ et on calcule successivement les vecteurs v_j de la base de Krylov. L'algorithme est le suivant : Pour $k = 1, ..., m$

$$\left\{\begin{array}{l} w = Av_k \\ \text{Si } k > 1 \text{ Alors Faire } \beta_k = h_{k-1,k} = h_{k,k-1} \text{ et } w = w - h_{k,k-1}v_{k-1} \\ \alpha_k = h_{k,k} = (w, v_k) \\ w = w - h_{k,k}v_k \\ h_{k+1,k} = \|w\| \\ v_{k+1} = w/h_{k+1,k} \end{array}\right.$$

La matrice H_m de l'algorithme d'Arnoldi se réduit ici à une matrice tridiagonale

$$T_m = \begin{pmatrix} \alpha_1 & \beta_2 & 0 & ... & 0 \\ \beta_2 & \alpha_2 & \beta_3 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_m \\ 0 & ... & 0 & \beta_m & \alpha_m \end{pmatrix}$$

avec

$$\begin{aligned} \alpha_k &= (v_k, Av_k) \quad \text{pour } k = 1, .., m \\ \beta_k &= (v_k, Av_{k-1}) \quad \text{pour } k = 2, .., m \end{aligned}$$

Les coefficients de la matrice T_m vérifient pour $k = 2, ..., m-1$

$$\beta_{k+1}v_{k+1} = Av_k - \alpha_k v_k - \beta_k v_{k-1}$$

La matrice T_m a les mêmes valeurs propres que A. L'itération sur m conduit à calculer à chaque pas m valeurs propres. Ces valeurs convergent vers les valeurs propres de A. On arrête l'itération lorsque la différence entre deux estimations successives des valeurs propres est devenue suffisamment petite. Lorsque la matrice A est non symétrique, la méthode de Lanczòs donne des valeurs incorrectes pour des valeurs propres multiples ou pour des valeurs propres proches les unes des autres. Dans ce cas, on emploie une méthode bi-Lanczòs qui consiste à construire deux bases des espaces de Krylov $\mathcal{K}_k(A, v)$ et $\mathcal{K}_k(A^t, \widetilde{v})$.

6.8 Calcul du polynôme caractéristique

6.8.1 Méthode de Krylov

La *méthode de Krylov* utilise le théorème de Cayley-Hamilton pour calculer le polynôme caractéristique

$$P(\lambda) = (-1)^n(\lambda^n + p_1\lambda^{n-1} + p_2\lambda^{n-2} + ... + p_{n-1}\lambda + p_n)$$

Soit A la matrice associée au polynôme P et définie par

$$P(\lambda) = \det(A - \lambda I) = \begin{vmatrix} -(\lambda + p_1) & -p_2 & -p_3 & \cdots & -p_n \\ 1 & -\lambda & 0 & \cdots & 0 \\ 0 & 1 & -\lambda & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -\lambda \end{vmatrix}$$

Si b est un vecteur arbitraire, on pose $b_0 = A^n b$, $b_1 = A^{n-1}b$, ..., $b_{n-1} = Ab$, $b_n = b$. En appliquant $P(A) = 0$, c'est-à-dire

$$A^n + \sum_{i=1}^{n} p_i A^{n-i} = 0$$

au vecteur b, on obtient, en répétant l'opération, un système de n équations à n inconnues $(p_1, p_2, ..., p_n)$

$$p_1 b_1 + p_2 b_2 + ... + p_n b_n = -b_0$$

qui se résout par une méthode de résolution de systèmes linéaires.

6.8.2 Méthode de Leverrier

La *méthode de Leverrier* utilise la trace matricielle pour calculer le polynôme caractéristique

$$P(\lambda) = (-1)^n(\lambda^n + p_1\lambda^{n-1} + p_2\lambda^{n-2} + ... + p_{n-1}\lambda + p_n)$$

Notons $\lambda_1, \lambda_2, ..., \lambda_n$ les valeurs propres de P (non nécessairement distinctes).

$$P(\lambda) = (\lambda_1 - \lambda)(\lambda_2 - \lambda)...(\lambda_n - \lambda)$$

En posant, pour $k = 1, 2, ..., n$,

$$s_k = tr(A^k) = \sum_{i=1}^{n} \lambda_i^k$$

Les valeurs p_k des coefficients du polynôme caractéristique sont données par les formules de Newton

$$\left\{ \begin{array}{l} -p_1 = s_1 \\ -2p_2 = s_2 + p_1 s_1 \\ ... \\ -k.p_k = s_k + p_1 s_{k-1} + ... + p_{k-1} s_1 \\ ... \\ -n.p_n = s_n + p_1 s_{n-1} + ... + p_{n-1} s_1 \end{array} \right.$$

qui est un système triangulaire qui se résout de proche en proche par la méthode de remontée.

Exemple. La matrice

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 3 \end{pmatrix}$$

a pour trace $s_1 = 6$. Les matrices

$$A^2 = \begin{pmatrix} 1 & 3 & 0 \\ 0 & 4 & 0 \\ 0 & -5 & 9 \end{pmatrix} \quad A^3 = \begin{pmatrix} 1 & 7 & 0 \\ 0 & 8 & 0 \\ 0 & -19 & 27 \end{pmatrix}$$

ont pour trace $s_2 = 14$ et $s_3 = 36$. Les équations $p_1 = s_1 = -6$, $-2p_2 = s_2 + p_1 s_1 = 14 - 36$ et $-3p_3 = s_3 + p_1 s_2 + p_2 s_1 = 36 - 84 + 66$, conduisent à $p_1 = -6$, $p_2 = 11$ et $p_3 = -6$. Le polynôme caractéristique est donc

$$P(\lambda) = -\lambda^3 - p_1\lambda^2 - p_2\lambda - p_3 = -\lambda^3 + 6\lambda^2 - 11\lambda + 6$$

6.8.3 Méthode de Faddeev

La *méthode de Faddeev*, aussi appelée *méthode de Souriau-Leverrier* utilise la trace matricielle pour calculer le polynôme caractéristique

$$P(\lambda) = (-1)^n(\lambda^n + p_1\lambda^{n-1} + p_2\lambda^{n-2} + ... + p_{n-1}\lambda + p_n)$$

En posant

$$\left\{ \begin{array}{l} A_1 = A \\ A_k = (A_{k-1} + p_{k-1}I)A \qquad k = 2, ..., n \end{array} \right.$$

Le calcul des coefficients du polynôme caractéristique s'obtient par l'expression

$$p_k = -\frac{1}{k} tr(A_k)$$

Exemple. La matrice

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 3 \end{pmatrix}$$

a pour polynôme caractéristique

$$P(\lambda) = -\lambda^3 + 6\lambda^2 - 11\lambda + 6$$

Retrouvons ce résultat en appliquant l'algorithme de Faddeev. On calcule le coefficient p_1 à partir de la trace de la matrice $p_1 = -tr(A) = -6$. Puis, la matrice

$$A_2 = A^2 + p_1 A = \begin{pmatrix} -5 & -3 & 0 \\ 0 & -8 & 0 \\ 0 & 1 & -9 \end{pmatrix}$$

donne la valeur $p_2 = -tr(A_2)/2 = 11$. Le calcul de A_3

$$A_3 = A^3 + p_1 A^2 + p_2 A = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix}$$

conduit à la valeur $p_3 = -tr(A_3)/3 = -6$. On retrouve bien l'expression du polynôme caractéristique.

6.9 Exercices

1. On considère la matrice

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

   Calculer les valeurs propres et les vecteurs propres de cette matrice en appliquant la méthode de Rutishauser, puis celle de Francis.

2. Calculer le polynôme caractéristique de la matrice

$$A = \begin{pmatrix} 1 & 3 & -1 \\ 2 & 1 & 1 \\ 0 & 2 & 0 \end{pmatrix}$$

   par la méthode de Krylov.

3. On considère la matrice

$$A = \begin{pmatrix} -2 & 0 & 0 \\ 0 & 3 & -1 \\ 0 & -1 & 3 \end{pmatrix}$$

Calculer les valeurs propres de cette matrice par la méthode de Givens.

7

Équations et systèmes d'équations différentielles

On rappelle, dans les premiers paragraphes, les résultats fondamentaux des équations différentielles. Les paragraphes suivants sont consacrés aux méthodes numériques usuelles. Dans le traitement numérique des équations différentielles, on distingue les *méthodes à pas séparés* (ou à un seul pas) qui permettent de calculer y_{n+1} à partir de la seule connaissance de y_n et les *méthodes à pas liés* (ou à pas multiples) qui nécessitent la connaissance de y_n, $y_{n-1}, ..., y_{n-p}$ pour calculer y_{n+1}. Les méthodes numériques de résolution (dites à *différences finies*) sont fondées sur le développement de Taylor.

7.1 Existence et unicité des solutions

Soit U un ouvert de $\mathbb{R}\times\mathbb{R}^n$ et $f : U \to \mathbb{R}^n$ une application continue. On considère une équation différentielle de la forme $\dot{x} = f(t, x)$ munie d'une condition initiale $x(t_0) = x_0 \in \mathbb{R}^n$ donnée, avec la notation $\dot{x} = dx/dt$. Une solution de cette équation est (par définition) une application *différentiable* $y : [a, b] \to \mathbb{R}^n$ vérifiant les conditions

$$\begin{cases} \dot{x} = f(t, x) & \forall\, t \in [a, b] \\ x(t_0) = x_0 \end{cases}$$

Ce problème admet, sous certaines conditions, une solution. On démontre en effet que si la fonction f est une fonction continûment différentiable alors il existe deux valeurs maximales (éventuellement infinies) t_1 et t_2

pour lesquelles la solution de l'équation différentielle existe et est unique pour tout t dans l'intervalle $[t_0 - t_1, t_0 + t_2]$. On démontre aussi que si f est une fonction continue dans $[a, b]\times\mathbb{R}^n$ et vérifie la condition de Lipschitz de rapport $k > 0$ suivante

$$\forall\, t \in [a, b],\ \forall\ x_1, x_2 \in \mathbb{R}^n, \quad \|f(t, x_2) - f(t, x_1)\| \leq k\, \|x_2 - x_1\|$$

alors le problème avec condition initiale (t_0, x_0) admet une solution unique dans $[a, b]$, donnée par

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds$$

L'équation est dite *autonome* lorsqu'elle ne dépend pas explicitement de la variable temporelle t, c'est-à-dire lorsqu'elle est de la forme $\dot{x} = f(x)$. Il importe de remarquer que la condition de Lipschitz assure l'unicité de la solution. Par exemple, l'équation $dx/dt = 2\sqrt{x}$ pour $t > 0$ et $x(0) = 0$ admet plusieurs solutions de la forme $x(t) = (t - a)^2$ si $t \geq a$ et $x(t) = 0$ si $t \leq a$, obtenues pour différentes valeurs du paramètre a, car la fonction $x(t)$ n'est pas lipschitzienne. Remarquons aussi que l'intervalle de définition d'une solution dépend de la condition initiale et que l'existence d'une solution est une propriété locale. L'équation $dx/dt = x^2$ et $x(1) = -1$ admet comme solution $x(t) = -1/t$ qui n'est pas définie en $t = 0$ bien que la fonction $f(t, x) = x^2$ soit continue.

7.2 Champs de vecteurs

L'équation différentielle

$$\dot{x} = f(t, x)$$

définit un champ de vecteur f. On appelle *champ de vecteurs* une application qui à tout point x associe un vecteur tangent en ce point. Plus généralement, un champ de vecteurs sur une variété différentiable M est une section différentiable du fibré tangent T_M sur la variété M. À tout champ de vecteur X, on peut associer une équation différentielle $\dot{x} = X(x)$.

Une *courbe intégrale* de l'équation différentielle $\dot{x} = f(t, x)$ est un arc paramétré dérivable γ de $[0, 1]$ dans $\mathbb{R}^n$ qui vérifie

$$\frac{d\gamma(t)}{dt} = f(t, \gamma(t))$$

Une courbe intégrale admet donc en chacun de ses points x, un vecteur tangent $f(t, x)$. Pour résoudre une équation différentielle il suffit donc de trouver toutes ses courbes intégrales.

Soit I un intervalle réel centré sur l'origine et U un ouvert d'un espace vectoriel E. On appelle *groupe local à un paramètre* tout difféomorphisme

$\phi : (t, x) \to \phi(t, x)$ de $I \times U$ sur E noté $\phi_t(x) = \phi(t, x)$ tel que l'application $t \to \phi_t$ de I dans E soit un homomorphisme de groupes, c'est-à-dire tel que (1) l'application ϕ_0 soit l'identité, (2) l'application composée vérifie $\phi_t \circ \phi_s = \phi_{t+s}$ et (3) l'application réciproque est donnée par $(\phi_t)^{-1} = \phi_{-t}$. On rappelle qu'un *difféomorphisme de classe* C^p ϕ_t est une application bijective telle que l'application ϕ_t et sa réciproque sont de classe C^p. On définit un groupe (global) à un paramètre en prenant pour l'intervalle I la droite réelle et pour l'ouvert U l'espace tout entier E. À chaque champ de vecteurs, on peut associer un groupe local à un paramètre. Pour un champ donné X, les courbes intégrales associées sont les courbes définies par les fonctions $\gamma_x(t) = \phi_t(x)$. On appelle *orbite* d'un point x du champ X, la courbe intégrale $\gamma_x(t)$ passant par x. Inversement, à un groupe de paramètres, on peut associer le champ de vecteur

$$X(t) = \frac{d}{dt}\phi(t, x)\bigg|_{t=0}$$

qui est le vecteur vitesse de l'arc paramétré $\gamma_x(t)$. Un champ de vecteurs est *complet* si toutes les courbes intégrales maximales sont paramétrées sur l'ensemble des réels tout entier (i.e. t varie de $-\infty$ à $+\infty$). La différence entre *local* et *global* est essentielle. En effet, tout champ de vecteurs engendre un groupe local à un paramètre de difféomorphismes, mais n'engendre pas nécessairement un groupe global. On démontre que pour qu'un champ soit complet, il faut et il suffit que ce champ soit engendré par un groupe (global) à un paramètre de difféomorphismes. Par exemple, le champ de vecteurs de classe C^∞ sur $\mathbb{R}^2$

$$X = x^2 \frac{\partial}{\partial x}$$

est associé à l'équation différentielle $\dot{x} = x^2$, avec pour condition initiale $x(0) = a$ $(a \neq 0)$. Il admet les courbes intégrales d'équations

$$x(t) = \frac{a}{1 - at}$$

Comme ces solutions ne sont pas définies pour la valeur $t = 1/a$, le champ n'est pas complet. En revanche, le champ de vecteurs

$$X = x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}$$

associé à l'équation différentielle

$$\begin{cases} \dot{x} = -y \\ \dot{y} = x \end{cases}$$

admet comme courbes intégrales $x(t) = x_0 cos(t) - y_0 sin(t)$ et $y(t) = x_0 sin(t) + y_0 cos(t)$. Le champ est complet, car les courbes intégrales sont définies à tout instant. Les courbes intégrales sont des cercles concentriques. Le groupe à un paramètre est le groupe des rotations planes d'angle t.

7.3 Inversion locale

Lorsque f est une application différentiable de $\mathbb{R}^n$ dans $\mathbb{R}^n$, la différentielle de $f = (f_1, f_2, ...f_n)$ au point $a = (a_1, a_2, ..., a_n)$ est l'application linéaire définie par la matrice jacobienne

$$Jac\, f(a) = \left(\frac{\partial f_j}{\partial x_i}\right)_{i,j}(a) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \cdots & \frac{\partial f_1}{\partial x_n}(a) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(a) & \cdots & \frac{\partial f_n}{\partial x_n}(a) \end{pmatrix}$$

Le jacobien est le déterminant de cette matrice. Donnons quelques définitions.

Une application f d'un ouvert U d'un espace vectoriel de dimension n dans un espace vectoriel E de dimension $n + p$ de la forme

$$f = (f_1(x_1, ...x_n), ..., f_{n+p}(x_1, ...x_n))$$

est une *immersion* lorsque la dérivée de f est injective, c'est-à-dire lorsque la matrice jacobienne est de rang n (on peut extraire une matrice carrée d'ordre n de déterminant non nul).

Une application f d'un ouvert U d'un espace vectoriel de dimension $n+p$ dans un espace vectoriel E de dimension p de la forme

$$f = (f_1(x_1, ...x_{n+p}), ..., f_p(x_1, ...x_{n+p}))$$

est une *submersion* lorsque la dérivée de f est surjective, c'est-à-dire lorsque la matrice jacobienne est de rang p (on peut extraire une matrice carrée d'ordre p de déterminant non nul).

Le *théorème du difféomorphisme local* assure l'existence de solutions locales d'une équation différentielle dès que f est suffisamment régulière. Soit U un ouvert, f une application de U dans un espace vectoriel E de classe C^p. Soit x un point de l'ouvert U pour lequel le jacobien de f en ce point est non nul. Alors il existe un ouvert U' inclus dans U et contenant x et un ouvert U'' contenant $f(x)$ tel que la restriction de f à U' soit un difféomorphisme de classe C^p de U' sur U''. Autrement dit, on peut toujours inverser localement la fonction f. Si de plus f est injective, alors f est un difféomorphisme global de U sur $f(U)$.

Le *théorème des fonctions implicites* permet de déterminer une réciproque locale et de calculer sa dérivée. Soit f une application de $\mathbb{R}^n$ dans $\mathbb{R}$ de classe C^1 au voisinage du point $a = (a_1, ..., a_n)$ telle que $f(a) = 0$. On suppose que la dérivée partielle de f en x_n au point a est non nulle $f'_{x_n}(a) \neq 0$. Alors pour un voisinage du point $(a_1, ..., a_{n-1})$, l'équation $f(x_1, ..., x_n) = 0$ admet une solution unique $h(x_1, ...x_{n-1})$ définie de l'intervalle local $]x_0 - \alpha, x_0 + \alpha[\times\,]x_1 - \alpha, x_1 + \alpha[\times\, ...\times\,]x_{n-1} - \alpha, x_{n-1} + \alpha[$

dans $]x_n - \beta, x_n + \beta[$ telle que

$$f(x_1, ..., x_{n-1}, h(x_1, ...x_{n-1})) = 0$$

et dont la dérivée est donnée par

$$dh(x_1, ...x_{n-1}) = -\frac{f'_{x_1}(x_1, ...x_{n-1}, h)}{f'_{x_n}(x_1, ...x_{n-1}, h)}dx_1 - .... - \frac{f'_{x_{n-1}}(x_1, ...x_{n-1}, h)}{f'_{x_n}(x_1, ...x_{n-1}, h)}dx_{n-1}$$

Par exemple, la fonction $f(x, y) = y^5 - 4y^4 + 4xy^2 - x^2$ s'annule au point (1,1). La dérivée de f en y au point (1,1) est non nulle $f'_y(1, 1) = -3$, par conséquent l'équation $f(x, y) = 0$ admet une solution locale $y = h(x)$ dont la dérivée vaut $-2/3$. La tangente au point $(a_1, a_2) = (1, 1)$ a pour équation $y - a_1 = h'(x)(x - a_2)$, soit $y = -2x/3 + 5/3$.

7.4 Équations différentielles linéaires

L'équation ou le système d'équations différentielles de la forme

$$\dot{X}(t) = AX(t) + B(t)$$

où A est une matrice et $B(t)$ une fonction continue admet une solution unique prenant en $t = t_0$ la valeur x_0, qui est donnée par

$$X(t) = e^{(t-t_0)A}.x_0 + \int_{t_0}^{t} e^{(t-s)A}.B(s)ds$$

On rappelle que l'*exponentielle d'une matrice* est définie par la série

$$e^A = \sum_{n=0}^{+\infty} \frac{A^n}{n!}$$

Le déterminant de l'exponentielle d'une matrice est égal à l'exponentielle de la trace de cette matrice, qui est la somme des valeurs propres

$$\det(e^A) = e^{tr(A)}$$

En pratique, on résout d'abord l'équation homogène

$$\dot{X}(t) = AX(t)$$

puis on détermine une solution particulière de l'équation globale. L'ensemble des solutions est obtenu par superposition de la solution de l'équation homogène augmentée d'une solution particulière. L'équation homogène se résout en mettant la matrice A sous sa forme de Jordan, qui est, dans une

base convenable, la somme d'une matrice diagonale et d'une matrice nilpotente $J = D + N$. Si P désigne la matrice de passage $A = P(D+N)P^{-1}$, les solutions sont de la forme $X(t) = Pe^{tD}e^{tN}P^{-1}X_0$ où $P^{-1}X_0$ est un vecteur arbitraire. Lorsque A est une matrice diagonalisable, la résolution de l'équation homogène s'effectue par un simple changement de variables. Puisque A est de la forme $A = PDP^{-1}$, l'équation $X' = AX$ est équivalente à l'équation $X' = PDP^{-1}X$, soit en posant $Y = PX$, cette équation devient $PY' = PDP^{-1}PY$, soit en multipliant à gauche par l'inverse de P, $Y' = DY$. Cette dernière équation est un système diagonal qui se résout simplement en $y_i' = \alpha_i e^{\lambda_i t}$ où les coefficients α_i sont des constantes arbitraires et les λ_i sont les valeurs propres de A. Le vecteur X, qui est relié au vecteur Y par la matrice de passage, est donné par la formule $X = P^{-1}Y$ qui résout l'équation homogène.

Les équations linéaires d'ordre p se ramènent à des systèmes d'équations. L'équation

$$x^{(p)} + a_{p-1}x^{(p-1)} + ... + a_1x' + a_0x = 0$$

est équivalente au système

$$\begin{cases} x' = x_1 \\ x_1' = x_2 \\ ... \\ x_{p-2}' = x_{p-1} \\ x_{p-1}' = x^{(p)} = -a_{p-1}x_{p-1} - ... - a_1x_1 - a_0x \end{cases}$$

Ce système est de la forme matricielle $X' = AX$. Si on désigne par $\lambda_1, \lambda_2, ...,$ λ_n, les valeurs propres de la matrice A, le système admet pour solution

$$\begin{pmatrix} x \\ x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} 1 & 1 & ... & 1 \\ \lambda_1 & \lambda_2 & ... & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & ... & \lambda_n^2 \\ \cdots & & & \cdots \\ \lambda_1^p & \lambda_2^p & ... & \lambda_n^p \end{pmatrix} \begin{pmatrix} c_1e^{\lambda_1 t} \\ c_2e^{\lambda_2 t} \\ c_3e^{\lambda_3 t} \\ \cdots \\ c_pe^{\lambda_p t} \end{pmatrix}$$

où les coefficients c_j sont des constantes. Les solutions sont donc de la forme

$$x(t) = c_1e^{\lambda_1 t} + ... + c_pe^{\lambda_p t}$$

Dans le cas d'une équation du second degré, les racines du polynôme caractéristique nous renseignent sur le type de solution. Considérons l'équation

$$a\ddot{x} + b\dot{x} + cx = 0$$

À cette équation différentielle est associé un polynôme P appelé polynôme caractéristique

$$P(\lambda) = a\lambda^2 + b\lambda + c$$

Ce polynôme P détermine l'équation caractéristique $P(\lambda) = 0$ de l'équation différentielle. Dans l'ensemble des complexes, les solutions de cette équation forment un espace vectoriel sur l'ensemble des complexes de dimension 2. Si le discriminant du polynôme caractéristique ($\Delta = b^2 - 4ac$) est non nul, l'équation caractéristique admet deux racines distinctes complexes λ_1 et λ_2. Les solutions sont de la forme $x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$, les coefficients c_i étant complexes. Si le discriminant est nul, l'équation caractéristique admet une racine double λ, et les solutions sont de la forme $x(t) = (c_1 + c_2 t)e^{\lambda t}$.

Dans l'ensemble des réels, si le discriminant est strictement positif, l'équation caractéristique admet deux racines réelles distinctes λ_1 et λ_2. Les solutions sont de la forme $x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$, les coeffcients c_i étant réels. Si le discriminant est nul, l'équation caractéristique admet une racine double λ, et les solutions sont de la forme $x(t) = (c_1 + c_2 t)e^{\lambda t}$. Enfin, si le discriminant est négatif, les deux racines sont complexes conjuguées de la forme $\lambda_1 = \alpha + i\beta$, et $\lambda_2 = \alpha - i\beta$, de sorte que la solution est de la forme $x(t) = e^{\alpha t}(c_1 \cos(\beta t) + c_2 \sin(\beta t))$, équation qui peut encore s'écrire en introduisant un facteur de phase $x(t) = Ce^{\alpha t} \cos(\beta t + \varphi)$.

Certaines équations non linéaires se ramènent à des équations linéaires par changement de variables. C'est le cas par exemple des *équations de Bernoulli*

$$\dot{x} = p(t)x + q(t)x^{\alpha}$$

les fonctions p et q étant continues. Lorsque α vaut 0 ou 1 l'équation est linéaire. Sinon, en posant $y = x^{1-\alpha}$, on se ramène à l'équation linéaire suivante

$$\frac{\dot{y}}{1-\alpha} = p(t)y + q(t)$$

L'*équation de Ricatti*

$$\dot{x} = a(t)x^2 + b(t)x + c(t)$$

se ramène à une équation de Bernoulli avec $\alpha = 2$, dès qu'on en connaît une solution particulière $x_1(t)$. En effet, il suffit de poser $x = x_1 + y$, et de reporter dans l'équation, pour montrer que la variable y vérifie l'équation de Bernoulli suivante

$$\dot{y} = (2a(t)x_1(t) + b(t))y(t) + a(t)y^2(t)$$

7.5 Points critiques

Un point a est un point *critique* (*stationnaire* ou *singulier*) de l'équation différentielle associée au champ f, si $f(a) = 0$. Si h est une fonction réelle définie sur un ouvert U contenant une variété M, on dit que a, un point de M, est un point critique de h sur M si la dérivée de h s'annule en

$a : h'(a) = 0$. Cette définition établit une condition suffisante pour que la fonction h ait un extremum relatif. Nous savons déjà que pour que h ait un extremum relatif en a, il faut que la dérivée de h en a s'annule. La forme bilinéaire symétrique définie lorsque h est de classe C^2 sur le fibré tangent $T_a(M)$ par

$$\phi(u,v) = (h \circ \phi)''(0)(\phi'(0)^{-1}(u), \phi'(0)^{-1}(v)) = Hess\,(u,v)$$

est appelée la *hessienne* de h en a. Soit $(e_1, ...e_r)$ une base orthogonale de $T_a(M)$

$$\phi(e_i, e_j) = k_i \delta_{ij}$$

Le point critique a est non dégénéré si la forme bilinéaire Hess est non dégénérée, c'est-à-dire si aucun k_i n'est nul. L'indice de a est l'indice de Hess, c'est-à-dire le nombre de k_i strictement négatifs. Si tous les points critiques de h sur M sont non dégénérés, h est appelée *fonction de Morse*. Si M est une variété compacte et si h est une fonction de Morse, h a un nombre fini de points critiques. La caractéristique d'Euler-Poincaré de M $(dimM = r)$ est liée aux points critiques par la formule

$$\chi(M) = \sum_{j=0}^{r} (-1)^j c_j(h)$$

c_j est le nombre de points critiques de h d'indice j. Les $c_j(h)$ sont plus grands que les nombres de Betti de M notés $b_j(M)$

$$c_j(h) \geq b_j(M)$$

7.6 Ensembles limites

Un point y est un point ω-limite pour le flot $\varphi(t,x)$ associé à l'équation différentielle $\dot{x} = f(x)$ s'il existe une suite croissante de réels t_n tendant vers l'infini telle que $\lim \varphi(t_n, x) = y$. L'ensemble ω-limite de x est l'ensemble

$$\omega(x) = \{y \in \mathbb{R}^n,\ \exists (t_n) \nearrow \infty,\ \lim_{n\to\infty} \varphi(t_n, x) = y\}$$

Un point y est un point α-limite pour le flot $\varphi(t,x)$ associé à l'équation différentielle $\dot{x} = f(x)$ s'il existe une suite décroissante de réels s_n tendant vers l'infini telle que $\lim \varphi(s_n, x) = y$. L'ensemble α-limite de x est l'ensemble

$$\alpha(x) = \{y \in \mathbb{R}^n,\ \exists (s_n) \searrow -\infty,\ \lim_{n\to\infty} \varphi(s_n, x) = y\}$$

Un ensemble A est *invariant* si

$$\forall x \in A,\ \varphi(t,x) \in A,\ \forall t$$

Un ensemble A est positivement invariant (resp. négativement invariant) si $\forall x \in A$, $\varphi(t,x) \in A$, $\forall t > 0$ (resp. $t < 0$). La courbe intégrale ou la *trajectoire* passant par x est l'ensemble

$$\gamma(x) = \bigcup_{t\in\mathbb{R}} \varphi(t,x)$$

et la *semi-trajectoire positive* est l'ensemble

$$\gamma^+(x) = \bigcup_{t\geq 0} \varphi(t,x)$$

On démontre que l'ensemble A est invariant si et seulement si $\gamma(x) \subset A$, $\forall x \in A$. L'ensemble ω-limite est l'intersection des fermetures des semi-trajectoires positives

$$\omega(x) = \bigcap_{y\in\gamma(x)} \overline{\gamma^+(y)}$$

Si $\omega(x)$ est invariant et si $\gamma^+(x)$ est borné, alors $\omega(x)$ est compact.

Exemples. (1) L'équation différentielle $\dot{x} = -x$ avec $x(0) = 0$ admet comme solution les courbes $x(t) = Ce^{-t}$. Le flot associé est donc $\varphi(t,x) = Ce^{-t}$. L'ensemble ω-limite est réduit à $\{0\}$. Si $x = 0$, $\alpha(x) = 0$, mais si x est positif ou négatif, l'ensemble α-limite est vide.
(2) Considérons le système d'équations

$$\begin{cases} \dot{x} = -y + x(1 - x^2 - y^2) \\ \dot{y} = x + y(1 - x^2 - y^2) \end{cases}$$

En coordonnées polaires, ce système équivaut à $\dot{r} = r(1 - r^2)$. Les points d'équilibre sont obtenus pour $r = 0$ et $r = 1$. L'ensemble ω-limite est égal au cercle de rayon 1, si r est non nul et se réduit à $\{0\}$ si $r = 0$. L'ensemble α-limite est vide si r est plus grand que 1, c'est-à-dire à l'extérieur du cercle unité et est égal au point origine, si r est inférieur à 1.

7.7 Stabilité de Lyapunov

On considère l'équation différentielle $\dot{x} = f(t,x)$ et on note $\varphi(t,x)$ le flot associé. On suppose que l'équation admet une solution $x(t)$ pour la condition initiale $x(t_0) = x_0$. Un point x ou une solution $x(t)$ est *stable au sens de Lyapunov* si pour tout ϵ positif, il existe un nombre δ positif tel que pour toute solution $y(t)$ de la même équation, l'inégalité $|x(t_0) - y(t_0)| < \delta$ entraîne que $|x(t) - y(t)| < \epsilon$, pour tout t supérieur à t_0. Autrement dit, la stabilité de Lyapunov demande à ce que les solutions qui sont proches des conditions initiales, restent proches lorsque le temps (ou la variable d'intégration) augmente. Lorsque x est un élément de $\mathbb{R}^n$, la notation des valeurs absolues représente la norme. Un point x ou une solution $x(t)$ est

dite *quasi asymptotiquement stable* si pour toutes les solutions proches de $x(t_0)$, vérifiant $|x(t_0) - y(t_0)| < \delta$, on a

$$\lim_{t\to\infty} |x(t) - y(t)| = 0$$

Un point ou une solution est *asymptotiquement stable* si elle est à la fois stable et quasi-asymptotiquement stable. Par exemple, pour l'équation différentielle donnée en coordonnées polaires $\dot{r} = 0$, $\dot{\theta} = 1$, dont les solutions $\dot{r} = r_0$, $\dot{\theta} = t + \dot{\theta}_0$ sont des cercles parcourus à vitesse constante (égale à 1), l'origine est stable, mais n'est pas quasi asymptotiquement stable. En revanche, les solutions de l'équation différentielle $\dot{x} = -x$ sont asymptotiquement stables. En effet, les solutions sont de la forme $x(t) = x_0 e^{-(t-t_0)}$. Deux solutions proches vérifient

$$\lim_{t\to\infty} |x(t) - y(t)| = \lim_{t\to\infty} |x_0 - y_0|\, e^{-(t-t_0)} = 0$$

Par changement de variables, on ramène l'étude de la stabilité au voisinage de l'origine. La *fonction de Lyapunov* V associée à l'équation différentielle $\dot{x} = f(x)$ est une fonction continûment différentiable définie sur un voisinage ouvert U de l'origine qui s'annule à l'origine $V(0) = 0$ et reste positive au voisinage de l'origine $V(x) > 0$, $\forall x \in \overline{U}\backslash 0$, et dont la dérivée le long des trajectoires (appelée *dérivée totale* ou *dérivée de Lie* selon le champ de vecteurs f associé à l'équation différentielle) reste négative

$$L_f V = \dot{V} = \dot{x}.\nabla V = f.\nabla V = \sum f_i(x).\frac{\partial V}{\partial x_i} \leq 0 \qquad \forall x \in U$$

Le *premier théorème de Lyapunov* affirme que s'il existe une fonction de Lyapunov V définie au voisinage de l'origine pour laquelle $x = 0$ est un point stationnaire de l'équation différentielle $\dot{x} = f(x)$, alors l'origine est stable au sens de Lyapunov.

Exemple. L'oscillateur non linéaire $\ddot{x} + c\dot{x} + ax + bx^3 = 0$, où les constantes a, b, c sont positives peut s'écrire sous la forme d'un système

$$\begin{cases} \dot{x} = y \\ \dot{y} = -ax - cy - bx^3 \end{cases}$$

La fonction $V(x, y) = 2ax^2 + bx^4 + 2y^2$ est positive et de dérivée totale $L_X V = \dot{V}(x, y) = -4cy^2$ toujours négative. $V(x, y)$ est une fonction de Lyapunov sur tout ouvert borné contenant l'origine. Le point stationnaire (0,0) est par conséquent stable.

Le *second théorème de Lyapunov* affirme que s'il existe une fonction de Lyapunov V définie au voisinage de l'origine pour laquelle $x = 0$ est un point stationnaire de l'équation différentielle $\dot{x} = f(x)$ telle que la dérivée

totale de V soit strictement négative pour tous les points du voisinage de 0 sauf en 0

$$\dot{V}(x) < 0, \quad \forall x \in U \backslash \{0\}$$

alors l'origine est asymptotiquement stable.

Exemple. Considérons l'oscillateur $\ddot{y} - a\dot{x}(2x - 1) + x = 0$, où a est une constante positive. Écrit sous la forme d'un système

$$\begin{cases} \dot{x} = y + a(x^2 - x) \\ \dot{y} = x \end{cases}$$

l'oscillateur admet une fonction de Lyapunov $V(x, y) = (x^2 + y^2)/2$ de dérivée totale $\dot{V}(x, y) = x.\dot{x} + y\dot{y} = ax^2(x - 1)$ toujours strictement négative au voisinage de 0. Le point stationnaire (0,0) est par conséquent asymptotiquement stable.

Le *théorème d'instabilité de N. Tchétaev*, parfois appelé *troisième théorème de Lyapunov* affirme que s'il existe une fonction V définie au voisinage de l'origine, pour laquelle $x = 0$ est un point stationnaire de l'équation différentielle $\dot{x} = f(x)$, de dérivée totale positive et s'il existe un point x_0 du voisinage de l'origine tel que $V(x_0)\dot{V}(x_0) > 0$, alors l'origine est un point instable.

Exemple. Considérons l'équation suivante

$$\begin{cases} \dot{x} = y + x^2 \\ \dot{y} = x + y^2 \end{cases}$$

La fonction $V(x, y) = (x^3 + y^3)/3 + xy$ est une fonction de dérivée de Lie

$$L_f V(x, y) = (y + x^2)^2 + (x + y^2)^2$$

positive. Le point de coordonnées $(\varepsilon, \varepsilon)$ avec $\varepsilon > 0$ vérifie la condition du théorème, par conséquent l'origine est un point instable.

Les résultats de stabilité s'étendent dans certains cas au problème linéarisé. Si $\dot{x} = f(x)$ a une linéarisation $\dot{x} = Ax$ où A est le jacobien de f en $x = 0$, et si la matrice A a n valeurs propres distinctes dont chacune a une partie réelle strictement négative, alors $x = 0$ est asymptotiquement stable.

7.8 Solutions périodiques. Théorie de Floquet

Floquet a étudié le comportement des solutions dans le cas périodique. Considérons tout d'abord l'équation $\dot{x} = a(t)x$ sur la droite réelle lorsque

$a(t)$ est une fonction de période T. Supposons que l'équation admette une solution

$$x(t) = x_0 \exp(\int_0^t a(s)ds) = x_0\varphi(t)$$

La fonction $\varphi(t)$ vérifie $\varphi(T)\varphi(t) = \varphi(t+T)$ et par conséquent $\varphi(nT) = \varphi(T)^n$. Le nombre $\varphi(T) = e^{\sigma T}$ est appelé *multiplicateur de Floquet* et σ est l'*exposant de Floquet* (défini à une constante $2ik\pi/T$ près). En notant

$$u(t) = \varphi(t)e^{-\sigma T}$$

la solution de l'équation différentielle s'écrit

$$x(t) = x_0 e^{\sigma T} u(t)$$

Par conséquent si la partie réelle de l'exposant de Floquet est positive, la solution tend vers zéro et si elle est négative, la solution diverge. Dans le cas multidimensionnel, le résultat est similaire. Notons $\phi(t)$ une matrice fondamentale, $\phi(0) = I$ solution du système d'équations $\dot{x} = A(t)x$ où $A(t)$ est une matrice d'ordre n à coefficients périodiques de période T. Les valeurs propres λ_i de la matrice fondamentale ϕ sont les *multiplicateurs de Floquet* et les nombres σ_i définis par $\lambda_i = e^{\sigma_i T}$ sont les *exposants de Floquet*. Le théorème de Floquet affirme que si tous les modules des multiplicateurs de Floquet sont inférieurs à 1 (ou si la partie réelle des exposants de Floquet est négative $Re(\sigma_i) < 0$), alors l'origine est asymptotiquement stable. Si au moins un des modules des multiplicateurs est plus grand que 1 (au moins une partie réelle des exposants de Floquet est positive), alors les solutions divergent. Si tous les modules des multiplicateurs de Floquet sont inférieurs ou égaux à 1 et si les multiplicateurs de modules égaux à 1 sont simples, alors la solution est stable.

7.9 Intégrales et fonctions elliptiques

Les fonctions elliptiques ressemblent étrangement aux fonctions circulaires. Elles sont définies à partir de l'inverse d'une intégrale elliptique et permettent la résolution de nombreuses équations différentielles, comme par exemple la résolution des oscillations d'un pendule de longueur l, soumis à une pesanteur g

$$l\frac{d^2\theta}{dt^2} = -g\sin\theta$$

Une *intégrale elliptique* est l'intégrale d'une fonction rationnelle $R(z, w)$ dans laquelle w^2 est un polynôme de degré 3 ou 4 en z. En général, une intégrale elliptique ne peut s'exprimer avec des fonctions élémentaires. On démontre qu'une intégrale elliptique s'écrit comme somme de fonctions élémentaires et d'une combinaison linéaire d'intégrales elliptiques du premier,

deuxième ou troisième type. Une intégrale elliptique du premier type dans sa forme normale de Legendre s'écrit pour $0 < k < 1$

$$u = \int_0^\phi \frac{1}{\sqrt{1 - k^2 \sin^2 t}} dt = F(\phi, k)$$

la variable $\phi = am\ u$ est appelée l'amplitude, k le module de l'intégrale u et $k' = 1 - k^2$ le module complémentaire. Cette intégrale est aussi égale à

$$u = \int_0^z \frac{dx}{\sqrt{(1 - x^2)(1 - k^2 x^2)}}$$

Les fonctions $F(\pi/2, k)$ et $F(\pi/2, k')$ sont solutions de l'équation différentielle

$$(1 - k^2)\frac{d^2u}{dk^2} + \frac{1 - 3k^2}{k}\frac{du}{dk} - u = 0$$

L'inverse de cette intégrale de première espèce définit la fonction elliptique de Jacobi $sn\ u$. Les trois fonctions elliptiques, introduites par Jacobi en 1829, sont définies par les relations

$$\begin{aligned} sn\ u &= \sin\phi \\ cn\ u &= \cos\phi \\ dn\ u &= \sqrt{1 - k^2 \sin^2 \phi} \end{aligned}$$

On introduit également la fonction $tn\ u = \tan\phi$. Ces fonctions dégénèrent en fonctions trigonométriques pour $k = 0$

$$sn(u, 0) = \sin u \qquad cn(u, 0) = \cos u$$

et en fonctions hyperboliques pour $k = 1$

$$sn(u, 1) = \tanh u \qquad cn(u, 1) = 1/\cosh u$$

Les fonctions elliptiques vérifient les relations usuelles

$$\begin{aligned} sn^2u + cn^2u &= 1 \\ dn^2u + k^2 sn^2u &= 1 \\ dn^2u - k^2 cn^2u &= 1 - k^2 \end{aligned}$$

Les dérivées sont données par les relations

$$\begin{aligned} \frac{d\ sn\ u}{du} &= cn\ u.dn\ u \\ \frac{d\ cn\ u}{du} &= -sn\ u.dn\ u \\ \frac{d\ dn\ u}{du} &= -k^2 sn\ u.cn\ u \end{aligned}$$

Les relations d'additivité

$$\begin{aligned} sn(u+v) &= \frac{sn\ u.cn\ v.dn\ v + sn\ v.cn\ u.dn\ u}{1-k^2 sn^2 u.sn^2 v} \\ cn(u+v) &= \frac{cn\ u.cn\ v - sn\ u.dn\ u.sn\ v.dn\ v}{1-k^2 sn^2 u.sn^2 v} \\ dn(u+v) &= \frac{dn\ u.dn\ v - k^2.sn\ u.cn\ u.sn\ v.dn\ v}{1-k^2 sn^2 u.sn^2 v} \end{aligned}$$

L'intégrale elliptique du second type est définie par les formules

$$E(\phi,k) = \int_0^{\phi} \sqrt{1-k^2\sin^2 t}dt = \int_0^z \sqrt{\frac{1-k^2x^2}{1-x^2}}dx$$

Les fonctions $E(\pi/2,k)$ et $E(\pi/2,k')$ sont solutions de l'équation différentielle

$$(1-k^2)\frac{d^2u}{dk^2} + \frac{1-k^2}{k}\frac{du}{dk} + u = 0$$

L'intégrale elliptique de troisième type est définie par

$$\begin{aligned} R(\phi,n,k) &= \int_0^{\phi} \frac{dt}{(1+n\sin^2 t)\sqrt{1-k^2\sin^2 t}} \\ &= \int_0^z \frac{dx}{(1+nx^2)\sqrt{(1-x^2)(1-k^2x^2)}} \end{aligned}$$

Exemple. L'équation différentielle

$$y'' = ay + by^3$$

admet comme solutions les fonctions de la forme

$$y(x) = c.sn(\lambda v, k)$$

où $v = x - x_0$, λ et x_0 sont des constantes, c est déterminé par l'équation $c^2 = -2(\lambda^2+a)/b$ et $k^2 = -(\lambda^2+a)/\lambda^2$.

7.10 Transcendantes de Painlevé

En considérant les équations différentielles du second ordre de la forme

$$\frac{d^2y}{dx^2} = f(y, y', x)$$

dans laquelle la fonction f est une fonction rationnelle en y et y', dérivée de y par rapport à x et analytique en x, Painlevé, Gambier et Picard ont cherché à classer les équations non linéaires du second ordre. Les équations

différentielles de ce type qui ont des points singuliers non paramétriques peuvent être classées en cinquante équations canoniques. Ces cinquante équations se résolvent avec des fonctions classiques ou transcendantes à l'exception de six équations, appelées *transcendantes de Painlevé*, qui nécessitent l'introduction de nouvelles fonctions (voir [Ince,1956], ou [Davis, 1962]). Les solutions des cinq premières transcendantes de Painlevé sont des fonctions analytiques en x. La première transcendante de Painlevé est solution de l'équation

$$y'' = 6y^2 + x$$

La deuxième transcendante de Painlevé est solution de l'équation (a est une constante complexe)

$$y'' = 2y^3 + xy + a$$

La troisième transcendante de Painlevé est solution de l'équation (a, b, c, d sont des constantes complexes $bd \neq 0$)

$$y'' = y'^2/y + e^x(ay^2 + b) + e^{2x}(cy^3 + d/y)$$

La quatrième transcendante de Painlevé est solution de l'équation (a, b sont des constantes complexes)

$$y'' = y'^2/2y + 3y^3/2 + 4xy^2 + 2(x^2 - a)y + b/y$$

La cinquième transcendante de Painlevé est solution de l'équation (a, b, c, d sont des constantes complexes $bd \neq 0$)

$$\begin{aligned} y'' &= y'^2\left(\frac{1}{2y} + \frac{1}{y-1}\right) - \frac{y'}{x} + \frac{(y-1)^2}{x^2}\left(ay + \frac{b}{y}\right) \\ &\quad + c\frac{y}{x} + d\frac{y(y+1)}{y-1} \end{aligned}$$

La solution de la sixième équation

$$\begin{aligned} y'' &= \frac{1}{2}y'^2\left(\frac{1}{y} + \frac{1}{y-1} + \frac{1}{y-x}\right) - \left(\frac{1}{x} + \frac{1}{x-1} + \frac{1}{y-x}\right)y' \\ &\quad + \frac{y(y-1)(y-x)}{x^2(x-1)^2}\left(a + b\frac{x}{y^2} + c\frac{x-1}{(y-1)^2} + d\frac{x(x-1)}{(y-x)^2}\right) \end{aligned}$$

admet trois points critiques, $x = 0, 1$ et ∞.

Les transcendantes de Painlevé jouent un rôle important dans la théorie des systèmes intégrables et des solitons.

7.11 Hyperbolicité. Variété centrale

On démontre que l'existence d'une transformation locale de changement de coordonnées d'un système $\dot{x} = f(x)$ par une forme linéarisée

$\dot{x} = Df(0)x$, obtenue à partir de la dérivée de f, dépend des valeurs propres de la matrice $Df(0)$, de même que la structure locale des variétés centrales, stables et instables. Lorsque le point stationnaire est hyperbolique, on a plusieurs résultats remarquables qui sont présentés dans ce paragraphe.

Soit $(\lambda_1, ..., \lambda_n)$ les valeurs propres de la matrice $Df(0)$. On dit que la matrice $Df(0)$ est résonante s'il existe des entiers $(m_1, ..., m_n)$ dont la somme est supérieure à 2 tels qu'il existe une valeur propre λ_s qui soit égale au produit scalaire des vecteurs $m = (m_1, ..., m_n)$ et $\lambda = (\lambda_1, ..., \lambda_n)$

$$\sum_{j=0}^{n} m_j \lambda_j = \lambda_s$$

La quantité $|m| = m_1 + ... + m_n \geq 2$ est appelée *ordre de la résonance*. On démontre que si $Df(0)$ est non résonante et diagonalisable, alors il existe un changement de coordonnées locales au voisinage de l'origine, $y = u(x)$ tel que $\dot{y} = Df(0)y$.

Un point stationnaire est *hyperbolique* si la dérivée $Df(x)$ en ce point n'a pas de valeur propre nulle ou purement imaginaire ($Re(\lambda_i) \neq 0$), i.e. toutes les valeurs propres ont leur partie réelle non nulle.

Pour un système linéaire $\dot{x} = Ax$, admettant un point critique en $x = 0$, on définit la *variété centrale* E^c comme l'espace invariant engendré par les vecteurs propres associés aux valeurs propres de la matrice A telles que $Re(\lambda_i) = 0$. La *variété stable* E^s est l'espace invariant engendré par les vecteurs propres associés aux valeurs propres de la matrice A telles que $Re(\lambda_i) < 0$ et la *variété instable* E^i est l'espace invariant engendré par les vecteurs propres associés aux valeurs propres de la matrice A telles que $Re(\lambda_i) > 0$. Si la variété centrale est réduite à l'ensemble vide, alors la variété stable est l'ensemble

$$E^s = \{x \in \mathbb{R}^n, \lim_{t \to +\infty} e^{tA} x = 0\}$$

La variété instable est obtenue de la même façon en remplaçant t par $-t$.

Pour un système non linéaire quelconque $\dot{x} = f(x)$, de flot intégral $\phi(t, x)$, un point stationnaire x et U un voisinage de x, on définit la variété locale stable

$$W^s = \{y \in U, \lim_{t \to +\infty} \phi(t, y) = x, \phi(t, y) \in U, \forall t \geq 0\}$$

et la variété locale instable par

$$W^i = \{y \in U, \lim_{t \to -\infty} \phi(t, y) = x, \phi(t, y) \in U, \forall t \leq 0\}$$

On démontre que si $x = 0$ est un point hyperbolique stationnaire de $\dot{x} = f(x)$, E^s et E^i les variétés stables et instables du système linéarisé $\dot{x} = Df(0)x$, alors il existe deux variétés locales stables et instables W^s et W^i de même dimension que E^s et E^i respectivement et tangentes à E^s et E^i à l'origine.

Exemple. Le système d'équations

$$\begin{cases} \dot{x} = x \\ \dot{y} = -y + x^2 \end{cases}$$

admet l'origine (0,0) comme point stationnaire. Le système linéarisé $\dot{x} = x$, $\dot{y} = -y$ admet pour variété stable l'ensemble des points $x = 0$, et pour variété instable l'ensemble des points $y = 0$. La variété locale est obtenue à partir d'un développement en série de

$$y(x) = A(x) = \sum_{i \geq 2} a_i x^i$$

en substituant dans l'équation

$$\dot{y} = -y + x^2 = -\sum_{i \geq 2} a_i x^i + x^2$$

et comme

$$\dot{y} = \dot{x} \frac{\partial A}{\partial x} = \sum_{i \geq 2} i a_i x^i$$

on obtient en égalant terme à terme les deux expressions précédentes $-a_2 + 1 = 2a_2$ et $-a_j = j a_j$ si $j \geq 3$, d'où $a_2 = 1/3$ et $a_j = 0$ pour $j \geq 3$. La variété locale instable est donc l'ensemble

$$W^i = \{(x, y),\ y = \frac{1}{3} x^3\}$$

On vérifie sur cet exemple que la variété locale est tangente à la variété E^i.

Les théorèmes de *Hartman-Grobman* conduisent à la classification des champs de vecteurs au voisinage d'un point stationnaire hyperbolique. Si $x = 0$ est un point hyperbolique stationnaire de $\dot{x} = f(x)$, alors il existe une application inversible continue h définie sur un voisinage de 0 (i.e. un homomorphisme local) qui prend localement les orbites du flot non linéaire pour celles du flot linéarisé $exp(tDf(0))$. On a équivalence topologique entre ces flots. Les points hyperboliques se conservent sous l'effet d'une perturbation. Si $x = 0$ est un point stationnaire hyperbolique de $\dot{x} = f(x)$, alors l'équation perturbée $\dot{x} = f(x) + \varepsilon u(x)$ où u est une fonction indéfiniment dérivable admet un point stationnaire hyperbolique au voisinage de l'origine du même type que le point hyperbolique du système non perturbé. Si le couple (m_s, m_i) désigne les dimensions de la variété stable et de la variété instable du système d'origine, c'est-à-dire le nombre, compté avec leur multiplicité, de valeurs propres dont la partie réelle est négative (resp. positive) du point stationnaire hyperbolique $x = 0$ du système $\dot{x} = f(x)$, alors le même couple correspond aux dimensions des variétés stables et instables du système perturbé.

7.12 Classification des flots bidimensionnels

Un flot bidimensionnel vérifie l'équation différentielle

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Désignons par A la matrice du système ci-dessus. Le polynôme caractéristique de ce système s'écrit

$$P(x) = x^2 - tr(A)x + \det(A)$$

Son discriminant est $\Delta = tr(A)^2 - 4\det(A)$. Les différentes possibilités conduisent à définir les cas suivants.
Cols. Si $det(A) < 0$, alors les valeurs propres de A sont réelles et de signes opposés et le flot est hyperbolique. On dit que les courbes intégrales forment un *col.* Notons $\lambda < 0 < \mu$, les valeurs propres de A. Les orbites suivent l'hyperbole $x = x_0(y/y_0)^{\mu/\lambda}$. L'axe $x = 0$ est la variété instable. L'axe $y = 0$ est la variété stable.
Nœuds. Si $det(A) > 0$ et $tr(A)^2 \geq 4\det(A)$, alors les valeurs propres de A sont réelles et de même signe. On dit qu'il s'agit d'un nœud qui est *attractif* ou *stable* si $tr(A) < 0$ et *répulsif* ou *instable* si $tr(A) > 0$. Le nœud est dit *propre* lorsque A est un multiple de l'identité, et *impropre* dans le cas contraire.
Foyers. Si $tr(A) \neq 0$ et si $tr(A)^2 < 4\det(A)$, alors les valeurs propres de A sont complexes de partie réelle $tr(A)/2$ et de partie imaginaire non nulle. On dit que les courbes forment un *foyer attractif* ou *stable* si $tr(A) < 0$ et *répulsif* ou *instable* si $tr(A) > 0$.
Centres. Si $det(A) < 0$ et $tr(A) = 0$, alors les valeurs propres de A sont imaginaires pures $\pm i\lambda$, $\lambda > 0$. Les courbes forment un *centre.* Les orbites sont les cercles de centre 0, périodiques de même période $2\pi/\lambda$.
Col-nœuds. Si $det(A) = 0$, alors A n'est pas inversible. On se trouve dans un cas dégénéré. Si $tr(A) \neq 0$ les valeurs propres sont 0 et λ, on parle de *nœud-col.* Les orbites sont l'axe $y = 0$ et les demi-droites $x = a$, $y > 0$ et $x = a$, $y < 0$. Si $tr(A) = 0$ et A non nulle, les orbites sont les points $(0, b)$ et les droites $x = a$.

7.13 Théorème de Poincaré-Bendixson

Deux sous-espaces vectoriels U et V d'un espace vectoriel E de dimension finie sont *transverses* si $U + V = E$ ou si

$$dim(U \cap V) = dim(U) + dim(V) - dim(E)$$

En général, la somme des dimensions des deux sous-espaces est inférieure à la dimension de E, mais lorsque cette somme est égale à la dimension de

E, dire que U et V sont transverses signifie simplement que U et V sont supplémentaires ($U \oplus V = E$). Deux sous-variétés U et V de E sont dites transverses si pour tout point a de l'intersection $U \cap V$ les sous-espaces vectoriels tangents T_aU et T_aV sont transverses. Dans l'espace usuel de dimension 3, un point est transverse à une surface s'il ne lui appartient pas. Deux courbes sont transverses si elles ne se rencontrent pas. Une courbe et une surface sont transverses si la courbe n'est nulle part tangente à la surface. Deux surfaces sont transverses si en leurs points communs leurs plans tangents sont distincts. Une sous-variété U de V est transverse en a à un champ de vecteurs X et a n'est pas un point singulier de X si le sous-espace tangent T_aU est supplémentaire à la droite $\mathbb{R}X(a)$ dans T_aV. Dans ce cas, U est de codimension 1 dans V en a. Une *transversale locale* (ou *section de Poincaré*) est un segment de droite coupé par des trajectoires dans le même sens.

Une démonstration du *théorème de Poincaré-Bendixson* repose sur un lemme qui affirme que si une trajectoire coupe une transversale locale plusieurs fois, alors les points d'intersection se déplacent sur la transversale. Le théorème de Poincaré-Bendixson affirme que si une trajectoire $\gamma(x_0)$ à partir d'un certain temps entre et reste dans une région D fermée bornée qui ne contient pas de point stationnaire, alors il existe au moins une orbite périodique dans D qui est un ensemble ω-limite de x_0.

Exemple 1. Considérons le système

$$\begin{cases} \dot{x} = -y - x(x^2 + y^2) \\ \dot{y} = x - y(x^2 + y^2) \end{cases}$$

qui s'écrit en coordonnées polaires $\dot{r} = -r(r-1)$ et $\dot{\theta} = 1$. Dans le domaine D compris entre les cercles de rayon $1/2$ et 2, il n'y a pas de point stationnaire (le seul point étant l'origine) et comme $\dot{r} > 0$ si $r < 1/2$ et $\dot{r} < 0$ si $r > 2$, toutes les trajectoires restent dans D. Le théorème de Poincaré affirme qu'il existe un cercle limite (qui est ici le cercle de rayon 1).

Exemple 2. Appliquons à l'oscillateur d'équation

$$\ddot{x} + f(x)\dot{x} + g(x) = 0$$

le théorème de Poincaré-Bendixson. Notons $F(x)$ une primitive de la fonction f

$$F(x) = \int_0^x f(u)du$$

et écrivons l'équation sous la forme du système

$$\begin{cases} \dot{x} = y - F(x) \\ \dot{y} = g(x) \end{cases}$$

Supposons que : (1) la quantité $xg(x) > 0$ pour x non nul, $g(0) = 0$, $f(0) < 0$ et $g'(0) > 0$, (2) $sgn(x)F(x) > k > 0$ pour $|x|$ suffisamment grand et (3) la fonction $G(x)$ primitive de $g(x)$ tend vers l'infini quand $|x| \to \infty$, alors, dans ces conditions, on démontre que le système admet au moins une orbite périodique.

7.14 Stabilité structurelle. Théorème de Peixoto

Deux champs de vecteurs f et g sur des variétés U et V sont *orbitalement équivalents* s'il existe un homéomorphisme h de U dans V qui transforme chaque orbite de f en une orbite de g, en conservant le sens de la paramétrisation temporelle.

$$\forall x, t_1, \exists t_2,\ h \circ \varphi_f(t_1, x) = \varphi_g(t_2, h(x))$$

Un champ de vecteurs f est *structurellement stable* si f et $f + \varepsilon u$ sont orbitalement équivalents pour tout ε appartenant à un intervalle arbitraire $[0, a]$ et u est une fonction quelconque. Autrement dit, le champ f est structurellement stable s'il existe un voisinage U de f dans l'ensemble des champs de vecteurs tel que tout élément de U soit orbitalement équivalent à f.

Une orbite périodique $u(t)$ du système $\dot{x} = f(x)$ de période T est dite *hyperbolique* si aucun des multiplicateurs de Floquet de l'équation $\dot{v} = Df(u(t))v$ n'est sur le cercle unité ($|\lambda_i| \neq 1$), sauf un, qui peut être égal à l'unité ($\lambda = 1$). Le *théorème de Peixoto* (1962) donne la classification des systèmes structurellement stables en dimension 2. Il affirme que si f est un champ de vecteurs sur une variété compacte orientée de dimension 2, alors f est structurellement stable si et seulement si

(1) tous les points stationnaires et orbites périodiques sont hyperboliques et en nombre fini

(2) pour tout couple (ω, ω') d'orbites fermées, les sous-variétés $W^s(\omega)$ et $W^i(\omega')$ sont transverses, en particulier, il n'existe pas d'orbite reliant des cols

(3) pour tout point x dont l'orbite n'est pas fermée, il existe un voisinage ouvert U de x et un réel T tel que $|t| \geq T$ entraîne que $\varphi(t, U) \cap U = \emptyset$. En particulier, il n'existe pas d'orbites récurrentes fermées, c'est-à-dire d'orbites fermées contenues dans l'un de ses ensembles limites.

Les champs de vecteurs vérifiant les deux premières conditions du théorème de Peixoto sont appelés *champs de Kupka-Smale* et ceux qui vérifient les trois conditions sont appelés *champs de Morse-Smale*. Les champs de Morse-Smale forment dans l'ensemble des champs de vecteurs sur une variété orientable une partie ouverte dense. La généralisation du théorème de Peixoto à des dimensions plus grandes se heurte à des difficultés liées à l'apparition du chaos. Anosov a construit en 1962 un difféomorphisme

structurellement stable du tore avec des points périodiques denses et Smale a donné en 1967 l'exemple du "fer à cheval". Ces deux exemples sont des systèmes structurellement stables en dimension supérieure à deux.

7.15 Bifurcations

Lorsque le champ de vecteurs associé à l'équation différentielle dépend d'un paramètre μ et que pour une valeur particulière de ce paramètre μ_0 le champ n'est pas un champ de Kupka-Smale, on parle de *bifurcation*. Ce type de situation se produit lorsqu'un point singulier n'est pas hyperbolique, lorsqu'une orbite périodique est non hyperbolique ou lorsqu'une variété stable et une variété instable se coupent non transversalement. Lorsqu'en un point stationnaire, un couple de valeurs propres conjuguées du champ linéarisé est purement imaginaire, on dit que la bifurcation est une *bifurcation de Hopf*. Lorsque ce champ possède une valeur propre nulle de multiplicité 1, toutes les autres étant de parties réelles non nulles, on parle de *bifurcation col-nœud*. Dans le cas d'orbites périodiques, la bifurcation est *supercritique* lorsque les orbites bifurquées sont stables et *sous-critique* lorsqu'il n'y a pas d'orbite périodique.

Exemple. Considérons le système

$$\begin{cases} \dot{x} = \mu x - \omega y - (x^2 + y^2)x \\ \dot{y} = \omega x + \mu y - (x^2 + y^2)y \end{cases}$$

L'équation caractéristique du système linéarisé

$$(\mu - \lambda + i\omega)(\mu - \lambda - i\omega) = 0$$

conduit à deux valeurs propres $\lambda = \mu \pm i\omega$. Si $\mu < 0$, l'origine est un foyer stable et si $\mu > 0$, l'origine est un foyer instable. Lorsque le paramètre $\mu = 0$, les valeurs propres sont de parties réelles nulles, une bifurcation est éventuellement possible. En écrivant le système en coordonnées polaires (r,θ)

$$\begin{cases} \dot{r} = \mu r - r^3 \\ \dot{\theta} = \omega \end{cases}$$

on constate que le système admet lorsque $\mu > 0$ une orbite périodique stable $r = \sqrt{\mu}$. Le système bifurque en l'origine de $\mu = 0$ vers $\mu > 0$. C'est une *bifurcation de Hopf supercritique.*

7.16 Système de Lorenz

Le *système de Lorenz* s'écrit

$$\left\{ \begin{array}{l} \dot{x} = \sigma(y - x) \\ \dot{y} = \rho x - y - xz \\ \dot{z} = xy - \beta z \end{array} \right.$$

où σ, ρ et β sont des paramètres positifs. Le système linéarisé

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} -\sigma & \sigma & 0 \\ \rho & -1 & 0 \\ 0 & 0 & -\beta \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

admet au plus trois points critiques. Si $0 < \rho < 1$, le système admet un seul point stationnaire (0,0,0) qui est asymptotiquement stable. Quand ρ passe la valeur 1, nous allons voir que nous avons affaire à une bifurcation. Si $\rho > 1$, le système admet trois points critiques $(0, 0, 0)$ ($\sqrt{\beta(\rho - 1)}$, $\sqrt{\beta(\rho - 1)}$, $\rho - 1$) et ($-\sqrt{\beta(\rho - 1)}$, $-\sqrt{\beta(\rho - 1)}$, $\rho - 1$). Le point $(0, 0, 0)$ admet trois valeurs propres réelles, deux négatives et une positive. Les valeurs propres des deux autres points stationnaires vérifient l'équation

$$\lambda^3 + (1 + \sigma + \beta)\lambda^2 + \beta(\rho + \sigma)\lambda + 2\sigma\beta(\rho - 1) = 0$$

Notons a la valeur

$$a = \frac{\sigma(\sigma + \beta + 3)}{\sigma - \beta - 1}$$

Lorsque $1 < \rho < a$, les trois valeurs propres ont leurs parties réelles négatives. Lorsque $\rho = a$, deux des valeurs propres sont purement imaginaires, on a alors une bifurcation de Hopf. Enfin, si $\rho > a$, une valeur propre est réelle négative et les deux autres sont complexes avec une partie réelle positive. Ces deux points critiques sont donc instables. Ce cas correspond au schéma classique de l'attracteur de Lorenz ($\beta = 8/3$, $\rho = 28$, $\sigma = 10$, $a \approx 24,74...$) qui a une orbite qui n'est pas fermée.

Pour le point critique $(0, 0, 0)$, le système admet trois valeurs propres $\lambda_1 = -\beta$, λ_2 et λ_3 qui vérifient l'équation $\lambda^2 + (1 + \sigma)\lambda + \sigma(1 - \rho) = 0$. Lorsque $\rho = 1$, ces trois valeurs propres se réduisent aux expressions $\lambda_1 = -\beta$, $\lambda_2 = 0$, et $\lambda_3 = -(1 + \sigma)$. L'espace propre associé à la valeur propre λ_1 est engendré par le vecteur propre de coordonnées $(0, 0, 1)$, celui de λ_2 est engendré par le vecteur propre de coordonnées $(1, 1, 0)$ et celui de λ_3 correspond au vecteur propre $(\sigma, -1, 0)$. La variété stable E^s est engendrée par les deux vecteurs propres associés aux deux valeurs propres négatives (λ_1 et λ_3). La variété centrale est engendrée par le vecteur propre de λ_2. La matrice de passage P formée des vecteurs propres permet de diagonaliser la partie linéaire du système par le changement de variables $X = PU$ où

$X = (x, y, z)$ et U est le nouveau vecteur $U = (u, v, w)$. Le système de Lorenz dans ces nouvelles coordonnées devient

$$\begin{pmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -(1+\sigma) & 0 \\ 0 & 0 & -\beta \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} + \begin{pmatrix} -\sigma(u+\sigma v)w/(1+\sigma) \\ \sigma(u+\sigma v)w/(1+\sigma) \\ (u-v)(u+\sigma v) \end{pmatrix}$$

Dans ces nouvelles coordonnées, la variété stable E^s est l'axe des u, la variété centrale E^c est engendrée par le vecteur $(0, v, w)$ et la variété instable est vide. La variété centrale W^c du système de Lorenz est tangente à la variété E^c du système linéarisé. C'est l'ensemble des points (u, v, w) tels que $(v, w) = (g(u), h(u))$. En développant en série formelle $v = a_2u^2 + a_3u^3 + a_4u^4 + ...$ et $w = b_2u^2 + b_3u^3 + b_4u^4 + ...$ et en reportant dans le système ci-dessus, on détermine les coefficients a_i et b_j. On trouve que la variété centrale est donnée au cinquième ordre par les relations

$$\begin{aligned} v &= \frac{u^3}{(1+\sigma)^2\beta} + o(u^5) \\ w &= \frac{1}{\beta}u^2 + \frac{2\sigma}{(1+\sigma)\beta^3} + o(u^5) \end{aligned}$$

En reportant ces valeurs dans l'équation $\dot{u} = -\sigma(u+\sigma v)w/(1+\sigma)$, on trouve que l'équation déterminant la variété centrale est l'équation différentielle

$$\dot{u} = \frac{-\sigma}{(1+\sigma)\beta}\left(u^3 + \frac{\sigma\beta + 2(1+\sigma)\beta}{2\beta(1+\sigma)^2}u^5\right) + o(u^6)$$

Le coefficient de u^3 étant négatif, le théorème de la variété centrale assure que l'origine est un point asymptotiquement stable.

Si maintenant le paramètre ρ est voisin de 1 ($\rho = 1 + \varepsilon$) on a, à l'ordre deux en ε, une expression approchée des valeurs propres $\lambda_2 \approx \lambda\varepsilon/(1+\sigma)$ et $\lambda_3 \approx -(1+\sigma) + \sigma\varepsilon/(1+\sigma)$. Si $\varepsilon < 0$, les trois valeurs propres ont leurs parties réelles non nulles. L'origine est un point stationnaire hyperbolique et stable. Si $\varepsilon > 0$, l'origine est un point stationnaire hyperbolique. C'est un col instable. Enfin, si $\varepsilon = 0$, l'origine n'est pas hyperbolique et une bifurcation se produit.

7.17 Méthodes d'Euler

La *méthode de Leonhard Euler* (1707-1783) est une méthode à pas séparé du premier ordre. Elle consiste à remplacer l'opérateur de dérivation d/dx par le schéma discret $(y_{i+1} - y_i)/h$. La résolution du problème

$$\begin{cases} y'(x) = f(x, y) \\ y(x_0) = y_0 \end{cases}$$

conduit au schéma

$$\begin{cases} x_{i+1} = x_i + h \\ y_{i+1} = y_i + hf(x_i, y_i) \end{cases}$$

En pratique, la méthode d'Euler n'est pas utilisée, car elle n'offre pas une précision suffisante. Cette méthode est convergente et du premier ordre, car l'erreur de consistance vaut

$$|y(x_i) - y_i| = \frac{1}{2}h^2 f'(c, y_i) \quad \text{avec } c \in [x_{i-1}, x_i]$$

Mais la méthode explicite est souvent instable. Par exemple, si la fonction f est linéaire $f(x, y) = -ay$ avec $a > 0$, le schéma d'Euler

$$y_{i+1} = y_i - ahy_i = (1 - ah)y_i$$

est instable dès que $h > 2/a$, car dans ce cas, y_i tend vers l'infini lorsque i tend vers l'infini. En revanche, le schéma rétrograde

$$y_{i+1} = y_i - hy'_{i+1} = \frac{y_i}{1 + ah}$$

conduit à une méthode implicite qui est universellement stable puisque y_{i+1} tend vers zéro, quand le pas h tend vers l'infini.

Dans la méthode des trapèzes, la fonction f est remplacée par une fonction affine par morceaux. Le schéma de discrétisation

$$y_{i+1} = y_i + \frac{h}{2}(y'_{i+1} + y'_i)$$

conduit à une méthode d'ordre 2.

7.18 Méthodes de Runge-Kutta

Carl Runge (1856-1927) et Martin Kutta (1867-1944) ont proposé en 1895 de résoudre le problème de Cauchy

$$\begin{cases} y'(x) = f(x, y) \\ y(x_0) = y_0 \end{cases}$$

en introduisant un schéma numérique de la forme

$$\begin{cases} x_{i+1} = x_i + h_i \\ y_{i+1} = y_i + h.\Phi(x_i, y_i, h_i) \end{cases}$$

où la fonction d'incrémentation Φ est une approximation de $f(x, y)$ sur l'intervalle $[x_i, x_{i+1}]$. Supposons donnés un entier r, une matrice A dont

les éléments triangulaires supérieurs sont nuls et un vecteur $b = (b_1, ..., b_r)$, l'algorithme de Runge-Kutta est le suivant

$$\begin{cases} y_{i+1} = y_i + h.(b_1k_1 + ...b_rk_r) \\ x_{i+1} = x_i + h \\ k_j = f(x_i + c_jh,\ y_i + h(a_{j1}k_1 + ... + a_{jr}k_r)) \end{cases}$$

Le vecteur b vérifie $b_1 + ... + b_r = 1$. Les coefficients c_i sont les sommes des éléments d'une ligne de la matrice A. Les éléments supérieurs de A sont nuls $a_{ij} = 0$ si $j \geq i$. Dans ces méthodes, le pas h peut facilement varier (h_i). Une méthode de Runge-Kutta est entièrement déterminée par la donnée de l'entier r, du vecteur b et de la matrice A.

Méthodes d'ordre 1. Pour $b = 1$ et $a_{11} = 0$, l'algorithme $y_{i+1} = y_i + h.f(x_i, y_i)$ se réduit à la méthode d'Euler. Pour $b = 1$ et $a_{11} = 1$, l'algorithme de Runge-Kutta $y_{i+1} = y_i + h.f(x_{i+1}, y_{i+1})$ conduit à la méthode d'Euler rétrograde (méthode implicite).

Méthodes d'ordre 2. Pour déterminer toutes les méthodes d'ordre 2, cherchons une fonction Φ de la forme

$$\Phi = b_1k_1 + b_2k_2$$

où les coefficients k_1 et k_2 sont donnés par

$$\begin{aligned} k_1 &= f(x_i, y_i) \\ k_2 &= f(x_i + ch, y_i + ahk_1) \end{aligned}$$

Développons y_{i+1} au voisinage du point (x_i, y_i),

$$y_{i+1} = y(x_i) + hf(x_i, y(x_i)) + \frac{h^2}{2}(\frac{\partial f}{\partial x}(x_i, y(x_i)) + f(x_i, y(x_i)))\frac{\partial f}{\partial y}(x_i, y(x_i))$$

De même, développons k_2 au voisinage de (x_i, y_i)

$$y_{i+1} = y_i + hb_1k_1 + hb_2(f(x_i, y_i) + hc\frac{\partial f}{\partial x}(x_i, y_i) + ahf(x_i, y_i)\frac{\partial f}{\partial y}(x_i, y_i)) + O(h^2)$$

En identifiant les deux expressions, il vient

$$\begin{cases} y(x_i) = y_i \\ f(x_i, y_i) = (b_1 + b_2)f(x_i, y_i) \\ \frac{1}{2}f'_x + \frac{1}{2}f'_yf = b_2(cf'_x + af'_yf) \end{cases}$$

On en déduit que $b_1 + b_2 = 1$ et $b_2c = b_2a = 1/2$. Soit en posant $b_2 = \theta$ et $b_1 = 1 - \theta$ et $c = a = 1/2$, on retrouve les trois cas standards. La méthode d'Euler s'obtient pour $\theta = 0$

$$y_{i+1} = y_i + hf(x_i, y_i)$$

La méthode de Heun (ou d'Euler-Cauchy) est obtenue pour $\theta = 1/2$

$$\begin{cases} y_{i+1} = y_i + h(k_1 + k_2)/2 \\ k_1 = f(x_i, y_i) \\ k_2 = f(x_i + h, y_i + hk_1) \end{cases}$$

La méthode de Runge-Kutta (proprement dite) est obtenue pour $\theta = 1$

$$\begin{cases} y_{i+1} = y_i + hk_2 \\ k_1 = f(x_i, y_i) \\ k_2 = f(x_i + h/2, y_i + hk_1) \end{cases}$$

Méthodes d'ordre 3. L'algorithme de Runge-Kutta classique correspond au cas $b = (1/6, 2/3, 1/6)$ et à la matrice

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ -1 & 2 & 0 \end{pmatrix}$$

L'algorithme effectue à chaque pas le calcul de trois facteurs k_i

$$\begin{cases} y_{i+1} = y_i + h(k_1 + 4k_2 + k_3)/6 \\ k_1 = f(x_i, y_i) \\ k_2 = f(x_i + h/2,\ y_i + hk_1/2) \\ k_3 = f(x_i + h,\ y_i - hk_1 + 2hk_2) \end{cases}$$

Pour améliorer l'efficacité du calcul, on utilise des méthodes à pas variable, c'est-à-dire des méthodes dans lesquelles le pas varie à chaque itération. Une des méthodes classiques consiste à employer deux *méthodes de Runge-Kutta emboîtées*. La première méthode d'ordre r sert à calculer la solution approchée, tandis que la seconde méthode d'ordre r' sert à estimer l'erreur de consistance pour contrôler le pas. On dit que la méthode est d'ordre (r', r). Proposée en 1957, la *méthode de Merson* est la première méthode de Runge-Kutta emboîtée. Elle consiste à calculer

$$\begin{cases} k_1 = f(x_i, y_i) \\ k_2 = f(x_i + h/3,\ y_i + hk_1/3) \\ k_3 = f(x_i + h/3,\ y_i - hk_1/6 + hk_2/6) \\ k_4 = f(x_i + h/2,\ y_i + hk_1/8 + 3hk_2/8) \\ k_5 = f(x_i + h,\ y_i + hk_1/2 - 3hk_3/2 + 2hk_1) \\ y_{i+1} = y_i + h(k_1 + 4k_2 + k_3)/6 \\ y^*_{i+1} = y_i + h(k_1 - 3k_3 + 4k_4)/2 \end{cases}$$

L'erreur

$$\Delta_i = \left| y_{i+1} - y^*_{i+1} \right|$$

est évaluée à chaque pas. Si ε désigne la tolérance acceptée, l'algorithme de Merson divise le pas par facteur 2 quand $\Delta_i > \varepsilon$, multiplie le pas par

deux quand $\Delta_i \leq \varepsilon/64$, et conserve le pas actuel dans les autres cas. La méthode de Merson est d'ordre 5 pour le calcul de la solution et d'ordre 4 pour le contrôle du pas. Dans la méthode proposée par Fehlberg en 1969, on calcule

$$\left\{\begin{array}{l} k_1 = f(x_i, y_i) \\ k_2 = f(x_i + h/4,\ y_i + hk_1/4) \\ k_3 = f(x_i + 3h/8,\ y_i + 3hk_1/32 + 9hk_2/32) \\ k_4 = f(x_i + 12h/13,\ y_i + 439hk_1/216 - 8hk_2 + \ldots \\ \qquad \ldots + 3680hk_3/513 - 845hk_4/4104) \\ k_5 = f(x_i + h/2,\ y_i - 8hk_1/27 + 2hk_2 - 3544hk_3/2565 + \ldots \\ \qquad \ldots + 1859hk_4/4104 - 11hk_5/40) \\ y_{i+1} = y_i + h(25k_1/216 + 1408k_3/2565 + 2197k_4/4104 - k_5/5) \\ y^*_{i+1} = y_i + h(16k_1/135 + 6656k_3/12825 + 28561k_4/56430 + \ldots \\ \qquad \ldots - 9k_5/50 + 2k_6/55) \end{array}\right.$$

D'autres méthodes d'ordre plus élevé ont été proposées. Dormand et Prince [Dormand, 1980] ont proposé une méthode d'ordre (7,8). Plus récemment, Cash et Karp [Cash, 1990] ont proposé une méthode d'ordre (4, 5). La difficulté de ce genre d'algorithmes réside dans le fait d'ajuster au mieux les coefficients des deux méthodes. Dans la méthode de Cash-Karp, on calcule

$$\left\{\begin{array}{l} k_1 = hf(x_i, y_i) \\ k_2 = hf(x_i + h/5,\ y_i + k_1/5) \\ k_3 = hf(x_i + 3h/10,\ y_i + 3k_1/40 + 9k_2/40) \\ k_4 = hf(x_i + 3h/5,\ y_i + 3h/10 - 9k_2/10 + 6k_3/5) \\ k_5 = hf(x_i + h,\ y_i - 11k_1/54 + 5k_2/2 - 70k_3/27 + 35k_4/27) \\ k_6 = hf(x_i + 7h/8,\ y_i + 1631k_1/55296 + 175k_2/512 + 575k_3/13824 + ... \\ \qquad \ldots + 44275k_4/110592 + 253k_5/4096) \\ y_{i+1} = y_i + (37k_1/378 + 250k_3/621 + 125k_4/594 + 512k_6/1771) \\ y^*_{i+1} = y_i + (2825k_1/27648 + 18575k_3/48384 + 13525k_4/55296 + +k_6/4) \end{array}\right.$$

Si on pose

$$y_{i+1} = y_i + c_1k_1 + ... + c_6k_6 + O(h^6)$$

et

$$y^*_{i+1} = y_i + c^*_1k_1 + ... + c^*_6k_6 + O(h^6)$$

Une estimation de l'erreur est donnée par

$$\Delta = y_{i+1} - y^*_{i+1} = \sum_{i=1}^{6}(c_i - c^*_i)k_i$$

7.19 Méthode de Newmark

La *méthode de Newmark* est une méthode très utilisée dans les codes de dynamique. C'est une méthode de résolution directe qui s'applique à

l'équation matricielle

$$M\ddot{x} + C\dot{x} + Kx = F$$

où M est la matrice de masse, C la matrice d'amortissement, K la matrice de rigidité, F la force généralisée. La solution est une fonction $x(t)$ dépendante du temps. Le schéma de Newmark se présente sous la forme

$$\left\{ \begin{array}{l} x_{i+1} = x_i + h\dot{x}_i + h^2\left((1/2-\beta)\ddot{x}_i + \beta\ddot{x}_{i+1}\right) \\ \dot{x}_{i+1} = \dot{x}_i + h\left((1-\gamma)\ddot{x}_i + \gamma\ddot{x}_{i+1}\right) \end{array} \right.$$

où β et γ sont deux paramètres. Lorsque ces deux paramètres sont nuls, on retrouve les formules de Taylor. Lorsque $\beta = 1/12$ et $\gamma = 1/2$, la méthode s'appelle *méthode de Fox-Goodwin*. On démontre que la méthode de Newmark est d'ordre 1 pour $\gamma \neq 1/2$ et d'ordre 2 pour $\gamma = 1/2$. La discrétisation de l'équation de la dynamique s'écrit

$$\begin{array}{rcl} (M + h\gamma C + h^2\beta K)\ddot{x}_{i+1} & = & F_{i+1} - C\left(\dot{x}_i + (1-\gamma)h\ddot{x}_i\right) \\ & & -K\left(x_i + h\dot{x}_i + h^2(1/2-\beta)\ddot{x}_i\right) \end{array}$$

La résolution de ce système associée aux prédictions des vitesses et des déplacements conduit aux valeurs de l'accélération $\ddot{x}_{i+1}$.

Une amélioration de ce schéma a été proposée par Hilbert, Hugues et Taylor et est connue sous le nom de *méthode HHT* ou *α-HHT*. Elle consiste à introduire un paramètre α dans l'équation de la dynamique

$$M\ddot{x}_{i+1} + (1+\alpha)C\dot{x}_{i+1} - \alpha C\dot{x}_i + (1+\alpha)Kx_{i+1} - \alpha Kx_i = (1+\alpha)F_{i+1} - \alpha F_i$$

Les valeurs $\ddot{x}_0$, $\dot{x}_0$ et $x_0 = Mh(F_0 - C\dot{x}_0 - Kx_0)$ initialisent l'algorithme. En général, on choisit $\beta = (1-\alpha)^2/4$ et $\gamma = 1/2 - \alpha$. Ce schéma est inconditionnellement stable pour $-1/3 \leq \alpha \leq 0$. Notons enfin que d'autres codes de mécanique utilisent la *méthode de Gear*, implicite à deux pas, appelée aussi *Backward Differential Formulas*, qui se définit par le schéma

$$\left\{ \begin{array}{l} \dot{x}_{i+1} = \left(3x_{i+1} - 4x_i + x_{i-1}\right)/2h \\ \ddot{x}_{i+1} = \left(3\dot{x}_{i+1} - 4\dot{x}_i + \dot{x}_{i-1}\right)/2h \end{array} \right.$$

7.20 Méthodes d'Adams

Dans les méthodes proposées par John Adams (1819-1892) en 1855, la fonction f est approchée par son polynôme d'interpolation.

$$\begin{array}{rcl} y(x_{i+1}) & = & y(x_i) + \displaystyle\int_{x_i}^{x_{i+1}} f(t, y(t))dt \\ & \simeq & y(x_i) + \displaystyle\int_{x_i}^{x_{i+1}} p_{n,r}(t)dt \end{array}$$

Dans les *méthodes d'Adams-Bashforth*, $p_{n,r}(x)$ est le polynôme d'interpolation de f aux points $x_{i-r}, x_{i-r+1}, ..., x_i$. Si on note les pentes $f_i = f(x_i, y_i)$, l'algorithme de la méthode d'Adams-Bashforth à $(r+1)$ pas, s'écrit

$$\begin{cases} y_{i+1} = y_i + h_i.(a_0 f_i + a_1 f_{i-1} + ... + a_r f_{i-r}) \\ x_{i+1} = x_i + h_i \end{cases}$$

Les coefficients a_j sont tabulés, par exemple à l'ordre $r = 3$, on a a_0 =55/24, $a_1 = -59/24$, $a_2 = 37/24$, $a_3 = -9/24$. L'erreur commise est égale à

$$c_r.h^{r+2}.y^{(r+2)}(\xi)$$

où $\xi \in [x_{i-r}, x_{i+1}]$. À l'ordre 3, le coefficient c_r vaut 251/720. La méthode d'Adams-Bashforth est une méthode explicite. Les formules donnant l'expression de y_{i+1} sont parfois appelées *formules d'Adams ouvertes*. La méthode d'Adams-Bashforth à $(r+1)$ pas est une méthode stable (si f est lipschitzienne) et d'ordre $(r+1)$.

Les *méthodes de Milne explicites* sont fondées sur le même principe que les méthodes d'Adams-Bashforth mais ici, le schéma donnant y_{i+1} est exprimé en fonction de y_{i-r}.

$$\begin{cases} y_{i+1} = y_{i-r} + h_i.(a_0 f_i + a_1 f_{i-1} + ... + a_r f_{i-r}) \\ x_{i+1} = x_i + h_i \end{cases}$$

Les coefficients a_j sont tabulés. Par exemple à l'ordre $r = 3$, on a a_0 =8/3, $a_1 = -4/3$, $a_2 = 8/3$, $a_3 = 0$. L'erreur commise est égale à

$$c_r.h^{r+2}.y^{(r+2)}(\xi)$$

où $\xi \in [x_{i-r}, x_{i+1}]$. À l'ordre 3, le coefficient c_r vaut 14/45.

Dans les *méthodes d'Adams-Moulton*, la fonction f est approchée par son polynôme d'interpolation aux points $x_{i-r}, ..., x_i, x_{i+1}$. La méthode est identique à la méthode d'Adams-Bashforth, mais ici le point x_{i+1} est pris en plus. De ce fait, la méthode d'Adams-Moulton est une méthode implicite : à chaque pas de calcul, on suppose connues les valeurs $f_{i-r}, ...f_i, f_{i+1}$.

$$\begin{cases} y_{i+1} = y_i + h_i.(a_0 f_{i+1} + a_1 f_i + ... + a_r f_{i-r}) \\ x_{i+1} = x_i + h_i \end{cases}$$

Les coefficients a_j sont tabulés. Par exemple à l'ordre 3, on a $a_0 = 251/720$, $a_1 = 646/720$, $a_2 = -264/720$, $a_3 = 106/720$, $a_4 = -19/720$. L'erreur commise est égale à

$$c_r.h^{r+3}.y^{(r+3)}(\xi)$$

où $\xi \in [x_{i-r}, x_{i+1}]$. À l'ordre 3, le coefficient c_r vaut $-3/160$. La méthode d'Adams-Moulton est une méthode implicite. Les formules donnant

l'expression de y_{i+1} sont parfois appelées *formules d'Adams fermées*. La méthode d'Adams-Moulton à $(r+1)$ pas est une méthode stable (si f est lipschitzienne) et d'ordre $(r+2)$.

Les *méthodes de Milne implicites* sont fondées sur le même principe que les méthodes d'Adams-Moulton mais ici, le schéma donnant y_{i+1} est exprimé en fonction de y_{i-r}.

$$\begin{cases} y_{i+1} = y_{i-r} + h_i.(a_0 f_{i+1} + a_1 f_i + ... + a_r f_{i-r}) \\ x_{i+1} = x_i + h_i \end{cases}$$

Les coefficients a_j sont tabulés. Par exemple à l'ordre $r = 3$, on a a_0 =14/45, a_1 =64/45, a_2 =24/45, a_3 =64/45, et a_4=14/45. L'erreur commise est égale à

$$c_r.h^{r+3}.y^{(r+3)}(\xi)$$

où $\xi \in [x_{i-r}, x_{i+1}]$. À l'ordre 3, le coefficient c_r vaut $-8/945$.

7.21 Méthodes de Rosenbrock

Les *méthodes de Rosenbrock* aussi appelées *méthodes de Kaps-Rentrop*, ont l'avantage de pouvoir traiter des systèmes raides et sont compétitives avec d'autres algorithmes plus compliqués lorsque la taille du système n'est pas trop élevée ($N < 10$). Considérons l'équation

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

Les méthodes de Rosenbrock à s pas se définissent par le schéma

$$y_{n+1} = y_n + c_1 k_1 + ... + c_s k_s$$

où

$$\begin{aligned} k_i &= hf\left(t_n + \alpha_i h, y_n + \sum_{j=1}^{i-1} \alpha_{ij} k_j\right) + \gamma_i h^2 \frac{\partial f}{\partial t}(t_n, y_n) + \\ &\quad + h \frac{\partial f}{\partial y}(t_n, y_n) \sum_{j=1}^{i} \gamma_{ij} k_j \end{aligned}$$

Les coefficients b_i, $\alpha_{i,j}$ et $\gamma_{i,j}$ sont donnés pour chaque schéma et

$$\alpha_i = \sum_{j=1}^{i-1} \alpha_{ij} \text{ et } \gamma_i = \sum_{j=1}^{i} \gamma_{ij}$$

Lorsque le système ne dépend pas du temps, l'équation $y' = f(y)$ est approchée par

$$y_{n+1} \approx y_n + h\left(f(y_n) + \frac{\partial f}{\partial t}(y_n)(y_{n+1} - y_n)\right)$$

ce qui conduit à la résolution d'un système de la forme

$$\left(I - h\frac{\partial f}{\partial y}(y_n)\right)(y_{n+1} - y_n) = hf(y_n)$$

appelée *méthode semi-implicite d'Euler*. En paramétrant ces équations, Rosenbrock propose le schéma suivant

$$\left(I - h\gamma_{ii}\frac{\partial f}{\partial y}(y_n)\right)k_i = hf\left(y_n + \sum_{j=1}^{i-1}\alpha_{ij}k_j\right) + h\frac{\partial f}{\partial y}(y_n)\sum_{j=1}^{i}\gamma_{ij}k_j$$

Kaps-Rentrop proposent de choisir pour coefficients $\gamma_{ii} = \gamma$, ce qui permet d'inverser facilement la matrice $(I - h\gamma\partial f(y_n)/\partial y)$ et de calculer les coefficients k_i.
Pour une méthode à $s = 2$ pas, on prendra

$$y_{i+1} = y_i + \frac{1}{2}(k_1 + k_2)$$

avec

$$\begin{aligned}(1 - \gamma hJ)k_1 &= f(y_i)\\ (1 - \gamma hJ)k_2 &= f(y_i + hk_1) - 2\gamma hJk_1\end{aligned}$$

J étant une approximation du jacobien de f et $\gamma = 1 + 1/\sqrt{2}$.
Pour une méthode à $s = 4$ pas, Kaps-Rentrop proposent de poser

$$g_i = \sum_{j=1}^{i-1}\gamma_{ij}k_j + \gamma k_i$$

et de résoudre le système

$$\begin{cases}(1/\gamma h - J_n)g_1 = f(y_n)\\ (1/\gamma h - J_n)g_2 = f(y_n + 2g_1) - 8g_1/h\\ (1/\gamma h - J_n)g_3 = f(y_n + (48g_1 + 6g_2)/25) + (372g_1 + 60g_2)/25h\\ (1/\gamma h - J_n)g_4 = f(y_n + (48g_1 + 6g_2)/25) - (112g_1 + 54g_2 + 50g_3)/125h\\ y_{n+1} = y_n + 19g_1/9 + g_2/2 + 25g_3/108 + 125g_4/108\end{cases}$$

où J_n est une approximation du jacobien. Une estimation de l'erreur de cette méthode est donnée par la différence entre les valeurs de y_{n+1} à l'ordre $s = 4$ et à l'ordre $s = 3$,

$$\Delta = y_{n+1}^{(4)} - y_{n+1}^{(3)} = \frac{17}{54}g_1 + \frac{7}{36}g_2 + \frac{125}{108}g_4$$

7.22 Méthodes de prédiction-correction

Dans les méthodes de prédiction-correction (PECE), un prédicteur Φ_1 fournit une première valeur approchée $\tilde{y}_{i+1}$ de y_{i+1} à partir de la connaissance de $y_{i-r}, ..., y_i, f_{i-r}, ..., f_i$. Cette valeur approchée est utilisée pour évaluer une approximation $\tilde{f}_{i+1}$ de $f(x_{i+1}, \tilde{y}_{i+1})$. Une nouvelle formule de y_{i+1} donne une valeur corrigée en utilisant la valeur approchée précédemment calculée de $\tilde{f}_{i+1}$. Pour démarrer l'algorithme, comme les premiers termes ne sont pas connus, on utilise en général une méthode de Runge-Kutta. Les méthodes de prédiction-correction, aussi appelées *méthodes PECE* se composent de quatre équations Prédiction (P), Évaluation (E), Correction (C) et de nouveau Évaluation (E).

$$\begin{cases} \text{Prédiction} & \tilde{y}_{i+1} = \Phi_1(y_{i-r}, ..., y_i, f_{i-r}, ..., f_i) \\ \text{Évaluation} & \tilde{f}_{i+1} = f(x_{i+1}, \tilde{y}_{i+1}) \\ \text{Correction} & y_{i+1} = \Phi_2(y_{i-r}, ..., y_i, f_{i-r}, ..., f_i, \tilde{f}_{i+1}) \\ \text{Évaluation} & f_{i+1} = f(x_{i+1}, y_{i+1}) \end{cases}$$

Afin de gagner du temps, la dernière évaluation est parfois omise : c'est la méthode PEC. Dans les méthodes d'Adams, le prédicteur est une méthode d'Adams-Bashforth d'ordre $(r+1)$ et le correcteur utilise une formule d'Adams-Moulton d'ordre $(r+2)$. Par exemple, à l'ordre 4, on emploiera

$$\begin{cases} \tilde{y}_{i+1} = y_i + \frac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}) \\ \tilde{f}_{i+1} = f(x_{i+1}, \tilde{y}_{i+1}) \\ y_{i+1} = y_i + \frac{h}{24}(9\tilde{f}_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}) \\ f_{i+1} = f(x_{i+1}, y_{i+1}) \end{cases}$$

La méthode de Milne d'ordre 4 est illustrée ci-après

$$\begin{cases} \tilde{y}_{i+1} = y_{i-3} + \frac{4h}{3}(2f_i - f_{i-1} + 2f_{i-2}) \\ \tilde{f}_{i+1} = f(x_{i+1}, \tilde{y}_{i+1}) \\ y_{i+1} = y_i + \frac{h}{3}(\tilde{f}_{i+1} + 4f_i + f_{i-1}) \\ f_{i+1} = f(x_{i+1}, y_{i+1}) \end{cases}$$

À l'ordre 6, la méthode de Milne est donnée par les schémas suivants

$$\begin{cases} \tilde{y}_{i+1} = y_{i-5} + \frac{3h}{10}(11f_i - 14f_{i-1} + 26f_{i-2} - 14f_{i-3} + 11f_{i-4}) \\ \tilde{f}_{i+1} = f(x_{i+1}, \tilde{y}_{i+1}) \\ y_{i+1} = y_i + \frac{2h}{45}(7\tilde{f}_{i+1} + 32f_i + 12f_{i-1} + 32f_{i-2} + 7f_{i-3}) \\ f_{i+1} = f(x_{i+1}, y_{i+1}) \end{cases}$$

7.23 Exercices

1. *Équation intégro-différentielle.* Soit f une fonction définie sur un intervalle $I = [t_0, t_0 + T]$, $T > 0$ satisfaisant la condition de Lipschitz :

$$\forall t \in I, \quad \forall x, y \in \mathbb{R} \quad |f(t,x) - f(t,y)| \leq L|x - y|$$

et soit k une fonction de $I \times I \to \mathbb{R}$. On considère le problème intégro-différentiel suivant : Trouver une fonction y de classe C^1 telle que

$$\begin{cases} y'(t) = f(t, y(t)) + \int_{t_0}^{t} k(t,s)y(s)ds \\ y(t_0) = y0 \end{cases}$$

1) Vérifier que ce problème est équivalent au problème suivant : Trouver une fonction y de classe C^1 telle que

$$y(t) = y_0 + \int_{t_0}^{t} (f(u, y(u)) + \int_{t_0}^{u} k(u,s)y(s)ds)\, du$$

2) On considère la subdivision $t_i = t_0 + ih$ pour $i = 0, 1, \ldots, N$ et $h = T/N$, N étant fixé et l'approximation y_n de $y(t_n)$ donnée pour $n = 0, 1, \ldots, N-1$ par

$$y_{n+1} = y_n + hf(t_n, y_n) + h^2 \sum_{i=0}^{n-1} k(t_n, t_i)y_i$$

les sommes sur les indices négatifs étant nulles. Montrer que

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} (y'(t) - y'(t_n))\, dt + hy'(t_n)$$

En déduire que si on pose

$$s_1(y', h) = max\, |y'(t) - y'(s)|$$

le maximum étant pris sur les couples (s,t) tels que $|t-s| \leq h$ et

$$s_2(ky, h) = max\, |k(t,s)y(s) - k(t,u)y(u)|$$

le maximum étant pris sur les couples (s,u) tels que $|s-u| \leq h$, on peut majorer la quantité

$$\epsilon_n = y(t_{n+1}) - y(t_n) - hf(t_n, y(t_n)) - h^2 \sum_{i=0}^{n-1} k(t_n, t_i)y(t_i)$$

par

$$\epsilon_n \leq hs_1 + nh^2 s_2$$

En déduire que si y est de classe C^1, la somme

$$\sum_{i=0}^{N-1} |\epsilon_i|$$

tend vers zéro lorsque N tend vers l'infini.
3) On pose $e_n = |y(t_n) - y_n|$, $K^2 = max|k(s,t)|$ le maximum étant

pris sur tous les couples (s,t) de l'intervalle $I \times I$ et on définit les quantités :

$$\beta_n = Kh \sum_{i=0}^{n-1} e_i$$

Montrer que

$$e_{n+1} \leq (1+hL)e_n + Kh\beta_n + |\epsilon_n|$$

et

$$\beta_{n+1} \leq Khe_n + \beta_n$$

4) En admettant que

$$e_n \leq \sum_{i=0}^{n-1} e^{M(t_n - t_{i+1})} |\epsilon_i|$$

où $M = L + K$, démontrer que l'on peut estimer l'erreur par

$$|y(t_n) - y_n| \leq \frac{e^{MT}-1}{M} s_1 + \frac{e^{MT}-1-MT}{M^2} s_2$$

2. On considère le système

$$\begin{cases} x' = 2y(z-1) \\ y' = -x(z-1) \\ z' = xy \end{cases}$$

Déterminer les points stationnaires et étudier la stabilité, au sens de Lyapunov, du système en ces points.

3. On considère le système

$$\begin{cases} x' = xy \\ y' = -y + 3x^2 \end{cases}$$

Étudier la stabilité de Lyapunov du système au point $(0,0)$.

4. Soit a, b des paramètres positifs, on considère le système appelé *brusselator*

$$\begin{cases} x' = a - (b+1)x + x^2 y \\ y' = bx - x^2 y \end{cases}$$

défini pour $x, y \geq 0$. Étudier la possibilité pour que le *brusselator* présente une bifurcation de Hopf.

8
Équations aux dérivées partielles

Ce chapitre traite des problèmes théoriques liés à la résolution d'équations aux dérivées partielles. La notion de solution forte ou faible conduit à considérer différents aspects du problème. On rappelle les principaux résultats concernant les distributions et les opérateurs pseudo-différentiels et on introduit les espaces de Sobolev. Au plan numérique on étudie les *méthodes de différences finies*. Dans ces méthodes, chaque dérivée est approchée par une expression discrétisée. L'équation différentielle est ainsi remplacée par une expression discrète appelée *schéma numérique*. L'étude porte alors sur les conditions de convergence de ces schémas vers la solution exacte et de leur stabilité. Les erreurs accumulées au fil du calcul pouvant conduire à une solution numérique qui s'éloigne progressivement de la solution exacte.

8.1 Problèmes aux limites

Soit Ω un domaine de $\mathbb{R}^n$ et le $\partial\Omega$ bord de ce domaine. On considère un opérateur différentiel $\mathcal{L}$ et l'équation

$$\mathcal{L}u(x,t) = f(x,t)$$

Pour résoudre cette équation dans laquelle u est l'inconnue et f une donnée sur $\Omega \times \mathbb{R}$, on lui adjoint des conditions aux limites. Dans le *problème de Dirichlet* ou *premier problème aux limites* on cherche une solution de l'équation qui prend des valeurs données sur le bord de Ω. On cherche donc

à résoudre le système

$$\begin{cases} \mathcal{L}u = f & x \in \Omega \\ u = g & x \in \partial\ \Omega \end{cases}$$

En général la fonction g est (au moins) continue. Dans le *problème de Neumann* ou *deuxième problème aux limites*, on cherche une solution de l'équation différentielle dont on connaît la valeur du gradient sur le bord du domaine de résolution. Notant n la normale unitaire dirigée vers l'extérieur de Ω, on cherche donc à résoudre le problème

$$\begin{cases} \mathcal{L}u = f & x \in \Omega \\ D_n u = g & x \in \partial\ \Omega \end{cases}$$

expression dans laquelle on a noté $D_n = \partial u/\partial n = \nabla u.n$. Dans le *problème de Dirichlet-Neumann* ou *troisième problème aux limites*, on cherche une fonction qui vérifie la troisième condition au bord

$$\begin{cases} \mathcal{L}u = f & x \in \Omega \\ D_n u + au = g & x \in \partial\ \Omega \end{cases}$$

où a est une fonction de x. Enfin, une équation de la forme

$$\mathcal{L}u + \lambda u = 0$$

est un *problème aux valeurs propres*. Déterminer les solutions de ce type d'équation revient à déterminer les vecteurs propres de l'opérateur $\mathcal{L}$. Sous certaines conditions, on démontre que chaque problème admet une solution unique. On doit alors préciser ce qu'on entend par "solution", car u peut être une fonction différentiable (solution forte) ou une distribution (solution faible).

8.2 Espaces de Lebesgue

L'espace de Lebesgue L^p $(1 \leq p < \infty)$ est défini comme l'espace quotient de $\mathcal{L}^p$ par la relation d'équivalence des fonctions égales presque partout. La norme

$$\|f\|_p = \left(\int |f|^p \, d\mu(x) \right)^{1/p}$$

permet de définir l'espace de Lebesgue comme l'ensemble des fonctions de norme finie

$$L^p = \{u,\ \|u\|_p < \infty\}$$

L'espace L^2 est l'espace des fonctions de carré intégrable. Le dual (topologique) de l'espace L^p (i.e. l'ensemble des applications linéaires de L^p dans

$\mathbb{R}$) est l'espace L^q avec $1/p + 1/q = 1$. L'espace L^p est un espace complet. Si Ω est un ouvert de $\mathbb{R}^n$, l'espace $\mathcal{D}(\Omega)$ des fonctions C^∞ à support compact contenu dans Ω est dense dans l'espace L^p, c'est-à-dire que toute fonction de p-norme finie est approchable par une suite de fonctions régulières. Notons enfin que le dual de l'espace $L^1(\Omega, \mathbb{R})$ est l'espace $L^\infty(\Omega, \mathbb{R})$ des fonctions essentiellement bornées.

8.3 Distributions

Soit Ω un ouvert de $\mathbb{R}^n$. On définit pour un entier k, l'espace $C^k(\Omega)$ comme l'espace des fonctions continues dont toutes les dérivées jusqu'à l'ordre k sont continues sur Ω. L'ensemble des fonctions indéfiniment dérivables est l'intersection des ensembles $C^k(\Omega)$, et on note

$$C^\infty(\Omega) = \bigcap_{k=0}^{\infty} C^k(\Omega)$$

L'ensemble $\mathcal{D}(\Omega)$ est l'espace des fonctions C^∞ à support compact contenu dans Ω, $\mathcal{D}(\Omega) = C_0^\infty(\Omega)$. On note $\alpha = (\alpha_1, ..., \alpha_n)$ un multi-indice, $\alpha_j \in \mathbb{N}$ et

$$D^\alpha = \partial_1^{\alpha_1} .... \partial_n^{\alpha_n} \quad \text{avec} \quad \partial_j = \frac{\partial}{\partial x_j} \quad D_j = -i\partial_j$$

l'opérateur de différentiation, $|\alpha| = \alpha_1 + ... + \alpha_n$ est la longueur du multi-indice. L'ensemble $C^{k,\alpha}(\Omega)$ pour $0 < \alpha \leq 1$ est l'ensemble des fonctions de $C^k(\Omega)$ telles que toutes les dérivées d'ordre k sont höldériennes d'ordre k au voisinage de tout point de Ω.

Une *distribution* T à valeurs réelles ou complexes est une application de $\mathcal{D}$ dans $\mathbb{R}$ ou $\mathbb{C}$, $\varphi \to \langle T, \varphi \rangle$ linéaire et continue au sens où si on désigne par φ_n une suite de fonctions non nulles en dehors d'un ensemble borné, telles que si la suite des dérivées d'ordre k de φ_n $(k = 0, 1, ...)$ converge uniformément vers la dérivée de φ d'ordre k, alors $\langle T, \varphi_n \rangle$ converge vers $\langle T, \varphi \rangle$. L'ensemble des distributions forme un espace vectoriel noté $\mathcal{D}'$. On appelle *distribution régulière* une distribution associée à une fonction f localement sommable sur $\mathbb{R}^n$, encore notée T_f ou plus simplement f et définie par

$$\langle f, \varphi \rangle = \int f(x)\varphi(x)dx$$

On définit de la même manière une distribution associée à une mesure μ

$$\langle \mu, \varphi \rangle = \int \varphi(x)d\mu(x)$$

Ainsi la distribution de Dirac au point a est définie par

$$\langle \delta_a, \varphi \rangle = \int_{-\infty}^{\infty} \delta(x-a)\varphi(x)dx = \varphi(a)$$

Le produit d'une distribution $T \in \mathcal{D}'(\Omega)$ par une fonction $a \in C^\infty(\Omega)$ est la distribution aT définie pour tout $\varphi \in \mathcal{D}(\Omega)$, par

$$\langle aT, \varphi \rangle = \langle T, a\varphi \rangle$$

La dérivée d'une distribution est définie par l'expression

$$\left\langle \frac{\partial T}{\partial x_i}, \varphi \right\rangle = -\left\langle T, \frac{\partial \varphi}{\partial x_i} \right\rangle$$

En ce sens, toute distribution est indéfiniment dérivable. Un opérateur différentiel d'ordre m est défini par

$$P(x, D) = \sum_{|\alpha| \leq m} a_\alpha(x) D^\alpha$$

où x est un point de $\mathbb{R}^n$, a_α des fonctions de $\mathbb{R}^n$ sur $\mathbb{R}$ et α un multi-indice de $\mathbb{N}^n$.

Le *symbole principal* de l'opérateur P est défini par

$$\sigma(x, \xi) = \sum_{|\alpha|=m} a_\alpha(x).\xi^\alpha$$

Si P est un opérateur différentiel, une solution élémentaire (ou fondamentale) E est une solution de l'équation $PE = \delta$. Soit $\mathcal{S}(\mathbb{R}^n)$ l'espace des fonctions u de classe C^∞ sur $\mathbb{R}^n$ telles que pour tout α, β de $\mathbb{N}^n$,

$$\lim_{|x| \to \infty} \left| x^\alpha \partial^\beta u(x) \right| = 0$$

L'espace $\mathcal{S}'(\mathbb{R}^n)$ des *distributions tempérées* est le dual topologique de $\mathcal{S}(\mathbb{R}^n)$, c'est-à-dire l'espace des formes linéaires continues sur $\mathcal{S}(\mathbb{R}^n)$. Une distribution T appartient à $\mathcal{S}'(\mathbb{R}^n)$ si et seulement si il existe un entier naturel m, un multi-indice α et une fonction continue bornée sur $\mathbb{R}^n$ tels que

$$T = \partial^\alpha [(1 + |x|^2)^m f]$$

Pour une fonction $f \in L^1(\mathbb{R}^n)$, on note $\mathcal{F}f$ ou encore $\widehat{f}$ la transformée de Fourier

$$\widehat{f}(\xi) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-ix.\xi} f(x)dx$$

Dans cette expression $x.\xi$ désigne le produit scalaire des vecteurs x et ξ. La transformée de Fourier inverse est alors l'application

$$f(x) = \int_{\mathbb{R}^n} e^{ix.\xi} \widehat{f}(\xi) d\xi$$

Pour une distribution tempérée $T \in \mathcal{S}'(\mathbb{R}^n)$, on note $\mathcal{F}T$ ou encore $\widehat{T}$ la transformée de Fourier, définie par la relation, pour tout $\varphi \in \mathcal{S}(\mathbb{R}^n)$

$$\langle \mathcal{F}T, \varphi \rangle = \langle T, \mathcal{F}\varphi \rangle$$

8.4 Opérateurs pseudo-différentiels

Soit m un réel, notons S^m l'ensemble des fonctions $\sigma(x,\xi) \in C_0^\infty(\mathbb{R}^n \times \mathbb{R}^n)$ tel que pour tout multi-indice α et β, il existe une constante C qui dépend de ces multi-indices vérifiant

$$\left| D_x^\alpha D_\xi^\beta \ \sigma(x,\xi) \right| \leq C(1+|\xi|)^{m-|\beta|}$$

Toute fonction pour laquelle l'inégalité précédente est vraie pour toute valeur de m est appelée un *symbole*. Nous considérerons ici des symboles simples comme la fonction

$$\sigma(x,\xi) = \sum_{|\alpha|=m} a_\alpha(x).\xi^\alpha$$

dont les coefficients appartiennent à l'ensemble $C_0^\infty(\mathbb{R}^n)$, ensemble des fonctions indéfiniment différentiables à support compact. Un *opérateur pseudo-différentiel* sur un ouvert Ω de $\mathbb{R}^n$ est un opérateur $P : C_0^\infty(\Omega) \to C^\infty(\Omega)$ donné par

$$P\varphi(x) = \frac{1}{(2\pi)^n} \int e^{ix.\xi} \sigma(x,\xi) \widehat{\varphi}(\xi) d\xi$$

où φ est une fonction de $C_0^\infty(\Omega), x$ et ξ sont des réels de $\mathbb{R}^n$, $x.\xi$ est leur produit scalaire et $\widehat{\varphi}$ la transformée de Fourier de la fonction φ. La quantité $\sigma(x,\xi)$ est appelée le *symbole* de l'opérateur P. L'opérateur P est encore noté P_σ pour signifier qu'il s'agit de l'opérateur P associé au symbole σ. Si deux opérateurs P_σ et P_τ coïncident, alors les symboles sont égaux ($\sigma = \tau$).

Soit σ un élément de S^m, σ_j une suite de S^{m_j} où (m_j) est une suite de réels décroissante $m = m_0 > m_1 > ... > m$j tendant vers $-\infty$ lorsque j tend vers l'infini et tel que

$$\sigma - \sum_{j=0}^{n-1} \sigma_j \in S^{m_n} \qquad \forall n \in \mathbb{N}$$

Dans ces conditions, on appelle *extension asymptotique* du symbole σ, la somme infinie des σ_j et on note

$$\sigma \sim \sum_{j=0}^{\infty} \sigma_j$$

Si (f, g) désigne le produit scalaire de deux éléments de S

$$(f, g) = \int_{\mathbb{R}^n} f(x)\overline{g(x)}dx$$

on définit pour un opérateur pseudo-différentiel P_σ associé à un symbole σ l'adjoint P_σ^* de cet opérateur par l'égalité

$$(P_\sigma f, g) = (f, P_\sigma^* g)$$

Pour un symbole σ de S^m, l'opérateur adjoint est un opérateur pseudo-différentiel de symbole τ de S^m et d'extension

$$\tau(x, \xi) \sim \sum_{\mu} \frac{(-i)^{|\mu|}}{\mu} (\partial_x^\mu \ \partial_\xi^\mu \ \overline{\sigma})(x, \xi)$$

Si P est l'opérateur

$$P(x, D) = \sum_{|\alpha| \leq m} a_\alpha(x)\partial^\alpha$$

son adjoint est l'opérateur

$$P^* u = \sum_{|\alpha| \leq m} (-1)^{|\alpha|} \partial^\alpha (a_\alpha u)$$

Un symbole σ de S^m est *elliptique* s'il existe deux constantes C et R, telles que

$$\sigma(x, \xi) \geq C(1 + |\xi|)^m \qquad \forall\, |\xi| > R$$

Un opérateur pseudo-différentiel est elliptique si son symbole est elliptique.

8.5 Espaces de Sobolev

Soit $1 \leq p \leq \infty$ un réel, Ω un ouvert de $\mathbb{R}^n$, on considère l'*espace de Sobolev* défini par

$$W^{m,p}(\Omega) = \{u \in L^p(\Omega), \quad \forall \ |\alpha| \leq m, \ \partial^\alpha u \in L^p(\Omega)\}$$

Cet espace muni de la norme

$$\|u\|_{m,p} = \sum_{|\alpha| \leq m} \|\partial^\alpha u\|_{L^p}$$

et pour $p = \infty$

$$\|u\|_{m,\infty} = \max_{|\alpha| \leq m} \{\sup_{x \in \Omega} ess \ |\partial^\alpha u(x)|\}$$

Les espaces de Sobolev sont des espaces de Banach. Si $0 \leq m \leq n$, l'injection $W^{n,p}(\Omega) \subset W^{m,p}(\Omega)$ est continue. On note $W_0^{m,p}(\Omega)$ l'adhérence des fonctions $C_0^\infty(\Omega)$ dans $W^{m,p}(\Omega)$ pour la topologie définie par la norme usuelle de $W^{m,p}(\Omega)$. Pour $p = 2$, on note $H^m(\Omega)$ l'espace $W^{m,2}(\Omega)$ et on appelle *espaces d'énergie* les espaces $H^m(\Omega)$. Étant munis d'un produit scalaire, les espaces d'énergie sont des espaces de Hilbert. Pour $\Omega = \mathbb{R}^n$ et $p = 2$ on démontre la définition équivalente suivante

$$H^m(\mathbb{R}^n) = \{u \in L^2(\mathbb{R}^n), \ (1 + |\xi|^2)^{m/2} \hat{u} \in L^2(\mathbb{R}^n)\}$$

où $\hat{u}$ est la transformée de Fourier de la fonction u. Muni du produit scalaire

$$\langle u, v \rangle_m = \int_{\mathbb{R}^n} (1 + |x|^2)^m \hat{u}(x) \overline{\hat{v}(x)} dx$$

l'espace $H^m(\mathbb{R}^n)$ est un espace de Hilbert. La norme s'écrit

$$\|u\|_m = \left(\int_{\mathbb{R}^n} (1 + |x|^2)^m \, |\hat{u}(x)|^2 \, dx \right)^{1/2}$$

Lorsque m est un entier naturel, on démontre que les normes sont équivalentes

$$\|u\|_s^2 = \sum_{|\alpha| \leq m} \|\partial^\alpha u\|_{L^2}^2$$

On démontre les inclusions

$$\mathcal{D}(\Omega) \subset \mathcal{S}(\Omega) \subset C^\infty(\Omega)$$

et pour m entier naturel

$$\mathcal{D}(\Omega) \subset \mathcal{S}(\Omega) \subset H^m(\Omega) \subset ... \subset H^0(\Omega) = L^2(\Omega) \subset \mathcal{D}'(\Omega)$$

Pour un entier m non nul, on note $H_0^m(\Omega)$ l'adhérence de $\mathcal{D}(\Omega)$ dans $H^m(\Omega)$. On a

$$H_0^m(\Omega) = W_0^{m,2}(\Omega)$$

On démontre que l'ensemble $\mathcal{D}(\Omega)$ est dense dans $H_0^m(\Omega)$ muni de la norme de $H^m(\Omega)$. Le dual de $H_0^m(\Omega)$ est noté $H_0^{-m}(\Omega)$. Si m est un entier naturel et Ω un ouvert borné de $\mathbb{R}^n$, on démontre le *théorème de Rellich* qui affirme que $H_0^{m+1}(\Omega) \subset \ H_0^m(\Omega)$ est une injection compacte.

8.6 Variété des caractéristiques

La variété des caractéristiques S est la variété de $\mathbb{R}^n$ définie par l'équation

$$\phi(x) = 0$$

avec

$$\sigma(x, grad\phi(x)) = 0$$

et

$$grad\phi(x) \neq 0 \quad \text{pour} \quad x \in S$$

Dans le cas particulier où l'équation différentielle s'écrit sous la forme

$$\sum_{i=1}^{n} A_i \frac{\partial u}{\partial x_i} = f$$

où f est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^m$ et A_i une matrice carrée $m \times m$, l'équation des caractéristiques prend la forme

$$\det \left| \sum_{i=1}^{n} A_i \frac{\partial \phi}{\partial x_i} \right| = 0$$

Exemple 1. On considère le système de la dynamique des gaz

$$\begin{cases} \dfrac{\partial u}{\partial t} + \frac{1}{\rho} \dfrac{\partial p}{\partial y} = 0 \\ \dfrac{\partial p}{\partial t} + \rho c^2 \dfrac{\partial u}{\partial y} = 0 \end{cases}$$

Le système s'écrit sous la forme d'une équation matricielle

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \frac{\partial}{\partial t} \begin{pmatrix} u \\ p \end{pmatrix} + \begin{pmatrix} 0 & 1/\rho \\ \rho c^2 & 0 \end{pmatrix} \frac{\partial}{\partial y} \begin{pmatrix} u \\ p \end{pmatrix} = 0$$

L'équation des caractéristiques

$$\det \left| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \frac{\partial \phi}{\partial t} + \begin{pmatrix} 0 & 1/\rho \\ \rho c^2 & 0 \end{pmatrix} \frac{\partial \phi}{\partial y} \right| = 0$$

s'écrit

$$\left(\frac{\partial \phi}{\partial t} \right)^2 - c^2 \left(\frac{\partial \phi}{\partial y} \right)^2 = 0$$

soit en posant $dy/dt = (\dfrac{\partial \phi}{\partial t})/(\dfrac{\partial \phi}{\partial y})$,

$$(dy - cdt)(dy + cdt) = 0$$

Les caractéristiques sont donc les droites d'équation $y \pm ct = Cte$.

Exemple 2. L'équation du second ordre

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} = f$$

s'écrit sous la forme, en posant $X = \partial u/\partial x$ et $Y = \partial u/\partial y$

$$\begin{pmatrix} 0 & a \\ 1 & 0 \end{pmatrix} \frac{\partial}{\partial x} \begin{pmatrix} X \\ Y \end{pmatrix} + \begin{pmatrix} c & b \\ 0 & -1 \end{pmatrix} \frac{\partial}{\partial y} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}$$

L'équation des caractéristiques est alors

$$a\left(\frac{\partial \phi}{\partial x}\right)^2 + b\left(\frac{\partial \phi}{\partial x}\right)\left(\frac{\partial \phi}{\partial y}\right) + c\left(\frac{\partial \phi}{\partial y}\right)^2 = 0$$

soit encore

$$a(\frac{dy}{dx})^2 + b(\frac{dy}{dx}) + c = 0$$

8.7 Classification des équations

L'équation des caractéristiques permet la classification des équations

$$\det \left| \sum_{i=1}^{n} A_i \frac{\partial \phi}{\partial x_i} \right| = 0$$

Si l'équation des caractéristiques n'a aucune racine réelle, l'équation est dite *elliptique*. Si l'équation des caractéristiques a n solutions réelles, l'équation est dite *hyperbolique*. Si l'équation des caractéristiques a des solutions réelles et confondues, l'équation est dite *parabolique*. Dans le cas d'une équation du second ordre

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} = f$$

on a une analogie avec la classification des coniques selon la forme quadratique $q(x, y) = ax^2 + bxy + cy^2$. Si $b^2 - 4ac < 0$, l'équation est elliptique. C'est, par exemple, l'*équation de Poisson* (ou *de Laplace*, si $f = 0$)

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

Si $b^2 - 4ac = 0$, l'équation est parabolique. C'est, par exemple, l'*équation des ondes*

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

Si $b^2 - 4ac > 0$, l'équation est hyperbolique. C'est, par exemple, l'*équation de la chaleur*

$$\frac{\partial u}{\partial t} = c\frac{\partial^2 u}{\partial x^2}$$

Notons que le genre d'une équation peut varier selon les valeurs des variables. L'équation

$$\frac{\partial^2 u}{\partial x^2} + (x^2 - y^2)\frac{\partial^2 u}{\partial y^2} = 0$$

est hyperbolique si $|x| < |y|$, elliptique si $|x| > |y|$ et parabolique si $|x| = |y|$.

8.8 Problèmes équivalents

Soit Ω un domaine de $\mathbb{R}^n$, et $\partial\Omega$ son bord. On considère l'opérateur différentiel $\mathcal{L}$ et le problème de Dirichlet : trouver les solutions de l'équation

$$\begin{cases} \mathcal{L}u = -f & \text{sur } \Omega \\ u = u_0 & \text{sur } \partial\Omega \end{cases}$$

Sous certaines conditions, nous allons voir que ce problème est équivalent à deux autres problèmes : le problème variationnel, lui-même équivalent au problème de minimisation énergétique. Soit V un espace de Hilbert, appelé *espace des fonctions tests*, et A l'opérateur différentiel dérivé de $\mathcal{L}$. On note

$$a(u, v) = \int_\Omega AuAvdx$$

et

$$L(v) = \int_\Omega fvdx$$

Le *problème variationnel* équivaut à trouver une fonction u de V telle que

$$a(u, v) = L(v) \quad \forall\, v \in V$$

V est un espace de Hilbert, $a(u, v)$ une forme sesquilinéaire (linéaire en u et antilinéaire en v : $a(u, v_1 + \lambda v_2) = a(u, v_1) + \overline{\lambda}a(u, v_2)$) continue sur $V \times V$ et *coercive*, c'est-à-dire telle que

$$\exists\alpha > 0, \quad \forall\, v \in V, \quad \mathrm{Re}\, a(u, v) \geq \alpha \|v\|_V^2$$

et L une forme antilinéaire continue sur V. On démontre que si u est une solution du problème de Dirichlet, alors u est solution du problème variationnel. En général, la réciproque est fausse. Mais on peut toutefois trouver un espace $W \subset V$, tel que si u est élément de W, et solution du problème variationnel, alors u est solution du problème de Dirichlet. Les résultats reposent sur le *théorème de Lax-Milgram* qui affirme que si $a(u, v)$

est une forme sesquilinéaire continue, *coercive* sur V, c'est-à-dire vérifiant la majoration suivante

$$\exists M > 0, \quad \forall\, u, v \in V, \quad |a(u,v)| \leq M \,\|u\| \,.\, \|v\|$$

et la minoration de sa partie réelle

$$\exists \alpha > 0, \quad \forall\, v \in V, \quad \mathrm{Re}[a(u,v)] \geq \alpha \,\|u\|^2$$

et si $L(v)$ est une forme antilinéaire continue sur V, alors le problème variationnel qui consiste à trouver $u \in V$, tel que

$$a(u,v) = L(v)$$

admet une solution unique.

D'un point de vue pratique, pour déterminer l'écriture variationnelle d'une équation, on multiplie cette équation, en tenant compte des conditions limites initiales, par une fonction de test et on intègre. On procède ensuite à une intégration par parties (ou à une application de la formule de Stokes) de façon à diminuer le degré de différentiation de la fonction u.

Exemple 1. Écrire la formulation variationnelle de l'équation

$$\frac{d^4u}{dx^4} = f(x)$$

pour $x \in [a,b]$. La formulation forte conduit à écrire

$$\int_a^b v(x)\frac{d^4u}{dx^4}dx = \int_a^b v(x)f(x)dx$$

En intégrant deux fois par parties, on obtient

$$\int_a^b v(x)f(x)dx = \left[v(x)\frac{d^3u}{dx^3}\right]_a^b - \left[\frac{dv}{dx}\frac{d^2u}{dx^2}\right]_a^b + \int_a^b \frac{d^2u}{dx^2}\frac{d^2v}{dx^2}dx$$

Exemple 2. Écrire une formulation variationnelle sur un domaine Ω de $\mathbb{R}^2$ de l'équation

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

En utilisant la formule de Stokes (Green-Riemann), on a

$$\int_\Omega \frac{\partial}{\partial x}(v\frac{\partial u}{\partial x})dxdy = \int_\Omega (v\frac{\partial^2 u}{\partial x^2} + \frac{\partial u}{\partial x}\frac{\partial v}{\partial x})dxdy = \int_{\partial\Omega} v\frac{\partial u}{\partial x}\cos(\theta)ds$$

et

$$\int_\Omega \frac{\partial}{\partial y}(v\frac{\partial u}{\partial y})dxdy = \int_\Omega (v\frac{\partial^2 u}{\partial y^2} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y})dxdy = \int_{\partial\Omega} v\frac{\partial u}{\partial y}\sin(\theta)ds$$

où s est l'abscisse curviligne sur le bord du domaine et θ l'angle entre la normale extérieure $\overrightarrow{n}$ au domaine et le repère normé. La formule de Stokes conduit donc à l'expression

$$\int_\Omega v\Delta u \; dxdy = -\int_\Omega \overrightarrow{\nabla} u.\overrightarrow{\nabla} v \; dxdy + \int_{\partial\Omega} v\overrightarrow{n}.\overrightarrow{\nabla} u \; ds$$

en posant

$$\frac{\partial u}{\partial n} = \overrightarrow{n}.\overrightarrow{\nabla} u = \frac{\partial u}{\partial x}\cos(\theta) + \frac{\partial u}{\partial y}\sin(\theta)$$

La formulation variationnelle s'écrit

$$\int_\Omega u.vdxdy + \int_\Omega \overrightarrow{\nabla} u.\overrightarrow{\nabla} v \; dxdy + \int_{\partial\Omega} v\frac{\partial u}{\partial n}ds = \int_\Omega vfdxdy$$

Exemple 3. Écrire une formulation variationnelle sur l'intervalle [0,1], du système équationnel

$$\begin{cases} -\dfrac{d}{dx}(p(x)\dfrac{du}{dx}) = f(x) & \text{si } x \in [0,1] \\ \\ u(x) = a & \text{si } x = 0 \\ u'(x) = b & \text{si } x = 1 \end{cases}$$

L'introduction d'une fonction test

$$\int_0^1 \frac{d}{dx}(p(x)\frac{du}{dx})v(x)dx + \int_0^1 f(x)v(x)dx + v(0)(u(0)-a) + v(1)(u(1)-b) = 0$$

donne par intégration par parties

$$\begin{aligned} & \int_0^1 f(x)u'(x)v'(x)dx - \int_0^1 f(x)v(x)dx \\ = \; & v(0)(u(0) - a - p(0)u'(0)) + v(1)(u(1) - b - p(1)u'(1)) \end{aligned}$$

Exemple 4. Écrire une formulation du problème suivant : Soit Ω un domaine de $\mathbb{R}^n$ et $\partial\Omega$ son bord. On considère les opérateurs différentiels $\mathcal{L}$ et $\mathcal{B}$, et le problème suivant : trouver u vérifiant

$$\begin{cases} -\mathcal{L}u = f & \text{sur } \Omega \\ -\mathcal{B}u = g & \text{sur } \partial\Omega \end{cases}$$

On note $\mathcal{A}$ l'opérateur tel que $\partial\mathcal{A} = \mathcal{L}$. La formulation variationnelle forte s'écrit

$$\int_\Omega (\mathcal{L}u + f)vdx + \int_{\partial\Omega} (\mathcal{B}u + g)vdx = 0$$

On obtient une formulation faible par application de la formule de Stokes.

Soit $a(u,v)$ une forme hermitienne ($a(u,v) = \overline{a(v,u)}$) et L une forme antilinéaire continue, le *problème de minimisation* énergétique consiste à trouver $u \in V$ tel que

$$J(u) = \inf_{v \in V} J(v)$$

avec

$$J(v) = \frac{1}{2} a(u,v) - \operatorname{Re} L(v)$$

Le problème de minimisation énergétique équivaut à déterminer les fonctions u telles que

$$J(u) \leq J(v), \quad \forall v \in V$$

avec égalité si et seulement si $u = v$.

On démontre que *le problème variationnel équivaut au problème de minimisation*.

En effet, si u est solution du problème variationnel, $a(u,u) = L(u)$. Par conséquent,

$$J(u) = \frac{1}{2} a(u,u) - L(u) = -\frac{1}{2} a(u,u)$$

D'où la différence

$$\begin{aligned} J(v) - J(u) &= \frac{1}{2} a(v,v) - L(v) + \frac{1}{2} a(u,u) \\ &= \frac{1}{2} a(v,v) - a(u,v) + \frac{1}{2} a(u,u) \\ &= \frac{1}{2} [a(v,v) - a(u,v) - a(v,u) + a(u,u)] \\ &= \frac{1}{2} a(v-u, v-u) \geq 0 \end{aligned}$$

avec égalité si et seulement si $u = v$. Par conséquent, u est solution du problème énergétique. Inversement, supposons que u soit une solution du problème de minimisation. Soit w un élément de V et θ un nombre réel. Considérons la fonction

$$F(\theta) = 2J(u + \theta w)$$

Comme

$$\begin{aligned} J(u+\theta w) - J(u) &= a(u+\theta w, u+\theta w) - 2 \operatorname{Re} L(u+\theta w) \\ &\quad -a(u,u) + 2 \operatorname{Re} L(u) \\ &= \theta(a(w,u) + a(u,w)) + \theta^2 a(w,w) - 2\theta \operatorname{Re} L(w) \end{aligned}$$

la dérivée

$$\begin{aligned} F'(0) &= \lim_{\theta \to 0} \frac{F(\theta) - F(0)}{\theta} \\ &= 2 \operatorname{Re}(a(u,w) - L(w)) \end{aligned}$$

Comme F est stationnaire en $\theta = 0$, on a $F\prime(0) = 0$ pour tout $w \in V$. Par conséquent,

$$\mathrm{Re}(a(u,w) - L(w)) = 0, \quad \forall w \in V$$

Soit $w = \alpha + i\beta$, comme

$$\mathrm{Re}(a(u,\alpha) - L(\alpha)) + J(a(u,\beta) - L(\beta)) = 0$$

on en déduit que

$$\forall w \in V, \quad a(u,w) = L(w)$$

ce qui montre que le problème variationnel équivaut au problème de minimisation.

Quand on passe à la discrétisation du problème de Dirichlet, on considère un sous-espace fermé V_h de V, dans lequel on cherche la solution u_h. En appliquant le théorème de Lax-Milgram, on voit que la solution du problème discrétisé existe et est unique, et que le problème variationnel discrétisé

$$\text{trouver } u_h \in V_h \quad \text{tel que} \quad \forall v_h \in V_h, \quad a(u_h, v_h) = L(v_h)$$

équivaut au problème de minimisation discrétisé

$$\text{trouver } u_h \in V_h \quad \text{tel que} \quad J(u_h) = \inf_{v \in V_h} J(v)$$

avec

$$J(v) = \frac{1}{2}a(v,v) - L(v)$$

8.9 Schémas de discrétisation

Soit Ω un domaine de $\mathbb{R}^n$, on considère l'opérateur différentiel L et l'équation

$$Lu(x,t) = f \quad \text{sur} \quad \Omega$$

On découpe le domaine Ω en cellules élémentaires et on discrétise l'opérateur sur chaque cellule. L'expression ainsi obtenue L_h est appelée *schéma de discrétisation*. L'erreur de consistance (de troncature ou de discrétisation) est la différence entre la valeur discrétisée et la valeur exacte

$$\begin{aligned} e_h &= L_h u - Lu \\ &= u_{ij} - u(x_i, t_j) \end{aligned}$$

Plus précisément, on considère l'équation

$$\begin{cases} Lu(x,t) = 0 & \text{sur } \Omega \\ u(x,0) = u_0(x) & \text{sur } \partial\Omega \end{cases}$$

pour x réel et t positif. On construit une subdivision $x_0 < x_1 < ... < x_n$ avec $\Delta x = h = x_i - x_{i-1}$, et $t_0 < t_1 < ... < t_n$ avec $\Delta t = k = t_i - t_{i-1}$. On note u_{ij} une approximation de la solution exacte $u(x_i, t_j)$ et u_j le vecteur $u_j = (u_{1j}, u_{2j}, ..., u_{nj})$. L'équation $Lu(x,t)$ discrétisée

$$A_l u_{j+l} + A_{l-1} u_{j+l-1} + ... + A_0 u_j = 0$$

est appelée schéma de discrétisation à l niveaux en temps. Si A_l est la matrice identité, le schéma est dit *explicite*. Sinon, il est *implicite*. Le schéma est d'ordre p en temps et q en espace si l'erreur de consistance vérifie

$$e_h = O(\Delta t^p) + O(\Delta x^q)$$

Le schéma est *consistant* si l'erreur de consistance e_h tend vers zéro lorsque tous les pas de discrétisation tendent vers zéro.

Exemple. Considérons l'équation de la chaleur

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

et le schéma de discrétisation

$$\frac{u_{i,j+1} - u_{i,j}}{\Delta t} - \alpha \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2} = 0$$

Ce schéma est un schéma à deux niveaux en temps (j et $j+1$) explicite : si $u_{i-1,j}$, $u_{i,j}$ et $u_{i+1,j}$ sont connus à l'instant j, on peut calculer explicitement $u_{i,j+1}$ à l'instant $(j+1)$. La quantité $u_{i,j+1}$ est donnée par le schéma de discrétisation. Pour calculer l'ordre du schéma, écrivons le développement de Taylor

$$u_{i,j+1} = u_{i,j} + \Delta t \left(\frac{\partial u}{\partial t}\right)_{i,j} + \frac{\Delta t^2}{2} \left(\frac{\partial^2 u}{\partial t^2}\right)_{i,j} + O(\Delta t^3)$$

de même

$$\begin{aligned} u_{i+1,j} &= u_{i,j} + \Delta x \left(\frac{\partial u}{\partial x}\right)_{i,j} + \frac{\Delta x^2}{2} \left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} + \frac{\Delta x^3}{6} \left(\frac{\partial^3 u}{\partial x^3}\right)_{i,j} + \\ &\quad \frac{\Delta x^4}{24} \left(\frac{\partial^4 u}{\partial x^4}\right)_{i,j} + \frac{\Delta x^5}{120} \left(\frac{\partial^5 u}{\partial x^5}\right)_{i,j} + O(\Delta x^6) \end{aligned}$$

de même

$$\begin{aligned} u_{i-1,j} &= u_{i,j} - \Delta x \left(\frac{\partial u}{\partial x}\right)_{i,j} + \frac{\Delta x^2}{2} \left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} - \frac{\Delta x^3}{6} \left(\frac{\partial^3 u}{\partial x^3}\right)_{i,j} + \\ &\quad \frac{\Delta x^4}{24} \left(\frac{\partial^4 u}{\partial x^4}\right)_{i,j} - \frac{\Delta x^5}{120} \left(\frac{\partial^5 u}{\partial x^5}\right)_{i,j} + O(\Delta x^6) \end{aligned}$$

par addition, il vient

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2} - \frac{\Delta x^2}{12}\left(\frac{\partial^4 u}{\partial x^4}\right)_{i,j} + O(\Delta x^4)$$

L'erreur de consistance est donc

$$e_h = \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2} - \alpha\frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4} + O(\Delta t^2) + O(\Delta x^4)$$

Le schéma est d'ordre 2 en temps, d'ordre 4 en espace et est consistant puisque l'erreur de consistance tend vers zéro lorsque Δt et Δx tendent vers zéro. En revanche, si nous écrivons une discrétisation de la dérivée en x à l'instant t_j au lieu de l'écrire à l'instant t_{j+1}, on obtient le schéma suivant

$$\frac{u_{i,j+1} - u_{i,j}}{\Delta t} - \alpha\frac{u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}}{(\Delta x)^2} = 0$$

Si les éléments discrétisés sont connus jusqu'à l'instant j, on ne peut pas calculer $u_{i,j+1}$ par l'expression du schéma de discrétisation, car $u_{i+1,j+1}$ et $u_{i-1,j+1}$ sont inconnus. Le schéma donne implicitement $u_{i,j+1}$ à l'instant $(j+1)$. Il suffit d'écrire toutes les équations et de résoudre le système. Le schéma est dit *implicte*.

8.10 Convergence et stabilité

Un schéma de discrétisation est *convergent* si la solution numérique $u_{i,j}$ tend vers la solution exacte $u(x_i, t_j)$ lorsque les pas de discrétisation tendent vers zéro. Le schéma est *conditionnellement convergent* s'il converge pour une condition donnée. Soit H un espace vectoriel normé, A un opérateur de H. On note $u_{n+1} = Au_n$ un schéma numérique. On dit que ce schéma est *stable* s'il existe une constante K indépendante de n telle que

$$\|A^n\| \leq K$$

Le schéma est dit *universellement stable* ou *inconditionnellement stable* si le schéma est toujours stable, c'est-à-dire si K est bornée quels que soient les pas de discrétisation et *conditionnellement stable* si K est bornée pour certaines valeurs des pas de discrétisation. Autrement dit, un schéma est stable si les erreurs ne s'amplifient pas au fur et à mesure que le calcul progresse. D'autre part, remarquons qu'il y a autant de définitions de stabilité que de normes. En général, on parle de stabilité dans L^2 ou au sens de Neumann. On appelle *fonction ou matrice d'amplification*, la fonction ou la matrice S obtenue par transformée de Fourier de l'expression analytique du schéma numérique

$$u_{n+1} = Au_n$$

soit

$$\hat{u}_{n+1} = S(w)\hat{u}_n$$

où $\hat{u}_n$ est la transformée de Fourier

$$\hat{u}_n(w) = \int_{-\infty}^{+\infty} u_n(x)e^{-iwx}dx$$

On vérifie que la transformée de Fourier de $u_{i+k,j}$ satisfait

$$\hat{u}_{i+k,j}(w) = e^{ikw\Delta x}.\hat{u}_{i,j}(w)$$

En effet,

$$\hat{u}_{i+k,j}(w) = \int_{-\infty}^{+\infty} u(x_i + k\Delta x, t_j)e^{-iwx}\ dx_i$$

Par le changement de variable $y_i = x_i + k\Delta x$, on a

$$\begin{aligned}\hat{u}_{i+k,j}(w) &= e^{ikw\Delta x}.\int_{-\infty}^{+\infty} u(y_i, t_j)e^{-iwy_i}\ dy_i \\ &= e^{ikw\Delta x}.\hat{u}_{i,j}(w)\end{aligned}$$

On démontre qu'un schéma est stable si et seulement si tous les éléments de S^n restent bornés quand n tend vers l'infini

$$\|S^n\| \leq 1$$

Un schéma est dit *stable au sens de Neumann* si le rayon spectral de la matrice d'amplification est borné par 1. On démontre que si la norme de la matrice d'amplification est inférieure à 1 alors le schéma est stable et que réciproquement, lorsque la matrice S est normale ($SS^* = S^*S = S$), on a équivalence entre stabilité et condition de Neumann. Ce qui est toujours le cas lorsque S est une fonction. On démontre aussi le même résultat lorsque la matrice S s'écrit comme une somme $S = A + iB$ où A est une matrice symétrique et B une matrice antisymétrique ($B^t = -B$). On démontre encore les deux résultats suivants :

Si le déterminant de la matrice des vecteurs propres de la matrice d'amplification $S(w)$ garde un signe constant pour tout w, alors le schéma est stable si et seulement si le rayon spectral est inférieur à 1 : $\rho(S) \leq 1$.

Si tous les éléments de la matrice d'amplification $S(w)$ sont bornés pour tout w, et si toutes les valeurs propres de S, sauf peut-être une, sont strictement inférieures à 1, alors le schéma est stable si et seulement si le rayon spectral est inférieur à 1 : $\rho(S) \leq 1$.

Pour les problèmes linéaires bien posés, le *théorème de Lax* affirme l'équivalence des notions de convergence et de stabilité. Il affirme que pour qu'un schéma numérique d'un problème linéaire bien posé converge il faut et il suffit que ce schéma soit stable et consistant.

Exemple 1. Schéma implicite. Considérons l'équation de la chaleur

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

et le schéma implicite

$$\frac{u_{i,j+1} - u_{i,j}}{\Delta t} - \alpha \frac{u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}}{(\Delta x)^2} = 0$$

Par transformée de Fourier,

$$\hat{u}_{i,j+1} - \hat{u}_{i,j} - \alpha \frac{\Delta t}{(\Delta x)^2} [e^{iw\Delta x} \hat{u}_{i,j+1} - 2\hat{u}_{i,j+1} + e^{-iw\Delta x} \hat{u}_{i,j+1}] = 0$$

soit

$$\hat{u}_{i,j+1}(w) = S(w) \hat{u}_{i,j}(w)$$

avec

$$S(w) = \frac{1}{1 + 4\alpha \frac{\Delta t}{(\Delta x)^2} \sin^2(w \frac{\Delta x}{2})}$$

Comme $\sup_w |S(w)| \leq 1$, le schéma est universellement stable.

Exemple 2. Schéma de Richardson. Considérons l'équation de la chaleur et le schéma de discrétisation

$$\frac{u_{i,j+1} - u_{i,j-1}}{2\Delta t} - \alpha \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2} = 0$$

Par transformée de Fourier

$$\hat{u}_{i,j+1} - \hat{u}_{i,j-1} + 8\alpha \frac{\Delta t}{(\Delta x)^2} \sin^2 \left(w \frac{\Delta x}{2} \right) \hat{u}_{i,j} = 0$$

Soit, en posant $v_{i,j+1} = u_{i,j}$ et $a = -8\alpha \frac{\Delta t}{(\Delta x)^2} \sin^2 \left(w \frac{\Delta x}{2} \right)$

$$\begin{pmatrix} \hat{u}_{i,j+1} \\ \hat{v}_{i,j+1} \end{pmatrix} = \begin{pmatrix} a & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \hat{u}_{i,j} \\ \hat{v}_{i,j} \end{pmatrix}$$

Les valeurs propres de la matrice sont $(-a \pm \sqrt{4 + a^2})/2$, d'où le rayon spectral

$$\rho(S) = \frac{-a + \sqrt{4 + a^2}}{2}$$

Comme $\rho(S) > 1$, le schéma est toujours instable.

8.11 Exercices

1. On considère l'équation suivante

$$-\frac{d^2u}{dx^2} + b\frac{du}{dx} = -b$$

sur l'intervalle $]0,1[$ et les conditions limites $u(0) = u(1) = 0$. L'espace d'approximation est l'espace de Hilbert V défini par

$$V = \{v \in H_0^1(0,1) \ : \ v(0) = v(1)\}$$

Mettre le problème sous forme variationnelle et montrer que ce problème admet une solution unique sur V.

2. *Fléchissement d'une poutre*. Étant donné deux fonctions c et f continues sur l'intervalle $[0,1]$, la fonction $c(x)$ étant positive ou nulle sur cet intervalle, on considère le problème du fléchissement d'une poutre soumise à une force f sous la forme

$$\begin{cases} -u"(x) + c(x)u(x) = f(x) & 0 < x < 1 \\ u(0) = u(1) = 0 \end{cases}$$

On suppose la solution u deux fois continûment dérivable sur $[0,1]$. On note V l'espace de Sobolev $H_0^1([0,1])$, c'est-à-dire l'espace des fonctions continues sur $[0,1]$, nulles aux bornes de cet intervalle et continûment dérivables par morceaux. L'espace V est muni de la norme

$$\|v\| = \left(\int_0^1 (|v'(x)|^2 + |v(x)|^2)dx\right)^{1/2}$$

On note

$$a(u,v) = \int_0^1 (u'(x)v'(x) + c(x)u(x)v(x))dx$$

la forme bilinéaire $V \times V$ dans R et L la forme linéaire de V dans R

$$L(v) = \int_0^1 f(x)v(x)dx$$

1) Écrire la formulation variationnelle du problème de Dirichlet.
2) Montrer qu'il existe un nombre $\alpha > 0$ tel que

$$\alpha \|v\|^2 \leq a(v,v) \qquad \forall v \in V$$

En déduire que le problème variationnel admet une solution unique.
3) On note

$$J(v) = \frac{1}{2}a(v,v) - L(v)$$

Montrer que $u \in V$ est solution des équations

$$a(u,v) = L(v) \qquad \forall v \in V$$

si et seulement si

$$J(u) = \inf_{v \in V} J(v)$$

4) Déterminer la fonction u qui minimise sur V l'expression

$$H(v) = \int_0^1 (\frac{1}{2} x v^2(x) - x(x+1)(x-3)v(x)e^x + \frac{1}{2}(v'(x))^2) dx$$

5) Soit n un entier naturel non nul, on pose $h = 1/(n+1)$ et on définit une subdivision de l'intervalle $[0,1]$ aux nœuds $x_i = ih$ avec $0 \leq i \leq n+1$. On pose $c_i = c(x_i)$, $u_i = u(x_i)$ et $f_i = f(x_i)$. Écrire le système d'équations obtenu si on approche $u"(x_i)$ par le schéma

$$u"(x_i) \simeq \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta x)^2}$$

6) On suppose que $c_k \ll \frac{2}{h^2}$ pour tout $k = 1, \ldots, n$ et on admet que la matrice

$$A_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

a comme valeurs propres les nombres

$$\lambda_k = \frac{4}{h^2} \sin^2 \left(\frac{k\pi}{2(n+1)} \right)$$

Calculer le nombre de conditionnement de cette matrice. Que peut-on dire si n est grand ?

7) On suppose que $c = \omega^2$ où ω est une constante > 0 et que $f = 0$. On considère le schéma de Newmark suivant

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta x)^2} + \omega^2(\theta u_{i+1} + (1-2\theta)u_i + \theta u_{i-1}) = 0$$

où θ est un paramètre. Étudier directement la stabilité du schéma. Quelle valeur de θ préconisez-vous ?

Indication **:** On utilisera sans démonstration (question 2 de l'exercice 1 et question 7 du problème) le résultat suivant : Les conditions pour que les racines de l'équation $\lambda^2 - S\lambda + P = 0$ soient de module inférieur ou égal à 1 sont $P \leq 1$ si $\Delta \leq 0$ et $1 + S + P \geq 0$ si $\Delta > 0$.

9
Équations elliptiques

Lorsque le problème est bien posé, la solution d'une équation elliptique dépend entièrement des conditions limites. Ce comportement est typique des équations elliptiques et paraboliques. On dit que l'opérateur elliptique est *régularisant* : une donnée continue bornée conduit à une solution de classe C^∞. Nous prendrons l'*équation de Poisson* (ou *l'équation de Laplace* si $f = 0$) comme prototype des équations elliptiques linéaires homogènes

$$-\Delta u = f$$

avec

$$\Delta = \frac{\partial^2}{\partial x_1^2} + ... + \frac{\partial^2}{\partial x_n^2}$$

Pour l'équation de Laplace, le problème de Dirichlet ou le problème de Dirichlet-Neumann est bien posé. En revanche, le problème de Neumann est un problème mal posé ($n \geq 1$), comme on le vérifie facilement pour $n = 1$. L'équation $u''(x) = 0$ sur l'intervalle $[a, b]$ dotée des conditions limites $u'(a) = u_0$ et $u'(b) = v_0$ conduit à une solution de la forme $u''(x) = \alpha x + \beta$ qui, ou bien n'a pas de solution si $u_0 \neq v_0$, ou bien admet une infinité de solutions si $u_0 = v_0$. On démontre que les problèmes sont bien posés à l'aide du *principe du maximum*. Ce résultat permet aussi de démontrer un grand nombre de résultats sur l'existence, l'unicité et la régularité des solutions des problèmes elliptiques.

9.1 Fonctions harmoniques. Principe du maximum

Les fonctions harmoniques sur un ouvert Ω de $\mathbb{R}^n$ sont des fonctions dont le laplacien est nul $\Delta u = 0$. Elles vérifient le *principe du maximum* qui affirme que si une fonction réelle u est harmonique dans un ouvert de Ω de $\mathbb{R}^n$ et continue sur le bord $\partial\Omega$, alors u n'a ni maximum local strict ni minimum local strict dans Ω. Autrement dit, les valeurs de u dans Ω sont comprises entre l'inf et le sup des valeurs de u prises sur le bord $\partial\Omega$.

$$\inf_{y\in\partial\Omega} u(y) < u(x) < \sup_{y\in\partial\Omega} u(y)$$

Les fonctions harmoniques de $\mathbb{R}^n - \{0\}$ sont des fonctions centrales qui ne dépendent que de la norme de x. On démontre que ces fonctions sont de classe $C^\infty(\Omega)$ et analytiques sur Ω.

9.2 L'opérateur de Laplace

Au sens des distributions, l'équation de Laplace sur un domaine de $\mathbb{R}^n$

$$\Delta u = \delta$$

admet une solution fondamentale qui est donnée par les formules

$$E(x,y) = \begin{cases} \dfrac{1}{(n-2)s_n}\,|x-y|^{2-n} & \text{pour } n \geq 3 \\ \dfrac{1}{2\pi}\ln\dfrac{1}{|x-y|} & \text{pour } n = 2 \end{cases}$$

où $|x| = \left(\sum |x_i^2|\right)^{1/2}$ désigne la norme euclidienne de $\mathbb{R}^n$ et s_n est la surface de la sphère unité

$$s_n = \frac{2\pi^{n/2}}{\Gamma(n/2)}$$

Ce résultat se généralise à un opérateur elliptique de la forme

$$Lu = \sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{ij}(x)\frac{\partial u}{\partial x_i}\right)$$

La solution fondamentale pour $n \geq 3$ est

$$E(x,y) = \frac{1}{(n-2)s_n} d(x,y)^{2-n}$$

où d est une distance. Si les a_{ij} sont des constantes, alors

$$d(x,y) = \det(a_{ij})^{(n-2)/2}\left(\sum_{i,j=1}^{n}\frac{1}{a_{ij}}(x_i-y_i)(x_j-y_j)\right)^{1/2}$$

Au plan formel (i.e. sans se préoccuper des conditions de convergence), les solutions de l'équation de Poisson $\Delta u = f$ sont de la forme $u = E * f$, soit

$$u(x) = \int_{\partial\Omega} E(x,y) f(y) d\sigma(y)$$

où $d\sigma$ est une mesure sur la sphère unité de $\mathbb{R}^n$. Ce résultat se généralise aussi au cas des opérateurs *métaharmoniques*. Par exemple, l'opérateur *biharmonique* $\Delta^2 u = \Delta\Delta u$, qui intervient dans la vibration d'une plaque, admet une solution fondamentale sur $\mathbb{R}^2$ de la forme

$$E(x,y) = \frac{|y-x|^2}{8\pi} \ln\ |y-x|$$

En coordonnées polaires, le laplacien s'écrit sous la forme

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{n}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\Delta_S$$

où Δ_S est l'opérateur de Laplace sur la sphère unité. L'opérateur $-\Delta_S$ admet des valeurs propres

$$\lambda_k = k(k+n-1) \qquad k = 0, 1, 2...$$

auxquelles sont associés les espaces propres E_k de dimension

$$\dim E_k = \frac{2k+n-1}{n-1}\binom{k+n-2}{k}$$

9.3 Équations elliptiques linéaires

Pour le problème de Dirichlet sur un ouvert Ω de $\mathbb{R}^n$

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases}$$

on démontre que si Ω est un ouvert borné et si f est une fonction de $L^p(\Omega)$ avec $0 < p < 1$, alors il existe une unique solution $u \in W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$ qui vérifie l'estimation

$$\|u\|_{W^{2,p}(\Omega)} \leq C \|f\|_{L^p(\Omega)}$$

De plus, on a la propriété de régularité suivante : si Ω est un ouvert borné de classe C^{m+1}, $(m \geq 0)$ et si f est une fonction de $H^{m-1}(\Omega)$, alors la solution u est une fonction de $H^{m+1}(\Omega)$. Lorsque la condition sur le bord est de la forme $u = u_0$, il faut ajouter des conditions sur la fonction u_0. Le même résultat est valable si on suppose que $u_0 \in H^{m+1/2}(\Omega)$.

Pour le problème de Neumann sur un ouvert Ω de $\mathbb{R}^n$

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ \dfrac{\partial u}{\partial n} = u_1 & \text{sur } \partial\Omega \end{cases}$$

on démontre plusieurs théorèmes de régularité. En particulier, si Ω est un ouvert borné de classe C^{m+2}, $(m \geq 0)$ et si f est une fonction de $H^m(\Omega)$ et u_1 une fonction de $H^{m+1/2}(\Omega)$, alors la solution u est une fonction de $H^{m+2}(\Omega)$.

Exemple. On peut vérifier sur un cas simple que les solutions du problème de Dirichlet sont de classe C^∞ dès que les conditions initiales sont continues. Sur le disque de rayon R

$$\Omega = \{(r, \theta) : \ 0 \leq r \leq R\}$$

le problème de Dirichlet

$$\begin{cases} \Delta u = 0 & \text{dans } \Omega \\ u = f(\theta) & \text{sur } r = R \end{cases}$$

admet une solution unique de la forme

$$\begin{aligned} u(r, \theta) &= \frac{1}{2\pi} \int_0^{2\pi} f(\theta - \alpha) P(r, \alpha) d\alpha \\ &= f * G(r) \end{aligned}$$

où $P(r, \theta)$ est le noyau de Poisson et G la fonction de Green $G(r) = P(r, \theta)/2\pi$

$$P(r, \theta) = \frac{R^2 - r^2}{R^2 + r^2 - 2Rr\cos(\theta)}$$

Plus généralement, l'opérateur elliptique linéaire du second ordre

$$Lu = -div(A(x)\nabla u(x)) + div(bu(x)) + c(x)u(x) = 0$$

qui s'écrit encore

$$Lu = -\sum_{i,j=1}^{n} a_{ij}(x)\partial_{ij}u(x) + \sum_{i=1}^{n} b_i(x)\partial_i u(x) + c(x)u(x) = 0$$

où a_{ij} est une matrice $n \times n$, $b = (b_i)$ un vecteur de $\mathbb{R}^n$ et c une fonction, vérifie l'équation suivante sur un ouvert Ω de $\mathbb{R}^n$

$$Lu = f$$

On peut supposer que la matrice a_{ij} est une matrice symétrique, car $\partial_{ij}u = \partial_{ji}u$. Nous supposons en outre que cette matrice vérifie la *condition de coercivité (ou d'ellipticité stricte)*, c'est-à-dire qu'il existe une constante $\alpha > 0$ telle que pour tout point $\xi \in \Omega$

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_j\xi_i \geq \alpha \left|\xi\right|^2$$

Si la fonction $f \in H^{k-1}(\Omega)$ pour $k = 0, 1, 2,...$ une solution u de $H_0^1(\Omega)$ appartient à $H^k(\Omega)$ et on a pour tout $u \in H^{k+1}(\Omega) \cap H_0^1(\Omega)$

$$\left\|u\right\|_{H^{k+1}}^2 \leq c_1 \left\|Lu\right\|_{H^{k-1}}^2 + c_2 \left\|u\right\|_{H^k}^2$$

On démontre que si l'ouvert Ω est borné et de classe $C^{1,1}$ (c'est-à-dire que le bord est une fonction de classe $C^1(\Omega)$, dont la dérivée première est höldérienne d'ordre 1 au voisinage de tout point de Ω), si les coefficients a_{ij} sont dans $C(\overline{\Omega})$, si b et c sont $L^\infty(\Omega)$, si $c \geq 0$ et si la condition de coercivité est satisfaite alors pour $f \in L^p(\Omega)$, il existe une unique fonction $u \in W^{2,p}(\Omega) \cap W^{1,p}(\Omega)$ avec $1 < p < \infty$ solution du problème de Dirichlet

$$\begin{cases} Lu = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases}$$

De plus, cette solution vérifie l'estimation

$$\left\|u\right\|_{w^{2,p}(\Omega)} \leq C \left\|f\right\|_{L^p(\Omega)}$$

En particulier, si $p > n/2$ et si $u_0 \in C(\overline{\Omega})$, alors il existe une solution unique $u \in C(\overline{\Omega}) \cap W_{loc}^{2,p}(\Omega)$ au problème de Dirichlet

$$\begin{cases} Lu = f & \text{dans } \Omega \\ u = u_0 & \text{sur } \partial\Omega \end{cases}$$

On démontre aussi plusieurs résultats concernant la borne supérieure essentielle de u. Le *principe du maximum faible* affirme que pour un ouvert Ω borné connexe, une matrice a_{ij} vérifiant la condition de coercivité, des coefficients a_{ij}, b_i et une fonction c dans $C(\overline{\Omega})$, si $u \in C(\overline{\Omega}) \cap C^2(\Omega)$, et si l'opérateur L vérifie $Lu \leq 0$ sur Ω, alors la fonction u atteint sa borne supérieure sur le bord de Ω

$$\sup_{x\in\Omega} u(x) = \sup_{x\in\partial\Omega} u(x)$$

Le *principe du maximum fort* affirme que sous les mêmes conditions ou bien u est constant ou bien

$$u(x) < \sup_{y\in\partial\Omega} u(y) \qquad \forall x \in \Omega$$

En particulier, si u atteint un maximum positif ou nul sur l'intérieur de Ω, alors la fonction u est constante sur Ω.

9.4 Équations elliptiques non linéaires

Les équations elliptiques non linéaires se classent en trois catégories. Les *équations semi-linéaires* de la forme

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

les *équations quasi linéaires* de la forme

$$\sum_{i,j=1}^{n} a_{ij}(x, u, \nabla u)\partial_i \partial_j \; u = f(x, u, \nabla u)$$

et les *équations complètement non linéaires* de la forme

$$f(x, D^2 u) = 0$$

Lorsque Ω est un ouvert non vide, simplement connexe à frontière indéfiniment différentiable et f une fonction $C^\infty(\Omega \times \mathbb{R})$ vérifiant $\partial f/\partial u \geq 0$ et u_0 une fonction de classe $C^\infty(\partial\Omega)$, alors le problème de Dirichlet

$$\begin{cases} \Delta u = f(x, u) & \text{dans } \Omega \\ u = u_0 & \text{sur } \partial\Omega \end{cases}$$

admet une solution unique $u \in C^\infty(\Omega)$. En général, pour les équations non linéaires, on démontre des résultats de résolubilité locale. La régularité des solutions, quand elles existent, font intervenir de nouveaux espaces, comme par exemple les espaces de Zygmund.
Soit Ω un ouvert de $\mathbb{R}^n$ et $1 < p < (n+2)/(n-2)$, on considère le problème non linéaire suivant

$$\begin{cases} -\Delta u + \lambda \, |u|^{p-1} \, u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases}$$

dans lequel λ est un réel positif et f une fonction de L^q où q est le conjugué de p $(1/p + 1/q = 1)$. L'existence des solutions consiste à minimiser la *fonctionnelle d'énergie* $E(u)$ définie sur $u \in H_0^1(\Omega)$

$$E(u) = \frac{1}{2} \int_\Omega \; |\nabla u|^2 + \frac{\lambda}{p+1} \int_\Omega \; |u|^{p+1} - \int_\Omega \; fu$$

Le problème admet une infinité de solutions, car on démontre que la fonctionnelle d'énergie, qui n'est pas bornée inférieurement, possède une infinité de points critiques.

9.5 Méthode de Richardson-Liebmann

Dans cette méthode, l'équation de Poisson

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

est discrétisée selon l'expression

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{(\Delta y)^2} = f_{i,j}$$

ainsi que les conditions limites. Lorsque la discrétisation est la même en x et en y, $h = \Delta x = \Delta y$, la discrétisation s'écrit plus simplement

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = h^2 f_{i,j}$$

On obtient alors un système d'équations où les inconnues sont les valeurs $u_{i,j}$ de la fonction u en chacun des nœuds du maillage de discrétisation. On résout ce système par une méthode matricielle. La méthode est appelée *méthode de Liebmann* lorsque la résolution se fait par la méthode de Gauss-Seidel, et *méthode de Richardson* lorsque la résolution du système se fait par la méthode de Jacobi.

9.6 Méthodes de relaxation

Dans les *méthodes de relaxation*, l'équation de Poisson

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

écrite sous la forme discrétisée habituelle conduit à un système linéaire qui est résolu par une méthode de relaxation. À la k-ième itération, la méthode revient à calculer

$$u_{i,j}^{(k+1)} = (1 - w)u_{i,j}^{(k)} + w\xi_{i,j}^{(k)}$$

avec

$$\xi_{i,j}^{(k)} = \frac{1}{4}(f_{i,j}h^2 - u_{i+1,j}^{(k)} - u_{i-1,j}^{(k)} - u_{i,j+1}^{(k)} - u_{i,j-1}^{(k)})$$

9.7 Méthode par transformée de Fourier rapide

L'équation de Poisson

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

écrite sous la forme discrétisée

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = h^2 f_{i,j}$$

est modifiée par transformation de Fourier en

$$(e^{2i\pi m/I} + e^{-2i\pi m/I} + e^{2i\pi n/J} + e^{-2i\pi n/J} - 4)\hat{u}_{m,n} = h^2 \widehat{f}_{m,n}$$

soit

$$2(\cos(2\pi m/I) + \cos(2\pi n/J) - 2)\hat{u}_{m,n} = h^2 \widehat{f}_{m,n}$$

lorsque la discrétisation porte sur $x_0, ..., x_I$ et $y_0, ..., y_J$. La méthode consiste à calculer $\widehat{f}_{m,n}$ par

$$\widehat{f}_{m,n} = \sum_{l=0}^{I-1} \sum_{k=0}^{J-1} e^{2i\pi ml/I} e^{2i\pi nk/J} f_{l,k}$$

puis $\hat{u}_{m,n}$ par l'équation discrétisée

$$\hat{u}_{m,n} = \frac{h^2 \widehat{f}_{m,n}}{2(\cos(2\pi m/I) + \cos(2\pi n/J) - 2)}$$

enfin, on trouve $u_{i,j}$ par la formule d'inversion

$$u_{l,k} = \frac{1}{IJ} \sum_{m=0}^{I-1} \sum_{n=0}^{J-1} e^{-2i\pi lm/I} e^{-2i\pi nk/J} \hat{u}_{m,n}$$

9.8 Exercices

1. On considère l'équation de Laplace

   $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

   sur un domaine Ω délimité par les droites d'équations $x = 0$, $x = 4$, $y = 0$ et $y = 5$. On suppose que, sur la frontière, la fonction vaut

   $$u(x, y) = 5(x^2 - 2) + y^2$$

   Écrire pour différents schémas de discrétisation le système à résoudre. On supposera que le maillage est uniforme et de pas $h = 1$, en abscisse comme en ordonnée.

2. En utilisant le développement de Taylor, calculer l'ordre de la méthode de Liebmann pour l'équation de Laplace bidimensionnelle.

10
Équations paraboliques

Dans les problèmes paraboliques (comme dans les problèmes elliptiques), les solutions dépendent essentiellement des conditions initiales. Une donnée initiale continue bornée conduit à une solution de classe C^∞ : on dit que l'opérateur parabolique est *régularisant*. Le prototype des équations paraboliques linéaires est l'équation de la chaleur ou équation de la diffusion

$$\frac{\partial u}{\partial t} - \nabla.(a\nabla u) + cu = f$$

10.1 Équation de la chaleur

Considérons le problème suivant

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta u = 0 \\ u(x,0) = f(x) \end{cases}$$

Si $f \in S'(\mathbb{R}^n)$, $x \in \mathbb{R}^n$ réel et t dans $[0,T]$, alors le problème admet au sens des distributions une solution unique $u \in C^\infty(\overline{\mathbb{R}}^+, S'(\mathbb{R}^n))$ donnée par

$$u(x,t) = E(x,t) * f(x)$$

avec

$$E(x,t) = \frac{1}{(4\pi t)^{n/2}} e^{-|x|^2/4t}$$

Si f est une fonction de $L^2(\mathbb{R}^n)$, alors u est une fonction seulement continue. Dans le cas de variables réelles ($t,\ x \in \mathbb{R}$), la distribution

$$E(x,t) = \frac{1}{(4\pi ct)^{1/2}} e^{-x^2/4ct} H(t)$$

où $H(t)$ est la distribution de Heaviside (valant 1 pour $t > 0$ et 0 sinon) est solution fondamentale de l'équation $DE = \delta$ où D est l'opérateur

$$D = \frac{\partial}{\partial t} - c\frac{\partial^2}{\partial x^2}$$

Soit $\varphi \in \mathcal{D}(\mathbb{R}^n)$,

$$\begin{aligned} \langle DE, \varphi \rangle &= -\langle E, \partial_t \varphi - c\partial_{xx}\varphi \rangle \\ &= -\int_{[0,\infty[\times\mathbb{R}} (4\pi ct)^{-1/2} e^{-x^2/4ct} (\partial_t \varphi - c\partial_{xx}\varphi)\ dxdt \end{aligned}$$

Évaluons séparément les deux intégrales

$$\begin{aligned} I_1 &= \int_{[0,\infty[\times\mathbb{R}} (4\pi ct)^{-1/2} e^{-x^2/4ct} (\partial_t \varphi)\ dxdt \\ &= \lim_{\varepsilon\to 0} \int_0^\infty \int_\varepsilon^{+\infty} (4\pi ct)^{-1/2} e^{-x^2/4ct} (\partial_t \varphi)\ dxdt \end{aligned}$$

En intégrant par parties, on obtient

$$\begin{aligned} I_1 &= \lim_{\varepsilon\to 0} \int_0^\infty \int_\varepsilon^{+\infty} \partial_t \left((4\pi ct)^{-1/2} e^{-x^2/4ct} \right) \varphi\ dxdt + \\ &\quad + \int_{-\infty}^{+\infty} \left[(4\pi ct)^{-1/2} e^{-x^2/4ct} \varphi(x,t) \right]_\varepsilon^\infty dx \end{aligned}$$

soit

$$\begin{aligned} I_1 &= \lim_{\varepsilon\to 0} \frac{-1}{4\sqrt{\pi c}} \int_0^\infty \int_\varepsilon^{+\infty} \left(\frac{x^2}{2c^2 t^{5/2}} - \frac{1}{ct^{3/2}} \right) e^{-x^2/4ct} \varphi\ dxdt \\ &\quad - \int_{-\infty}^{+\infty} (4\pi c\varepsilon)^{-1/2} e^{-x^2/4c\varepsilon} \varphi(x,\varepsilon)\ dx \end{aligned}$$

De la même manière, on calcule la deuxième intégrale

$$\begin{aligned} I_2 &= \int_{[0,\infty[\times\mathbb{R}} (4\pi ct)^{-1/2} e^{-x^2/4ct} (c\partial_{xx}\varphi)\ dxdt \\ &= \lim_{\varepsilon\to 0} \int_0^\infty \int_\varepsilon^{+\infty} (4\pi ct)^{-1/2} e^{-x^2/4ct} (c\partial_{xx}\varphi)\ dxdt \end{aligned}$$

En intégrant deux fois par parties, et en utilisant le fait que la fonction φ s'annule à l'infini

$$I_2 = \lim_{\varepsilon\to 0} \frac{-1}{4\sqrt{\pi c}} \int_0^\infty \int_\varepsilon^{+\infty} \left(\frac{x^2}{2c^2t^{5/2}} - \frac{1}{ct^{3/2}} \right) e^{-x^2/4ct} \varphi \; dxdt$$

d'où

$$\begin{aligned} \langle DE, \varphi \rangle &= \lim_{\varepsilon\to 0} \int_{-\infty}^{+\infty} (4\pi c\varepsilon)^{-1/2} e^{-x^2/4c\varepsilon} \varphi(x,\varepsilon) \; dx \\ &= \lim_{\varepsilon\to 0} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi}} e^{-y^2} \varphi(2c\sqrt{\varepsilon}y, \varepsilon) \; dy \end{aligned}$$

par changement de variable $y = x/\sqrt{4c\varepsilon}$, on obtient finalement

$$\langle DE, \varphi \rangle = \varphi(0,0)$$

Par conséquent E est bien une solution élémentaire de l'opérateur D. Le problème considéré pour $x \in \mathbb{R}$ et $t > 0$ admet donc une solution unique

$$u(x,t) = \frac{1}{2\sqrt{\pi ct}} \int_{-\infty}^{+\infty} f(y) e^{-\frac{(x-y)^2}{4ct}} \; dy$$

Ce résultat permet de résoudre d'autres problèmes dans lesquels le domaine est restreint au cas des $x > 0$. Par exemple, le problème de Dirichlet

$$\begin{cases} \dfrac{\partial u}{\partial t} - c\Delta u = 0 & x > 0,\ t > 0 \\ u(x,0) = f(x) \\ u(0,t) = 0 \end{cases}$$

admet comme solution

$$u(x,t) = \frac{1}{2\sqrt{\pi ct}} \int_0^{+\infty} f(y) \left(e^{-\frac{(x-y)^2}{4ct}} - e^{-\frac{(x+y)^2}{4ct}} \right) \; dy$$

Le problème de Neumann

$$\begin{cases} \dfrac{\partial u}{\partial t} - c\Delta u = 0 & x > 0,\ t > 0 \\ u(x,0) = f(x) \\ \partial_x u(0,t) = 0 \end{cases}$$

admet pour solution

$$u(x,t) = \frac{1}{2\sqrt{\pi ct}} \int_0^{+\infty} f(y) \left(e^{-\frac{(x-y)^2}{4ct}} + e^{-\frac{(x+y)^2}{4ct}} \right) \; dy$$

10.2 Équation de la diffusion

Considérons le problème suivant

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nabla.(a\nabla u) + cu = 0 & x \in \Omega, \quad t > 0 \\ u(x,0) = u_0(x) & x \in \Omega \\ u(x,t) = 0 & x \in \partial\Omega \end{cases}$$

où (a_{ij}) est une matrice symétrique réelle définie positive de $C^\infty(\overline{\Omega})$ et $c(x)$ une fonction positive ou nulle appartenant aussi à $C^\infty(\overline{\Omega})$. On démontre que ce problème admet une solution si Ω est borné en utilisant une décomposition spectrale de l'opérateur

$$A = -\sum_{i,j=1}^{n} \partial_j a(x) \partial_i + c(x)$$

et si Ω n'est pas borné, on montre que l'opérateur A de domaine

$$D(A) = \{u \in H_0^1(\Omega), \quad A(u) \in L^2(\Omega)\}$$

est un opérateur *maximal accrétif* et le *théorème de Hille-Yosida* permet de conclure qu'il existe une unique solution. Notons qu'un opérateur A non borné de domaine $D(A)$ sur un espace de Hilbert H est dit *maximal accrétif* si $Re(Au, u) \geq 0$, pour tout $u \in D(A)$ et si pour un $\alpha > 0$, l'image de $(A + \alpha)$ est égale à l'espace H tout entier

$$Im(A + \alpha) = H$$

10.3 Équation parabolique non linéaire

Soit Ω ouvert non borné de $\mathbb{R}^n$, on considère l'équation

$$\frac{\partial u}{\partial t} - c\Delta u - \sum_{j=1}^{m} \partial_j h_j(u) = 0$$

avec $u(x,0) = f(x)$, u prend ses valeurs dans $\mathbb{R}^m$, ∂_j désigne la dérivée par rapport à u_j, les fonctions h_j sont dérivables et vérifient pour $1 \leq p \leq \infty$

$$|h_j(u)| \leq C(1 + |u|^2)^p$$

et

$$|\nabla h_j(u)| \leq C(1 + |u|^2)^{p-1}$$

On démontre que le problème admet pour $q \geq p$ et $q > n(p-1)$ si $f \in L^q(\Omega)$ une solution unique $u \in C([0,T], L^q(\Omega))$. Cette solution est indéfiniment dérivable $u \in C^\infty([0,T] \times \Omega)$. Dans le cas où u est une fonction à valeurs réelles ($m = 1$), f est une fonction de $L^\infty(\Omega)$, alors le problème admet une solution unique $u \in L^\infty([0,\infty[\times\Omega) \cap C^\infty([0,\infty[\times\Omega)$. Lorsque Ω est un ouvert borné, et u une fonction scalaire ($m = 1$), le problème de Dirichlet

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta u - F(t,x,u,\nabla u) = 0 & x \in \Omega, \quad t > 0 \\ u(x,0) = f(x) & x \in \Omega \\ u(x,t) = 0 & x \in \partial\Omega, \quad t > 0 \end{cases}$$

admet une solution unique. Si la fonction f est une fonction de classe C^1, bornée sur $\overline{\Omega}$, la solution est indéfiniment dérivable sur $[0,\infty[\times\Omega$.

10.4 Méthode du theta-schéma

L'équation de la chaleur

$$\frac{\partial u}{\partial t} = c\frac{\partial^2 u}{\partial x^2}$$

est discrétisée sous la forme

$$\begin{aligned} \frac{u_{i,j+1} - u_{i,j}}{\Delta t} &= \theta\ c\ \frac{u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1}}{(\Delta x)^2} + \dots \\ &\quad \dots + (1-\theta)\ c\ \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{(\Delta x)^2} \end{aligned}$$

La méthode du θ-schéma est appelée *méthode explicite* si $\theta = 0$, *méthode de Crank-Nicholson* si $\theta = 1/2$ et *méthode implicite* si $\theta = 1$. On démontre que si $0 \leq \theta < 1/2$, le θ-schéma est stable si

$$\frac{c\Delta t}{(\Delta x)^2} \leq \frac{1}{2(1-2\theta)}$$

et que si $1/2 \leq \theta \leq 1$, la méthode est universellement stable. En effet, en prenant la transformée de Fourier du θ-schéma,

$$\begin{aligned} \hat{u}_{i,j+1} - \hat{u}_{i,j} &= a\theta(\hat{u}_{i,j+1}e^{-ik\Delta x} - 2\hat{u}_{i,j} + \hat{u}_{i,j+1}e^{ik\Delta x}) \\ &\quad + a(1-\theta)(\hat{u}_{i,j}e^{-ik\Delta x} - 2u_{i,j} + \hat{u}_{i,j}e^{ik\Delta x}) \end{aligned}$$

avec

$$a = \frac{c\Delta t}{(\Delta x)^2}$$

On obtient finalement

$$\hat{u}_{i,j+1} = s(k)\hat{u}_{i,j}$$

$s(k)$ est la fonction d'amplification suivante

$$s(k) = \frac{1 - 4a(1-\theta)\sin^2(k\Delta x)}{1 + 4a\theta\sin^2(kx/2)}$$

La stabilité est assurée si et seulement si

$$|s(k)| \leq 1$$

ce qui équivaut à

$$-1 - 4a\theta\tau \leq 1 - 4a(1-\theta)\tau \leq 1 + 4a\theta\tau$$

avec $\tau = \sin^2(kx/2)$. Les nombres a et τ étant positifs, l'inégalité de droite est toujours vérifiée. L'inégalité de gauche conduit à

$$2a(1-2\theta)\tau \leq 1$$

Si $1 - 2\theta \leq 0$, le schéma est universellement stable car $-1 \leq s(k)$ et si $0 \leq \theta < 1/2$, l'inégalité n'est vraie quel que soit τ si et seulement si

$$2a(1-2\theta) \leq 1$$

c'est-à-dire si

$$a = \frac{c\Delta t}{(\Delta x)^2} \leq \frac{1}{2(1-2\theta)}$$

10.5 Méthode de Crank-Nicholson

Considérons l'équation parabolique générale du premier ordre

$$\frac{\partial u}{\partial t} = a(x)\frac{\partial^2 u}{\partial x^2} + b(x)\frac{\partial u}{\partial x} + c(x)u + d(x)$$

Le schéma de Crank-Nicholson

$$\begin{aligned}\frac{u_{i,j+1} - u_{i,j}}{\Delta t} &= a_i\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{2(\Delta x)^2} + b_i\frac{u_{i+1,j} - u_{i-1,j}}{4\Delta x} + c_i\frac{u_{i,j}}{2} \\ &\quad + d_i + a_i\frac{u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1}}{2(\Delta x)^2} \\ &\quad + b_i\frac{u_{i+1,j+1} - u_{i-1,j+1}}{4\Delta x} + c_i\frac{u_{i,j+1}}{2}\end{aligned}$$

est universellement stable.

10.6 Méthode alternative de Peaceman-Rachford-Douglas

Pour résoudre une équation de la forme

$$\frac{\partial u}{\partial t} = a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial y^2}$$

Peaceman-Rachford proposent de remplacer l'équation différentielle par deux équations discrétisées utilisées alternativement sur des périodes $\Delta t/2$. Si on note $v_{i,j}$ le résultat intermédiaire, le schéma s'écrit

$$\begin{cases} \dfrac{v_{i,j} - u^n_{i,j}}{(\Delta t/2)} = a\dfrac{v_{i-1,j} - 2v_{i,j} + v_{i+1,j}}{(\Delta x)^2} + b\dfrac{u^n_{i,j-1} - 2u^n_{i,j} + u^n_{i,j+1}}{(\Delta y)^2} \\ \dfrac{u^{n+1}_{i,j} - v_{i,j}}{(\Delta t/2)} = a\dfrac{v_{i-1,j} - 2v_{i,j} + v_{i+1,j}}{(\Delta x)^2} + b\dfrac{u^{n+1}_{i,j-1} - 2u^{n+1}_{i,j} + u^{n+1}_{i,j+1}}{(\Delta y)^2} \end{cases}$$

En posant

$$\mathcal{L}_x u^n_{i,j} = \frac{u^n_{i-1,j} - 2u^n_{i,j} + u^n_{i+1,j}}{(\Delta x)^2}$$

et

$$\mathcal{L}_y u^n_{i,j} = \frac{u^n_{i,j-1} - 2u^n_{i,j} + u^n_{i,j+1}}{(\Delta y)^2}$$

et en définissant U_n comme la matrice $(u^n_{i,j})$ et V_n la matrice $(v^n_{i,j})$, les équations s'écrivent sous forme matricielle

$$\begin{cases} V_n - U_n = \frac{\Delta t}{2}(a\mathcal{L}_x V_n + b\mathcal{L}_y U_n) \\ U_{n+1} - V_n = \frac{\Delta t}{2}(a\mathcal{L}_x V_n + b\mathcal{L}_y U_{n+1}) \end{cases}$$

À l'étape n, connaissant U_n, on calcule V_n puis U_{n+1} par résolution de systèmes tridiagonaux.

10.7 Exercices

1. *Schéma rétrograde.* On considère l'équation de la chaleur

$$\frac{\partial u}{\partial t} = c\frac{\partial^2 u}{\partial x^2}$$

et le schéma rétrograde suivant

$$\frac{\frac{3}{2}u_{i,j+1} - 2u_{i,j} + \frac{1}{2}u_{i,j-1}}{\Delta t} = c\frac{u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}}{\Delta x^2}$$

Déterminer l'ordre du schéma. Le schéma est-il explicite ou implicite ? Étudier la stabilité du schéma.

2. *Schéma de Dufort et Frankel.* On considère l'équation de la chaleur et le schéma suivant

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

Étudier l'ordre du schéma. Le schéma est-il implicite ou explicite ? Étudier la stabilité de ce schéma.

11
Équations hyperboliques

Les problèmes hyperboliques ont un comportement différent des équations elliptiques ou paraboliques, car ils présentent un phénomène particulier qui est la présence de chocs. Nous envisagerons trois cas d'équations hyperboliques, deux équations hyperboliques linéaires – équation du transport et équation des ondes – de la forme

$$\frac{\partial^2 u}{\partial t^2} - \nabla.(a\nabla u) + cu = f$$

et comme prototype des équations hyperboliques non linéaires, l'équation de Burgers.

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0$$

11.1 Résultats fondamentaux

Les *chocs* ou *ondes de chocs* sont les singularités de la solution d'une équation aux dérivées partielles. Contrairement aux équations elliptiques et paraboliques qui n'ont pas de chocs, les équations hyperboliques peuvent avoir des chocs. Si l'équation hyperbolique est linéaire et admet des chocs, alors ceux-ci figurent dans les conditions initiales ou les conditions limites. Les chocs se propagent le long des caractéristiques. En revanche, si l'équation hyperbolique est non linéaire, des chocs qui ne figurent pas dans les données (conditions initiales ou conditions limites) peuvent apparaître par

focalisation des caractéristiques. La solution d'une équation hyperbolique linéaire ne dépend que partiellement des conditions initiales.

Exemple 1. Considérons l'équation du transport

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = \rho$$

avec comme conditions aux limites

$$\begin{cases} u(x,0) = 0 & \text{si } x < \rho \\ u(x,0) = \rho & \text{si } x > \rho \end{cases}$$

La donnée $u(x,t)$ est discontinue au point $(\rho, 0)$, car on suppose $\rho \neq 0$. L'équation des caractéristiques s'écrit

$$\frac{dt}{1} = \frac{dx}{c} = \frac{du}{\rho}$$

Les caractéristiques sont des droites. La caractéristique qui passe par le point $(x_0, 0)$ a pour équation

$$x = ct + x_0$$

Les solutions sont données par

$$\begin{cases} u(x,0) = \rho(x - x_0)/c & \text{si } x_0 < \rho \\ u(x,0) = \rho(x - x_0)/c + \rho & \text{si } x_0 > \rho \end{cases}$$

Le long d'une caractéristique

$$\lim_{x_0 \to \rho^-} u(x,t) \neq \lim_{x_0 \to \rho^+} u(x,t)$$

La fonction $u(x,t)$ est donc discontinue le long d'une caractéristique. Par conséquent, cet exemple illustre le fait qu'une discontinuité dans les données entraîne une discontinuité de la solution le long des courbes caractéristiques.

Exemple 2. Considérons l'équation du transport

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = \rho$$

avec comme conditions aux limites

$$\begin{cases} u(x,0) = 0 & \text{si } x < \rho \\ u(x,0) = x - \rho & \text{si } x > \rho \end{cases}$$

La donnée $u(x,t)$ est continue au point $(\rho, 0)$, mais $\partial u/\partial x$ est discontinue en ce point. La caractéristique qui passe par le point $(x_0, 0)$ a pour équation $x = ct + x_0$. Les solutions s'écrivent

$$\begin{cases} u(x,0) = \rho(x - x_0)/c & \text{si } x_0 < \rho \\ u(x,0) = \rho(x - x_0)/c + x_0 - \rho & \text{si } x_0 > \rho \end{cases}$$

soit en éliminant x_0,

$$\begin{cases} u(x,0) = \rho t & \text{si } x_0 < \rho \\ u(x,0) = (\rho - c)t + x - \rho & \text{si } x_0 > \rho \end{cases}$$

On voit alors que si $x_0 < \rho$, c'est-à-dire à gauche de la droite caractéristique, la dérivée $\partial u/\partial x = 0$ et que si $x_0 > \rho$, c'est-à-dire à droite de la droite caractéristique, la dérivée vaut $\partial u/\partial x = 1$. Par conséquent, $\partial u/\partial x$ est discontinue le long de la caractéristique. Conclusion : La singularité de la fonction $u(x,t)$ se propage le long des courbes caractéristiques.

Exemple 3. Considérons l'équation des ondes

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

avec pour conditions limites

$$\begin{cases} u(x,t) = 0 & \text{si } x = ct \text{ et } x \leq 1 \\ u(x,t) = x - 1 & \text{si } x = ct \text{ et } x \geq 1 \\ u(x,t) = x & \text{si } x = -ct \end{cases}$$

La solution de cette équation

$$\begin{cases} u(x,t) = \frac{1}{2}(x - ct) & \text{si } x \leq 1 \\ u(x,t) = x - 1 & \text{si } x \geq 1 \end{cases}$$

est une *solution faible* car la fonction $u(x,t)$ est continue, mais n'est pas dérivable au point $(x = 1, t = 1/c)$. En ce point, la condition limite n'est pas dérivable : cette singularité se retrouve dans la solution.

Exemple 4. Considérons l'équation hyperbolique non linéaire de J.M. Burgers

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0$$

Le problème de Cauchy

$$\begin{cases} \dfrac{\partial u}{\partial t} + u \dfrac{\partial u}{\partial x} = 0 & \text{si } x \in \mathbb{R}, \text{ et } t > 0 \\ u(x,0) = u_0(x) \end{cases}$$

admet une solution $u(x,t)$ définie paramétriquement par l'équation des caractéristiques (D_λ)

$$x = u_0(\lambda)t + \lambda$$

Sur ces courbes, la solution est constante : $u = u_0(\lambda)$. Étudions sur des exemples, le fait que le problème soit bien ou mal posé et sous quelles conditions les chocs apparaissent.

Condition d'entropie. Supposons que la fonction u_0 soit la fonction de Heaviside.

$$\begin{cases} u_0(\lambda) = 0 & \text{si } \lambda \leq 0 \\ u_0(\lambda) = 1 & \text{si } \lambda > 0 \end{cases}$$

Dans ces conditions, une solution s'écrit

$$\begin{cases} u(x,t) = 0 & \text{si } x \leq 0 \\ u(x,t) = 1 & \text{si } 0 < t \leq x \\ u(x,t) = x/t & \text{si } 0 \leq x \leq t \end{cases}$$

Bien que la donnée initiale soit discontinue, la solution proposée est continue. Mais ce système admet une deuxième solution avec choc

$$\begin{cases} u(x,t) = 0 & \text{si } x \leq t/2 \\ u(x,t) = 1 & \text{si } x > t/2 \end{cases}$$

Les chocs se déplacent le long de la droite d'équation $x = t/2$. Afin d'éliminer cette solution, nous imposons une condition supplémentaire : c'est la condition d'entropie :

$$\forall x,\ \forall t > 0, \quad u(x-0,t) \geq u(x+0,t)$$

Nous acceptons les chocs à travers lesquels u diminue et nous refusons ceux à travers lesquels u augmente.

Focalisation des caractéristiques. Considérons la donnée initiale

$$\begin{cases} u_0(\lambda) = 0 & \text{si } \lambda \leq 0 \\ u_0(\lambda) = -\lambda^2 & \text{si } \lambda > 0 \end{cases}$$

Les caractéristiques se coupent. Si on note $f(x,t,\lambda) = x - u_0(\lambda) - \lambda$, l'enveloppe des caractéristiques a pour équation le système paramétrique $f = 0$ et $\partial f/\partial\lambda = 0$, soit

$$\begin{cases} t = \dfrac{-1}{u_0(\lambda)} \\ x = \lambda - \dfrac{u_0(\lambda)}{u_0'(\lambda)} \end{cases}$$

C'est-à-dire, ici, la branche d'hyperbole d'équation $t = 1/4x$. On sait que sur chaque caractéristique, la solution est constante et égale à la pente $u_0(\lambda)$, par conséquent si les caractéristiques focalisent, au point d'intersection la fonction $u(x,t)$ prend au moins deux valeurs ce qui est inacceptable (le problème est mal posé). De plus, lorsque u a des discontinuités de ses dérivées, l'équation aux dérivées partielles n'est plus définie car ses dérivées n'existent pas. Afin de remédier à ce problème, on transforme l'équation

aux dérivées partielles en une formulation variationnelle : Pour toute fonction ϕ de classe $C^1(\mathbb{R}^2)$ à support borné

$$\int_{-\infty}^{+\infty} dx \int_0^{+\infty} (u\frac{\partial\phi}{\partial t} + \frac{u^2}{2}\frac{\partial\phi}{\partial x})dt + \int_{-\infty}^{+\infty} u_0(x)\phi(x,0)dx = 0$$

On démontre alors que si u est une solution de classe C^1 du problème de Cauchy, u vérifie l'équation variationnelle et que inversement, si u vérifie l'équation variationnelle et si u est de classe C^1, alors u est une solution classique du problème de Cauchy.

Condition de Rankine-Hugoniot. On dit que u est une *discontinuité de première espèce* le long d'une courbe $\mathcal{C}$, si u n'est pas continue, u et ses dérivées admettent une dérivée à droite et une dérivée à gauche qui sont des fonctions continues de l'abscisse curviligne sur $\mathcal{C}$. On démontre que si u est une fonction de classe C^1 par morceaux présentant des discontinuités de première espèce sur une courbe $\mathcal{C}$, alors la pente du choc est égale à la valeur moyenne des valeurs de part et d'autre du choc.

$$\left.\frac{dx}{dt}\right|_{\mathcal{C}} = \frac{u^- + u^+}{2}$$

Pour l'équation de Burgers généralisée, la condition de Rankine-Hugoniot exprime la continuité du flot $f(x)$ à travers la courbe des discontinuités $x = x(t)$.

Problème bien posé. On démontre que si u_0 est une fonction mesurable bornée, le problème de Cauchy reformulé en

$$\begin{cases} \dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} = 0 & \text{si } x \in \mathbb{R}, \text{ et } t > 0 \\ u(x,0) = u_0(x) & \\ \left.\frac{dx}{dt}\right|_{\mathcal{C}} = (u^- + u^+)/2 & \text{Condition Rankine-Hugoniot} \\ \forall x,\ \forall t > 0, \quad u(x-0,t) \geq u(x+0,t) & \text{Condition d'entropie} \end{cases}$$

admet une solution unique. Ce problème est équivalent au problème formulé en termes variationnels

$$\begin{cases} \forall \phi \in C^1(\mathbb{R}^2) \text{ à support borné} \\ \\ \int_{-\infty}^{+\infty} dx \int_0^{+\infty} (u\frac{\partial\phi}{\partial t} + \frac{u^2}{2}\frac{\partial\phi}{\partial x})dt + \int_{-\infty}^{+\infty} u_0(x)\phi(x,0)dx = 0 \\ \\ \forall x,\ \forall t > 0, \quad u(x-0,t) \geq u(x+0,t) \end{cases}$$

qui admet une solution unique, si u_0 est une fonction mesurable bornée.

11.2 Équation du transport

L'équation du transport est le prototype des équations hyperboliques linéaires du premier ordre.

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0$$

Elle admet une solution de la forme

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

qui représente l'évolution d'une onde progressive se propageant à la vitesse c.

11.2.1 Schéma de Lax

Pour l'équation de transport,

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0$$

le schéma de Lax (1954)

$$\frac{u_{i,j+1} - \frac{1}{2}(u_{i-1,j} + u_{i+1,j})}{\Delta t} + c\frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x} = 0$$

est un schéma à un niveau de temps, stable et convergent s'il vérifie la *condition de Courant-Friedrichs-Lewy* (*CFL*)

$$a = c\frac{\Delta t}{\Delta x} \leq 1$$

Par transformation de Fourier, on vérifie que la fonction d'amplification du schéma est

$$s(k) = \cos(k\Delta x) - ia\sin(k\Delta x)$$

11.2.2 Schéma décentré

Pour l'équation de transport,

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0$$

le schéma décentré

$$\frac{u_{i,j+1} - u_{i,j}}{\Delta t} + c\frac{u_{i,j} - u_{i-1,j}}{\Delta x} = 0$$

est un schéma stable et convergent sous la condition *CFL*

$$a = c\frac{\Delta t}{\Delta x} \leq 1$$

La fonction d'amplification du schéma s'écrit

$$s(k) = (1 - a + a\cos(k\Delta x)) - ia\sin(k\Delta x)$$

11.2.3 Schéma saute-mouton

Pour l'équation de transport,

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0$$

le schéma saute-mouton est un schéma explicite à trois niveaux de temps

$$\frac{u_{i,j+1} - u_{i,j-1}}{2\Delta t} + c\frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x} = 0$$

qui est stable et convergent sous la condition *CFL*

$$a = c\frac{\Delta t}{\Delta x} \leq 1$$

La matrice d'amplification du schéma s'écrit

$$\begin{pmatrix} \hat{u}_{i+1} \\ \hat{u}_i \end{pmatrix} = \begin{pmatrix} \alpha & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \hat{u}_i \\ \hat{u}_{i-1} \end{pmatrix}$$

avec

$$\alpha = -2ia\sin(k\Delta x)$$

11.2.4 Schéma de Lax-Wendroff

Pour l'équation de transport,

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0$$

le schéma de Lax-Wendroff est obtenu à partir d'un développement de Taylor au deuxième ordre

$$u(x_i, t_{j+1}) = u(x_i, t_j) + \Delta t\frac{\partial u}{\partial t}(x_i, t_j) + \frac{\Delta t^2}{2}\frac{\partial^2 u}{\partial t^2}(x_i, t_j)$$

Comme

$$\frac{\partial u}{\partial t} = -c\frac{\partial u}{\partial x}$$

et

$$\frac{\partial^2 u}{\partial t^2} = -c\frac{\partial^2 u}{\partial t \partial x} = -c(-c\frac{\partial^2 u}{\partial x^2}) = c^2\frac{\partial^2 u}{\partial x^2}$$

le développement devient

$$u(x_i, t_{j+1}) = u(x_i, t_j) + c\Delta t\frac{\partial u}{\partial x}(x_i, t_j) + c^2\frac{\Delta t^2}{2}\frac{\partial^2 u}{\partial x^2}(x_i, t_j)$$

En utilisant les discrétisations usuelles

$$\frac{\partial u}{\partial x}(x_i, t_j) = \frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x}$$

et

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_j) = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2}$$

le schéma s'écrit

$$\frac{u_{i,j+1} - u_{i,j}}{\Delta t} = c\frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x} + c^2\frac{\Delta t}{2}\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2}$$

Le schéma de Lax-Wendroff est un schéma explicite, d'ordre 2 en espace et d'ordre 2 en temps, stable et convergent sous la condition de Courant-Friedrichs-Lewy

$$a = c\frac{\Delta t}{\Delta x} \leq 1$$

La fonction d'amplification du schéma s'écrit

$$s(k) = (1 - a^2 + a^2\cos(k\Delta x)) - ia\sin(k\Delta x)$$

11.3 Équation des ondes

L'équation des ondes est le prototype des équations hyperboliques linéaires du deuxième ordre

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}$$

Elle admet pour solution, toute fonction de la forme

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

où f et g sont des fonctions arbitraires de classe C^2, représentant la somme d'une onde progressive et d'une onde régressive. Plus précisément, considérons le problème de Cauchy suivant

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = c^2\dfrac{\partial^2 u}{\partial x^2} & \text{si } x \in \mathbb{R}, \text{ et } t > 0 \\ \\ u(x,0) = u_0(x) \\ \\ \dfrac{\partial u}{\partial t}(x,0) = u_1(x) \end{cases}$$

Si u_0 est de classe C^p et si u_1 est de classe C^{p-1}, le problème de Cauchy admet une solution classique de classe C^p

$$u(x,t) = u_0(x + ct)/2 + u_0(x - ct)/2 + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(\tau)d\tau$$

En termes de distribution, l'équation

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = u_0(x)\delta(t) + u_1(x)\delta'(t)$$

admet une solution faible

$$u(x,t) = E(x,t) * u_0(x) + \frac{\partial E}{\partial t} * u_1(x)$$

avec

$$E(x,t) = \frac{1}{2}\chi_{|x|<ct} = \begin{cases} 1/2 & \text{si} \quad |x| < ct \\ 0 & \text{si} \quad |x| > ct \end{cases}$$

Remarquons que si on pose $v = \partial u/\partial x$ et $w = \partial u/c\partial t$, l'équation des ondes s'écrit sous la forme du système

$$\begin{cases} \dfrac{\partial v}{\partial t} = c\dfrac{\partial w}{\partial x} \\ \dfrac{\partial w}{\partial t} = \dfrac{\partial v}{\partial x} \end{cases}$$

Plus généralement, le problème

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} = c^2 \Delta u + f(x,t) \qquad \text{si } x \in \mathbb{R}^n, \text{ et } t \in \mathbb{R} \\ \\ u(x,0) = u_0(x) \\ \\ \dfrac{\partial u}{\partial t}(x,0) = u_1(x) \end{cases}$$

admet une solution qui, si on note $f_t(x)$ la fonction $f(x,t)$, est donnée par la formule suivante

$$\begin{aligned} u(x,t) &= \frac{1}{c}E_{ct} * u_1(x) + \partial_t \left(\frac{1}{c} E_{ct} * u_0 \right)(x) + \\ &\quad + \int_0^t \frac{1}{c} E_{c(t-s)} * f_t(x) ds \end{aligned}$$

$E_t(x)$ est la fonction notée aussi $E(x,t)$ qui admet comme transformée de Fourier la fonction

$$\widehat{E}_t(\xi) = \frac{\sin(t\,\|\xi\|)}{\|\xi\|}$$

11.3.1 Méthode du theta-schéma

Considérons l'équation des ondes

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

et écrivons le θ-schéma associé à l'équation des ondes sous la forme

$$\frac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{(\Delta t)^2}-c^2\frac{\theta Au_{i,j+1}+(1-2\theta)Au_{i,j}+\theta Au_{i,j-1}}{(\Delta x)^2}=0$$

avec

$$Au_{i,j}=u_{i+1,j}-2u_{i,j}+u_{i-1,j}$$

Le schéma est explicite si $\theta=0$ et implicite dans tous les autres cas. Si $0\leq\theta\leq 1/4$, le schéma est stable si

$$c\frac{\Delta t}{\Delta x}\leq\frac{1}{\sqrt{1-4\theta}}$$

si $1/4\leq\theta\leq 1$, le schéma est universellement stable. En effet, par transformation de Fourier, en remarquant que

$$(Au_{i,j})\hat{}=(e^{ik\Delta x}+e^{ik\Delta x}-2)\hat{u}_{i,j}=-4\sin^2(k\Delta x)\hat{u}_{i,j}$$

Le schéma s'écrit

$$(1+\alpha^2\theta)\hat{u}_{i,j+1}-(2-(1-2\theta)\alpha^2)\hat{u}_{i,j}+(1+\alpha^2\theta)\hat{u}_{i,j-1}=0$$

avec

$$\alpha=2c\frac{\Delta t}{\Delta x}\sin^2(k\Delta x)$$

d'où la matrice d'amplification

$$\begin{pmatrix}\hat{u}_{i,j+1}\\ \hat{v}_{i,j+1}\end{pmatrix}=\begin{pmatrix}a&-1\\1&0\end{pmatrix}\begin{pmatrix}\hat{u}_{i,j}\\ \hat{v}_{i,j}\end{pmatrix}$$

avec

$$a=\frac{2-(1-2\theta)\alpha^2}{1+\alpha^2\theta}$$

Le polynôme caractéristique $\lambda^2-a\lambda+1=0$ admet comme discriminant $\Delta=a^2-4$. Ce discriminant est négatif si et seulement si $\alpha^2(1-4\theta)\leq 4$ ce qui équivaut à la condition

$$c\frac{\Delta t}{\Delta x}\leq\frac{1}{\sqrt{1-4\theta}}$$

Si $\theta\geq 1/4$, le discriminant est négatif, les racines sont complexes conjuguées et de module 1 (car le produit des racines vaut 1), le schéma est toujours stable. Si $\theta<1/4$, le rayon spectral de la matrice d'amplification est inférieur à 1 si et seulement si

$$c\frac{\Delta t}{\Delta x}\leq\frac{1}{\sqrt{1-4\theta}}$$

11.3.2 Schéma de Lax

Pour l'équation des ondes écrite sous forme d'un système

$$\begin{cases} \dfrac{\partial v}{\partial t} = c\dfrac{\partial w}{\partial x} \\ \dfrac{\partial w}{\partial t} = \dfrac{\partial v}{\partial x} \end{cases}$$

le schéma de Lax

$$\begin{cases} v_{i,j+1} = \dfrac{1}{2}(v_{i+1,j} + v_{i-1,j}) + c\dfrac{\Delta t}{2\Delta x}(w_{i+1,j} - v_{i-1,j}) \\ w_{i,j+1} = \dfrac{1}{2}(w_{i+1,j} + w_{i-1,j}) + c\dfrac{\Delta t}{2\Delta x}(v_{i+1,j} - v_{i-1,j}) \end{cases}$$

est un schéma à deux niveaux, du premier ordre, stable si la condition *CFL* est vérifiée

$$c\frac{\Delta t}{\Delta x} \leq 1$$

Par transformée de Fourier, on vérifie que la matrice d'amplification du schéma s'écrit

$$S = \begin{pmatrix} \cos(k\Delta x) & ia \\ ia & \cos(k\Delta x) \end{pmatrix}$$

avec

$$a = c\frac{\Delta t}{\Delta x}\sin(k\Delta x)$$

Comme le discriminant du polynôme caractéristique est négatif $\Delta = -a^2$, les racines sont complexes conjuguées et de module

$$|\lambda|^2 = 1 - \sin^2(k\Delta x)(1 - c^2\frac{\Delta t^2}{\Delta x^2})$$

Le rayon spectral est donc inférieur à 1, si la condition *CFL* est vérifiée.

11.3.3 Schéma saute-mouton

Le schéma saute-mouton pour l'équation des ondes écrite sous forme d'un système du premier ordre

$$\begin{cases} v_{i,j+1} = v_{i,j-1} + c\dfrac{\Delta t}{\Delta x}(w_{i+1,j} - w_{i-1,j}) \\ w_{i,j+1} = w_{i,j-1} + c\dfrac{\Delta t}{\Delta x}(v_{i+1,j} - v_{i-1,j}) \end{cases}$$

est un schéma à trois niveaux, explicite et du second ordre, stable s'il vérifie la condition *CFL*.

11.3.4 Schéma de Lax-Wendroff

Le schéma de Lax-Wendroff, avec $\lambda = \Delta t/\Delta x$

$$\begin{cases} v_{i,j+1} = v_{i,j} + c\dfrac{\lambda}{2}(w_{i+1,j} - w_{i-1,j}) + \dfrac{c^2\lambda^2}{2}(v_{i+1,j} - 2v_{i,j} + v_{i-1,j}) \\ w_{i,j+1} = w_{i,j} + c\dfrac{\lambda}{2}(v_{i+1,j} - v_{i-1,j}) + \dfrac{c^2\lambda^2}{2}(w_{i+1,j} - 2w_{i,j} + w_{i-1,j}) \end{cases}$$

est un schéma stable sous la condition *CFL*. Sa matrice d'amplification s'écrit

$$S = \begin{pmatrix} 1+a & ib \\ ib & 1+a \end{pmatrix}$$

avec $a = c^2\lambda^2(\cos(k\Delta x) - 1)$ et $b = c\lambda\sin(k\Delta x)$.

11.4 Équation de Burgers

L'équation de J.M. Burgers

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$$

se généralise sous la forme

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0$$

Lorsque f est une fonction convexe, l'équation de Burgers associée à la condition initiale

$$u(x,0) = \begin{cases} u_g & \text{si } x < 0 \\ u_d & \text{si } x > 0 \end{cases}$$

admet une solution unique $u(x,t) = w(x/t, u_g, u_d)$, où w est la fonction suivante appelée *solveur de Riemann* et g la fonction telle que $f'(g(x)) = x$

$$w(y,u,v) = \begin{cases} u & \text{si } y < f'(u) \\ g(y) & \text{si } f'(u) < y < f'(v) \\ v & \text{si } y > f'(v) \end{cases}$$

Dans la suite, nous supposerons que f est convexe et de classe C^2.

11.4.1 Schéma de Lax-Friedrichs

L'équation de Burgers

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0$$

admet comme schéma de discrétisation, le schéma de Lax suivant

$$\frac{u_{i,j+1} - \frac{1}{2}(u_{i-1,j} + u_{i+1,j})}{\Delta t} + \frac{f(u_{i+1,j}) - f(u_{i-1,j})}{2\Delta x} = 0$$

Ce schéma est un schéma explicite à deux niveaux en temps du premier ordre, stable s'il vérifie la condition de *Courant-Friedrichs-Lewy* (*CFL*)

$$\frac{\Delta t}{\Delta x} \sup |f'(u_{i,j})| \leq 1$$

En effet, posons

$$\alpha = \sup |f'(u_{i,j})|$$

en utilisant la relation

$$f(u_{i+1,j}) - f(u_{i-1,j}) = \frac{\alpha}{2}(u_{i+1,j} - u_{i-1,j})$$

le schéma linéarisé s'écrit sous la forme

$$u_{i,j+1} = \frac{1}{2}(u_{i-1,j} + u_{i+1,j}) - \frac{\Delta t}{2\Delta x}\alpha(u_{i+1,j} - u_{i-1,j})$$

En prenant la transformée de Fourier, on a

$$\hat{u}_{i,j+1} = s(k)\hat{u}_{i,j}$$

La fonction d'amplification, qui vaut

$$s(k) = \cos(k\Delta x) - i\alpha\frac{\Delta t}{\Delta x}\sin(k\Delta x)$$

est de module inférieur à 1 sous la condition *CFL*.

Remarquons que le schéma de Lax peut aussi s'écrire sous la forme

$$u_{i,j+1} = u_{i,j} - \frac{\Delta t}{2\Delta x}(f(u_{i+1,j}) - f(u_{i-1,j})) + \frac{1}{2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j})$$

qui correspond à la discrétisation de l'équation parabolique

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \varepsilon\frac{\partial^2 u}{\partial x^2}$$

où on a posé $\varepsilon = \Delta x$. Le schéma de Lax-Friedrichs introduit un terme supplémentaire qu'on appelle la *viscosité numérique* (par analogie avec la mécanique des fluides). Lorsque ε tend vers zéro, l'équation parabolique tend vers l'équation hyperbolique.

11.4.2 Schéma saute-mouton

Le schéma saute-mouton de l'équation de Burgers

$$u_{i,j+1} = u_{i,j-1} - \frac{\Delta t}{2\Delta x}(f(u_{i+1,j}) - f(u_{i-1,j}))$$

est un schéma explicite à trois niveaux, d'ordre 2, stable sous la condition *CFL*.

11.4.3 Schéma de Lax-Wendroff

Pour établir le schéma de Lax-Wendroff de l'équation généralisée de Burgers, écrivons un développement de Taylor à l'ordre 2

$$u(x,t+\Delta t) = u(x,t) + \Delta t \frac{\partial u}{\partial t}(x,t) + \frac{\Delta t^2}{2}\frac{\partial^2 u}{\partial t^2}(x,t) + O(\Delta t^3)$$

soit en termes discrétisés

$$u_{i,j+1} = u_{i,j} + \Delta t \left(\frac{\partial u}{\partial t}\right)_{i,j} + \frac{\Delta t^2}{2}\left(\frac{\partial^2 u}{\partial t^2}\right)_{i,j} + O(\Delta t^3)$$

Discrétisons en différences centrées

$$\left(\frac{\partial u}{\partial t}\right)_{i,j} = -\left(\frac{\partial f(u)}{\partial x}\right)_{i,j} = \frac{-f(u_{i+1,j}) + f(u_{i-1,j})}{2\Delta x} + O(\Delta x^2)$$

Remarquons que

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial t}\left(-\frac{\partial f(u)}{\partial x}\right) = -\frac{\partial}{\partial x}\left(f'(u)\frac{\partial u}{\partial t}\right) = \frac{\partial}{\partial x}\left(f'(u)\frac{\partial f(u)}{\partial x}\right)$$

Pour un $\theta \in [0,1]$, si on pose $\partial_x f(u) = \partial f(u)/\partial x$, on a

$$\begin{aligned}\frac{\partial}{\partial x}\left(f'(u)\partial_x f(u)\right) &= f'(u(x+\theta\Delta x,t))\frac{f(u(x+\Delta x,t) - f(u(x-\Delta x,t))}{\Delta x^2} \\ &\quad -f'(u(x+(\theta-1)\Delta x,t))\frac{f(u(x,t) - f(u(x-\Delta x,t))}{\Delta x^2} \\ &\quad +O(\Delta x)\end{aligned}$$

expression dans laquelle on a remplacé

$$\frac{\partial f(u)}{\partial x} = \frac{f(u(x+\Delta x,t) - f(u(x-\Delta x,t))}{\Delta x^2}$$

d'où l'expression discrétisée

$$\begin{aligned}\frac{\partial}{\partial x}\left(f'(u)\frac{\partial f(u)}{\partial x}\right)_{i,j} &= \frac{f'(u_{i+\theta,j})}{\Delta x^2}(f(u_{i+1,j}) - f(u_{i,j})) \\ &\quad -\frac{f'(u_{i+(\theta-1),j})}{\Delta x^2}(f(u_{i,j}) - f(u_{i-1,j}))\end{aligned}$$

ce qui conduit, pour $\theta = 1/2$, au schéma de Lax-Wendroff

$$\begin{aligned} u_{i,j+1} &= u_{i,j} - \frac{\lambda^2}{2}(f(u_{i+1,j}) - f(u_{i-1,j})) + \\ &\quad + \frac{\lambda^2}{2} f'(u_{i+1/2,j})(f(u_{i+1,j}) - f(u_{i,j})) \\ &\quad - \frac{\lambda^2}{2} f'(u_{i-1/2,j})(f(u_{i,j}) - f(u_{i-1,j})) \end{aligned}$$

avec $\lambda = \Delta t/\Delta x$ et

$$f'(u_{i\pm 1/2,j}) = \frac{f'(u_{i,j}) + f'(u_{i\pm 1,j})}{2}$$

Le schéma de Lax-Wendroff est un schéma explicite à deux niveaux du second ordre, stable sous la condition *CFL*.

11.4.4 Schéma d'Engquist-Osher

Pour l'équation de Burgers généralisée

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0$$

le schéma d'Engquist-Osher est une généralisation de celui de LaxWendroff.

$$u_{i,j+1} = u_{i,j} - \lambda(\Phi(u_{i,j}, u_{i+1,j}) - \Phi(u_{i-1,j}, u_{i,j}))$$

où $\lambda = \Delta t/\Delta x$, et le *flux numérique* est donné par

$$\Phi(u, v) = \frac{1}{2}(f(v) - f(u)) - \frac{\lambda}{2}\int_u^v |f'(\tau)|\, d\tau$$

Le terme intégral discrétise la viscosité numérique. Le schéma est du premier ordre, stable sous la condition *CFL*.

11.4.5 Schéma de Godunov

Pour l'équation de Burgers généralisée

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0$$

le schéma de Godunov introduit le *flux numérique* $\Phi(u, v)$ à partir du solveur de Riemann $w(0, u, v)$

$$\Phi(u, v) = w(0, u, v) = \begin{cases} f(u) & \text{si } f'(u) > 0 \\ f \circ g(0) & \text{si } f'(u) < 0 \text{ et } f'(v) > 0 \\ f(v) & \text{si } f'(v) < 0 \end{cases}$$

où g est la fonction telle que $f'(g(x)) = x$. Le schéma de Godunov

$$u_{i,j+1} = u_{i,j} - \frac{\Delta t}{\Delta x}(\Phi(u_{i,j}, u_{i+1,j}) - \Phi(u_{i-1,j}, u_{i,j}))$$

est un schéma du premier ordre, stable sous la condition *CFL*.

11.4.6 Schémas de Lerat-Peyret

Pour l'équation de Burgers généralisée

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0$$

les schémas S_α^β de Lerat-Peyret sont des schémas d'ordre 2, paramétrés par α et β. Selon les valeurs des paramètres, le schéma s'appelle aussi *schéma de Mac-Cormack* ($\alpha = 1$, $\beta = 0$), ou encore *schéma de Richtmeyer* ($\alpha = \beta = 1/2$). La méthode de résolution est une méthode de prédiction-correction, dans laquelle le prédicteur vaut

$$p_i = (1-\beta)u_{i,j} + \beta u_{i+1,j} - \alpha\frac{\Delta t}{\Delta x}(f(u_{i+1,j}) - f(u_{i,j}))$$

et le correcteur

$$\begin{aligned} u_{i,j+1} \quad = \quad & u_{i,j} - \frac{\Delta t}{2\alpha\Delta x}((\alpha-\beta)f(u_{i+1,j}) + (2\beta-1)f(u_{i,j}) \\ & +(1-\alpha-\beta)f(u_{i-1,j}) + f(p_i) - f(p_{i-1})) \end{aligned}$$

Les schémas de Lerat-Peyret sont stables sous la condition *CFL*.

11.5 Exercices

1. *Schéma décentré.* On considère l'équation des ondes

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}$$

et le schéma décentré suivant :

$$\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{(\Delta t)^2} = c^2\frac{u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}}{(\Delta x)^2}$$

Ce schéma est-il explicite ou implicite ? Étudier la stabilité du schéma.

2. *Équation des ondes couplée à une équation de la chaleur.* On considère le système d'équations suivant

$$\begin{cases} \dfrac{\partial u}{\partial t} - c\dfrac{\partial v}{\partial x} - d\dfrac{\partial w}{\partial x} = 0 \\ \dfrac{\partial v}{\partial t} - c\dfrac{\partial u}{\partial x} = 0 \\ \dfrac{\partial w}{\partial t} - \sigma\dfrac{\partial^2 w}{\partial x^2} - d\dfrac{\partial u}{\partial x} = 0 \end{cases}$$

dans lequel x est un réel et t représente le temps ($t > 0$).
1) Lorsque le paramètre d est nul, le système se découple en une équation des ondes et une équation de la chaleur. Lorsque $\sigma = 0$, montrer que l'élimination de v et w conduit à l'équation

$$\frac{\partial^2 u}{\partial t^2} - (c^2 + d^2)\frac{\partial^2 u}{\partial x^2} = 0$$

2) On considère le schéma numérique suivant, où θ est un paramètre de $[1/2, 1]$

$$\begin{cases} \dfrac{1}{\Delta t}(u_{i,n+1} - u_{i,n}) - \dfrac{c}{\Delta x}(v_{i+1,n} - v_{i,n}) - \dfrac{d}{\Delta x}(w_{i+1,n} - w_{i,n}) = 0 \\ \dfrac{1}{\Delta t}(v_{i,n+1} - v_{i,n}) - \dfrac{c}{\Delta x}(u_{i,n+1} - u_{i-1,n+1}) = 0 \\ \dfrac{1}{\Delta t}(w_{i,n+1} - w_{i,n}) - \dfrac{\sigma}{\Delta x^2}[\theta(w_{i+1,n+1} - 2w_{i,n+1} + w_{i-1,n+1}) + \\ (1-\theta)(w_{i+1,n} - 2w_{i,n} + w_{i-1,n})] - \dfrac{d}{\Delta x}(u_{i,n+1} - u_{i-1,n+1}) = 0 \end{cases}$$

Le schéma est-il implicite ou explicite ?
3) On pose $\gamma = c\Delta t/\Delta x$, $\delta = d\Delta t/\Delta x$, $\lambda = \sigma\Delta t/\Delta x^2$, et $\xi = k\Delta x$. Démontrer par transformation de Fourier que si on pose $X_n = (\widehat{u_n}, \widehat{v_n}, \widehat{w_n})$, on obtient un système de la forme

$$AX_{n+1} = BX_n$$

avec

$$A = \begin{pmatrix} 1 & 0 & 0 \\ \gamma(e^{-i\xi} - 1) & 1 & 0 \\ \delta(e^{-i\xi} - 1) & 0 & 1 + 4\lambda\theta\sin^2(\xi/2) \end{pmatrix}$$

et

$$B = \begin{pmatrix} 1 & \gamma(e^{i\xi} - 1) & \delta(e^{i\xi} - 1) \\ 0 & 1 & 0 \\ 0 & 0 & 1 - 4\lambda(1-\theta)\sin^2(\xi/2) \end{pmatrix}$$

4) L'étude de la stabilité de ce schéma se ramène à prouver que les racines de $det(\mu A - B) = 0$ sont de module inférieur ou égal à 1. Pour cela, on pose $\mu = \dfrac{1+z}{1-z}$. Donner une condition équivalente à la condition $|\mu| \leq 1$. En posant $z = e^{i\xi} - 1$, et en appliquant le théorème de Routh-Hurwitz au polynôme $Q(z) = (1-z^3)det(\mu A - B)$, montrer que la condition de stabilité s'exprime par la positivité des fonctions

$$\phi_1(x) = 1 + (2\lambda(2\theta - 1) - \delta^2 - \gamma^2)x - 2\lambda(2\theta - 1)\gamma^2 x^2$$

et

$$\phi_2(x) = 1 - \gamma^2 x$$

Une condition nécessaire et suffisante pour que les racines du polynôme $Q(z) = a_3 z^3 + a_2 z^2 + a_1 z + a_0$ (avec $a_0 > 0$) appartiennent au demi-plan $Re(z) \leq 0$ est que $a_3 \geq 0$, $a_2 \geq 0$ et $a_1 a_2 - a_0 a_3 \geq 0$ (Théorème de Routh-Hurwitz.)

5) Montrer qu'une condition nécessaire et suffisante pour la positivité de ϕ_1 est que $\phi_1(1) \geq 0$. En déduire que la condition de stabilité (dans le cas où $\theta \geq 1/2$) s'écrit

$$\frac{c^2 \Delta t^2}{\Delta x^2} + \frac{d^2 \Delta t^2}{\Delta x^2 + 2\sigma(2\theta - 1)\Delta t} \leq 1$$

3. *Équation des ondes avec viscosité.* On considère le problème visco-élastique suivant

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - c^2 \dfrac{\partial^2 u}{\partial x^2} - \sigma \dfrac{\partial^3 u}{\partial x^2 \partial t} = 0 \qquad x \in R, \quad t > 0 \\ \\ u(x,0) = u_0(x) \\ \\ \dfrac{\partial u}{\partial t}(x,0) = u_1(x) \end{cases}$$

On suppose que c et σ sont des nombres positifs ou nuls et on note A l'opérateur

$$Au_{i,j} = \frac{-u_{i+1,j} + 2u_{i,j} - u_{i-1,j}}{(\Delta x)^2}$$

et on considère le schéma de discrétisation

$$\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{(\Delta t)^2} + c^2 Au_{i,j} + \frac{\sigma}{\Delta t} A(u_{i,j} - u_{i,j-1}) = 0$$

1) Déterminer la matrice d'amplification du schéma.

2) Montrer que la condition de stabilité du schéma s'écrit

$$c^2 \frac{(\Delta t)^2}{(\Delta x)^2} + 2\sigma \frac{\Delta t}{(\Delta x)^2} \leq 1$$

3) Que retrouve-t-on si $\sigma = 0$? si $c = 0$?

12
Méthode des éléments finis

La méthode des éléments finis est apparue dans les années 50 et s'est développée grâce à la mise en place d'un nombre croissant d'éléments finis. Du fait de l'équivalence des problèmes variationnel et énergétique, plusieurs méthodes se sont développées en partant d'une formulation ou d'une autre. De plus, le choix de l'espace des fonctions tests V a largement contribué à la diversité des méthodes.

12.1 Principe de la méthode

Nous avons vu que le problème consistant à résoudre une équation différentielle

$$\begin{cases} \mathcal{L}u = f & x \in \Omega \\ u = u_0 & x \in \partial\ \Omega \end{cases}$$

supposée admettre une solution unique est équivalent à un problème variationnel affirmant la nullité d'une forme intégrale pour un ensemble hilbertien de fonctions de tests V. Ce problème est lui-même équivalent au problème de minimisation de l'énergie potentielle, qui consiste à trouver une fonction u telle que

$$J(u) \leq J(v), \quad \forall v \in V$$

La méthode des éléments finis se propose de déterminer la solution du problème variationnel sur un sous-espace discrétisé V_h de V. Elle consiste, à partir d'une équation différentielle, à écrire la formulation variationnelle

faible du problème. Puis, à construire un espace d'approximation $V_h \subset V$, en procédant au *maillage* du domaine, c'est-à-dire en découpant le domaine $\overline{\Omega} = \Omega \cup \partial\ \Omega$ de $\mathbb{R}^n$ en un nombre fini de sous-domaines, disjoints deux à deux, sur lesquels on choisit un nombre fini de points appelés *nœuds*. Les fonctions de V_h sont définies par morceaux sur chaque nœud intérieur au domaine, vérifient les conditions limites aux bords du domaine et s'expriment comme combinaisons linéaires d'éléments simples (en général des polynômes de degré 1, 2 ou 3) appelées *fonctions de forme*. Ces fonctions définies localement sur chaque nœud intérieur sont continues sur l'ensemble du domaine et vérifient les conditions aux limites. Dans le cas d'approximation par des éléments lagrangiens, les dérivées premières sont discontinues aux nœuds intérieurs, mais continues dans le cas d'éléments hermitiens. En exprimant la formule variationnelle par les éléments de V_h ainsi définis, on montre que l'équation se transforme en un système matriciel dans lequel les inconnues sont les valeurs de la fonction solution en chaque nœud. En choisissant des éléments de structures géométriques simples et identiques, le traitement matriciel peut être systématisé et effectué sur un seul élément de référence. On procède alors à la détermination des matrices de masse et de rigidité élémentaires associées à un élément, puis on assemble ces matrices en les plongeant dans une matrice unique représentant l'ensemble du domaine. Le système matriciel obtenu est de type bande, ce qui facilite le stockage des données. La résolution de ce système conduit à la détermination des valeurs de la solution des équations de départ en chaque nœud du maillage.

12.2 Formulation variationnelle

Pour illustrer la mise en œuvre de la méthode des éléments finis, nous traiterons dans les paragraphes suivants, l'équation

$$\alpha u(x) - \beta \frac{d^2u}{dx^2} = f(x)$$

pour x défini sur un intervalle $[a, b]$. Le problème est mis sous forme variationnelle forte

$$\alpha \int_a^b u(x)v(x)dx - \beta \int_a^b \frac{d^2u}{dx^2}v(x)dx = \int_a^b f(x)v(x)dx$$

soit en intégrant par parties la deuxième intégrale, la forme variationnelle faible s'écrit

$$\alpha \int_a^b u(x)v(x)dx + \beta \int_a^b \frac{du}{dx}\frac{dv}{dx}dx = \int_a^b f(x)v(x)dx + \beta \left[v(x)\frac{du}{dx}\right]_a^b$$

12.3 Maillage et fonctions de forme

Le domaine $\Omega = [a, b]$ est découpé en m sous-domaines élémentaires Ω_j, correspondant à la subdivision $a = x_1, x_2, ..., x_{m-1}, x_m = b$. On se propose de résoudre l'équation variationnelle sur chaque sous-domaine. Afin de discrétiser la forme intégrale obtenue, on remplace les fonctions u et v par une approximation sur une base de fonctions de forme

$$u(x) = \sum_{j=1}^{m} u_j N_j(x)$$

et

$$v(x) = \sum_{i=1}^{m} v_i N_i(x)$$

L'équation devient

$$\begin{aligned} & \alpha \sum_{i,j=1}^{m} u_i \int_a^b N_i(x) N_j(x) dx \; v_j + \beta \sum_{i,j=1}^{m} u_i \int_a^b \frac{dN_i}{dx} \frac{dN_j}{dx} dx \; v_j \\ = & \sum_{j=1}^{m} v_j \int_a^b f(x) N_j(x) dx + \beta (v(b)u'(b) - v(a)u'(a)) \end{aligned}$$

soit en notant

$$U = (u_1, u_2, ..., u_m), \quad V = (v_1, v_2, ..., v_m), \quad S = (-u'(a), 0, ..., 0, u'(b))$$

et en définissant la *matrice de masse* par

$$M_{i,j} = \int_a^b N_i(x) N_j(x) dx$$

la *matrice de rigidité*

$$K_{i,j} = \int_a^b \frac{dN_i}{dx} \frac{dN_j}{dx} dx$$

et le *vecteur de charge* par

$$F_j = \int_a^b f(x) N_j(x) dx$$

l'équation s'écrit sous forme matricielle

$$V(\alpha M + \beta K)U = V(F + \beta S)$$

d'où on déduit l'expression de U

$$U = (\alpha M + \beta K)^{-1}(F + \beta S)$$

Le problème est donc résolu. À partir de cette équation, il suffit de calculer les valeurs des matrices pour connaître les solutions aux nœuds du maillage. Pour cela, il faut aussi choisir les fonctions de forme. Nous présentons plusieurs choix dans les paragraphes suivants.

12.4 Matrices de masse et de rigidité élémentaires

En développant les fonctions de forme sur une base de polynômes (P_j), on définit les matrices de base élémentaires. La matrice de masse élémentaire

$$\widehat{M}_{i,j} = \int_a^b P_i(x)P_j(x)dx$$

La matrice de rigidité élémentaire

$$\widehat{K}_{i,j} = \int_a^b \frac{dP_i}{dx}\frac{dP_j}{dx}dx$$

Le vecteur de charge élémentaire

$$\widehat{F}_j = \int_a^b f(x)P_j(\frac{x-x_j}{h})dx$$

où h est le pas de discrétisation $h = x_j - x_{j-1}$.

12.5 Éléments finis lagrangiens d'ordre 1

Sur chaque domaine élémentaire $[x_k, x_{k+1}]$, cherchons une fonction $v(x)$ sous la forme $v(x) = ax + b$. La fonction devant satisfaire

$$\begin{cases} ax_k + b = w_{k+1} \\ ax_{k+1} + b = w_k \end{cases}$$

Elle est de la forme, pour tout $x \in [x_k, x_{k+1}]$,

$$v(x) = \frac{x - x_k}{x_{k+1} - x_k}w_k + \frac{x_{k+1} - x}{x_{k+1} - x_k}w_{k+1}$$

Notons

$$\begin{cases} P_0(x) = x \\ P_1(x) = 1 - x \end{cases}$$

Si le maillage est uniforme de pas h

$$v(x) = w_k P_0(\frac{x - x_k}{h}) + w_{k+1}P_1(\frac{x - x_k}{h})$$

Sur la maille $[x_k, x_{k+1}]$, les fonctions de forme valent

$$N_i(x) = P_i(\frac{x - x_k}{h})$$

avec $k = 1, ..., m$ et $i = 0, 1$.

$$\int_{x_k}^{x_{k+1}} N_i(x)N_j(x)dx = h\int_0^1 P_i(x)P_j(x)dx$$

On calcule facilement la matrice de masse élémentaire

$$\widehat{M} = \frac{1}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

et la matrice de rigidité élémentaire

$$\widehat{K} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Conditions de Neumann. Supposons que les conditions limites vérifient

$$u'(a) = u'(b) = 0$$

Pour calculer la matrice de masse globale, remarquons que

$$\int_a^b N_i(x)N_j(x)dx = \sum_{k=1}^{m}\int_{x_k}^{x_{k+1}} N_i(x)N_j(x)dx = \sum_{k=1}^{m} h\int_0^1 P_i(x)P_j(x)dx$$

Définissons une expansion de la matrice élémentaire, en remplaçant dans une matrice de zéros, les coefficients des k-ième et $(k+1)$-ième lignes et colonnes par les composantes de la matrice élémentaire $\widehat{M}$

$$M^{(k)} = \frac{1}{6}\begin{pmatrix} 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 2 & 1 & \cdots & 0 \\ 0 & \cdots & 1 & 2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

La matrice de masse globale s'obtient en assemblant les matrices élémentaires

$$\begin{aligned} M &= h\sum_{k=0}^{m} M^{(k)} \\ &= \frac{h}{6}\begin{pmatrix} 2 & 1 & 0 & \cdots \\ 1 & 2 & 0 & \cdots \\ 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} + \frac{h}{6}\begin{pmatrix} 0 & 0 & 0 & \cdots \\ 0 & 2 & 1 & \cdots \\ 0 & 1 & 2 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} + \\ &\quad \cdots + \frac{h}{6}\begin{pmatrix} 0 & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & 2 & 1 \\ 0 & \cdots & 1 & 2 \end{pmatrix} \end{aligned}$$

La matrice de masse globale est une matrice carrée $(m \times m)$ de la forme "*matrice bande*"

$$M = \frac{h}{6}\begin{pmatrix} 2 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & 4 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 4 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & 2 \end{pmatrix}$$

De la même façon, le calcul de la matrice de rigidité s'effectue après expansion de la matrice de rigidité élémentaire. En posant

$$K^{(k)} = \begin{pmatrix} 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 1 & -1 & \cdots & 0 \\ 0 & \cdots & -1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

on calcule la matrice de rigidité globale

$$K = h\sum_{k=0}^{m} K^{(k)} = \frac{1}{h}\begin{pmatrix} 1 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix}$$

Condition de Dirichlet. On suppose que les conditions aux limites sont données par

$$u(a) = u(b) = 0$$

Dans ce cas, on peut choisir les fonctions u et v dans le même espace, c'est-à-dire prendre $v(a) = v(b) = 0$. L'espace d'approximation est alors un espace de dimension $(m-2)$ (et non plus de dimension m). La subdivision devient $x_2, ..., x_{m-1}$, car on supprime les bornes a et b. La matrice de masse est obtenue à partir de la matrice de masse précédente en supprimant les premières et dernières lignes et colonnes de façon à obtenir une matrice

$(m-2)\times(m-2)$

$$M=\frac{h}{6}\begin{pmatrix} 4 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & 4 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 4 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & 4 \end{pmatrix}$$

La matrice de rigidité est obtenue de la même manière en supprimant les premières et les dernières lignes et colonnes

$$K=\frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix}$$

12.6 Éléments finis lagrangiens d'ordre 2

Sur chaque sous-domaine élémentaire, on approche la fonction $v(x)$ par un polynôme de degré 2. Sur chaque intervalle $[x_k, x_{k+1}]$, on prend pour $v(x)$ le polynôme d'interpolation de Lagrange sur les trois valeurs v_k, $v_{k+1/2}$, et v_{k+1}. Dans ces conditions, on démontre que l'approximation s'écrit

$$v(x)=v_kN_0(x)+v_{k+1/2}N_{1/2}(x)+v_{k+1}N_{k+1}(x)$$

avec

$$N_i(x)=P_{2i}(\frac{x-x_k}{h})$$

et

$$\begin{cases} P_0(x)=2(x-1)(x-1/2) \\ P_1(x)=-4x(x-1) \\ P_2(x)=2x(x-1/2) \end{cases}$$

La matrice de masse élémentaire vaut

$$\widehat{M}=\frac{1}{30}\begin{pmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{pmatrix}$$

La matrice de rigidité élémentaire

$$\widehat{K} = \frac{1}{3}\begin{pmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{pmatrix}$$

Condition de Neumann. La matrice de masse globale vaut

$$M = \frac{h}{30}\begin{pmatrix} 8 & 2 & -1 & & & & & & \\ 2 & 16 & 2 & & & & & & \\ -1 & 2 & 8 & 2 & -1 & & & & \\ & & & 2 & 16 & 2 & & & \\ & & & -1 & 2 & 8 & & & \\ & & & & & & \ddots & & \\ & & & & & & 8 & 2 & -1 \\ & & & & & & 2 & 16 & 2 \\ & & & & & & -1 & 2 & 8 \end{pmatrix}$$

La matrice de rigidité globale

$$K = \frac{1}{3h}\begin{pmatrix} 14 & -8 & 1 & & & & & & \\ -8 & 16 & -8 & & & & & & \\ 1 & -8 & 14 & -8 & 1 & & & & \\ & & -8 & 16 & -8 & & & & \\ & & 1 & -8 & 14 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 14 & -8 & 1 \\ & & & & & & -8 & 16 & -8 \\ & & & & & & 1 & -8 & 14 \end{pmatrix}$$

Condition de Dirichlet. Les matrices de masse et de rigidité s'obtiennent en supprimant la première et la dernière ligne ainsi que la première et la dernière colonne des matrices obtenues dans le cas des conditions aux limites de Neumann.

12.7 Éléments finis lagrangiens d'ordre 3

Pour obtenir les éléments lagrangiens d'ordre 3, il suffit de chercher sur chaque sous-domaine élémentaire, une fonction test polynomiale de degré 3. Sur chaque maille, on peut écrire

$$v(x) = v_k N_0(x) + v_{k+1/3} N_{1/3}(x) + v_{k+2/3} N_{2/3}(x) + v_{k+1} N_1(x)$$

avec

$$N_i(x) = P_{3i}(\frac{x - x_k}{h})$$

Les polynômes de base sont donnés par

$$\left\{\begin{array}{l} P_0(x) = -3(x-1/3)(x-2/3)(x-1) \\ P_1(x) = 27x(/2)(x-2/3)(x-1) \\ P_2(x) = P_1(1-x) \\ P_3(x) = P_2(1-x) \end{array}\right.$$

12.8 Éléments finis hermitiens

Au lieu d'utiliser l'approximation de Lagrange, nous utilisons ici l'interpolation d'Hermite en imposant à la fonction $v(x)$ et à sa dérivée d'être continues sur tout le domaine. Pour une approximation cubique, nous chercherons donc une fonction polynomiale d'ordre 3 de la forme $a_3x^3 + a_2x^2 + a_1x + a_0$ vérifiant

$$\left\{\begin{array}{l} v(x_k) = v_k \\ v'(x_k) = v'_k \\ v(x_{k+1}) = v_{k+1} \\ v'(x_{k+1}) = v'_{k+1} \end{array}\right.$$

où les nombres $v_k, v'_k, v_{k+1}, v'_{k+1}$ sont des nombres quelconques. Une telle fonction est unique et s'écrit

$$v(x) = v_kN_0(x) + v'_kN_1(x) + v_{k+1}N_2(x) + v'_{k+1}N_3(x)$$

avec

$$\left\{\begin{array}{ll} N_i(x) = P_i(\frac{x-x_k}{h}) & \text{si } i \text{ est pair} \\ N_i(x) = hP_i(\frac{x-x_k}{h}) & \text{si } i \text{ est impair} \end{array}\right.$$

Les polynômes de base étant donnés par

$$\left\{\begin{array}{l} P_0(x) = (x-1)^2(2x+1) \\ P_1(x) = x(x-1)^2 \\ P_2(x) = P_0(1-x) \\ P_3(x) = -P_1(1-x) \end{array}\right.$$

On calcule facilement la matrice de masse élémentaire

$$\widehat{M} = \frac{1}{420}\begin{pmatrix} 156 & 32 & 54 & -13 \\ 32 & 4 & 13 & -3 \\ 54 & 13 & 156 & -32 \\ -13 & -3 & -32 & 4 \end{pmatrix}$$

puis, la matrice de rigidité élémentaire

$$\widehat{K} = \frac{1}{30}\begin{pmatrix} 36 & 3 & -36 & 3 \\ 3 & 4 & -3 & -1 \\ -36 & -3 & 36 & -3 \\ 3 & -1 & -3 & 4 \end{pmatrix}$$

L'expansion de la matrice de masse s'écrit

$$\begin{aligned} M_{i,j}^{(k)} &= \int_{x_k}^{x_{k+1}} N_i(x)N_j(x)dx = h\int_0^1 h^{\varepsilon_i+\varepsilon_j} P_i(x)P_j(x)dx \\ &= h^{1+\varepsilon_i+\varepsilon_j}\widehat{M}_{i,j} \end{aligned}$$

avec $\varepsilon_i = 0$ si i est pair et $\varepsilon_i = 1$ si i est impair.

$$M_{i,j}^{(k)} = \frac{h}{420}\begin{pmatrix} 156 & 32h & 54 & -13h \\ 32h & 4h^2 & 13h & -3h^2 \\ 54 & 13h & 156 & -32h \\ -13h & -3h^2 & -32h & 4h^2 \end{pmatrix}$$

En posant

$$M_{i,j}^{(k)} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$$

avec

$$M_{11} = \begin{pmatrix} 156 & 32h \\ 32h & 4h^2 \end{pmatrix} \qquad M_{12} = \begin{pmatrix} 54 & -13h \\ 13h & -3h^2 \end{pmatrix}$$

et

$$M_{21} = \begin{pmatrix} 54 & 13h \\ -13h & -3h^2 \end{pmatrix} \qquad M_{22} = \begin{pmatrix} 156 & -32h \\ -32h & 4h^2 \end{pmatrix}$$

et

$$A_{11} = M_{11} + M_{22} = \begin{pmatrix} 312 & 0 \\ 0 & 8h^2 \end{pmatrix}$$

on écrit alors la matrice de rigidité globale sous la forme d'une matrice bande, carrée de dimension $2m \times 2m$.

$$M = \frac{h}{420}\begin{pmatrix} M_{11} & M_{12} & & & \\ M_{21} & A_{11} & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & A_{11} & M_{12} \\ & & & M_{21} & M_{22} \end{pmatrix}$$

correspondant au vecteur $u = (u_1, u_1', ..., u_m, u_m')$. Pour le problème de Dirichlet, la matrice de masse est une matrice $2(m-2)\times 2(m-2)$, obtenue à partir de la matrice précédente en supprimant les lignes correspondant aux coordonnées u_1 et u_m, c'est-à-dire en supprimant les premières et les avant-dernières lignes et colonnes. De la même façon, on calcule une expansion de la matrice de rigidité

$$K_{i,j}^{(k)} = \frac{1}{30h}\begin{pmatrix} 36 & 3h & -36 & 3h \\ 3h & 4h^2 & -3h & -h^2 \\ -36 & -3h & 36 & -3h \\ 3h & -h^2 & -3h & 4h^2 \end{pmatrix}$$

En posant

$$K_{i,j}^{(k)} = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$$

avec

$$K_{11} = \begin{pmatrix} 36 & 3h \\ 3h & 4h^2 \end{pmatrix} \qquad K_{12} = \begin{pmatrix} -36 & 3h \\ -3h & -h^2 \end{pmatrix}$$

et

$$K_{21} = \begin{pmatrix} 36 & -3h \\ 3h & -h^2 \end{pmatrix} \qquad K_{22} = \begin{pmatrix} 36 & -3h \\ -3h & 4h^2 \end{pmatrix}$$

et

$$B_{11} = K_{11} + K_{22} = \begin{pmatrix} 72 & 0 \\ 0 & 8h^2 \end{pmatrix}$$

on écrit alors la matrice de rigidité globale sous la forme d'une matrice bande, carrée de dimension $2m \times 2m$.

$$K = \frac{1}{30h} \begin{pmatrix} K_{11} & K_{12} & & & \\ K_{21} & B_{11} & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & B_{11} & K_{12} \\ & & & K_{21} & K_{22} \end{pmatrix}$$

correspondant au vecteur $u = (u_1, u_1', ..., u_m, u_m')$. Pour le problème de Dirichlet, la matrice de rigidité K est une matrice $2(m-2) \times 2(m-2)$, obtenue à partir de la matrice précédente en supprimant les lignes correspondant aux coordonnées u_1 et u_m, c'est-à-dire en supprimant les premières et les avant-dernières lignes et colonnes.

12.9 Méthodes des résidus pondérés

Dans la *méthode des résidus pondérés*, le résidu

$$R(v) = \mathcal{L}(v) - f$$

est évalué à partir d'une fonction d'essai

$$u(x) = \sum_{j=1}^{m} u_j N_j(x)$$

Les fonctions N_i sont connues et la fonction $u(x)$ vérifie les conditions aux limites. La forme de cette fonction est déterminée à partir d'une fonction

de test $v \in V$ telle que le poids moyen du résidu soit nul, c'est-à-dire telle que

$$\int_\Omega R(u(x)).v(x)\ dx = 0$$

On détermine ainsi les valeurs des coefficients u_i, si on se fixe un ensemble de fonctions tests. Selon le choix de ces fonctions, on distingue plusieurs méthodes de résidus pondérés.

Dans la *méthode de collocation par sous-domaines*, l'espace Ω est divisé en m sous-domaines deux à deux disjoints Ω_j. Sur chaque sous-domaine, la fonction test est choisie comme la fonction indicatrice de ce sous-domaine

$$v_j = 1_{\Omega_j}$$

fonction valant 1 si $x \in \Omega_j$ et 0 sinon. La nullité du résidu conduit à résoudre l'équation

$$\int_{\Omega_j} (\mathcal{L}v - f)dx$$

Dans la *méthode de collocation par points*, les m sous-domaines se réduisent à m points. Les fonctions tests sont de la forme

$$v_j(x) = \delta(x - x_j)$$

La difficulté est alors de choisir ces points de sorte qu'ils respectent les symétries du problème.

Exemple. Considérons l'équation

$$\begin{cases} u''(x) - u = -x & \text{sur } [0,1] \\ u(0) = u(1) = 0 \end{cases}$$

Si on choisit un seul point ($m = 1$) et la fonction de forme $N_1(x) = x(1-x)$, les solutions approchées sont de la forme

$$u(x) = u_1 N_1(x) = u_1 x(1 - x)$$

Le coefficient u_1 est inconnu. La fonction $u(x)$ vérifie les conditions aux limites $u(0) = u(1) = 0$. Le résidu est

$$R(x) = u''(x) - u(x) + x = x + u_1(x^2 - x - 2)$$

Si on choisit le point $x_1 = 1/2$, la fonction de forme vaut

$$v_1(x) = \delta(x - 1/2)$$

La minimalisation de l'intégrale

$$K = \int_0^1 R(x)\ v_1(x)\ dx = \frac{-9}{8}u_1 + \frac{1}{2} = 0$$

conduit à la valeur $u_1 = 2/9$. La solution du problème est donc approchée par la fonction

$$u(x) = 2x(1-x)/9$$

Si on choisit deux points, il faudra considérer des fonctions

$$u(x) = u_1 N_1(x) + u_2 N_2(x)$$

avec par exemple comme fonctions de forme $N_1(x) = x(1-x)$ et $N_2(x) = x^2(1-x)$. La fonction $u(x)$ vérifie les conditions initiales. En choisissant deux points, par exemple, $x_1 = 1/3$ et $x_2 = 2/3$, les conditions $R(1/3) = R(2/3)$ conduisent à déterminer les deux valeurs u_1 et u_2.

La *méthode des moindres carrés* est une méthode dans laquelle on cherche à minimiser la moyenne quadratique

$$I(v) = \int_\Omega (\mathcal{L}v - f)^2 \, dx = \int_\Omega R(v(x))^2 \, dx$$

Elle consiste à prendre comme fonctions tests

$$v_j = \frac{\partial R}{\partial u_j}$$

En notant $\mathcal{L}^*$ l'opérateur adjoint et en développant,

$$I(v) = \int_\Omega (\mathcal{L}^*\mathcal{L}v) \; v \; dx - 2\int_\Omega (\mathcal{L}^* f) \; v \; dx + \int_\Omega f^2 \; dx$$

La minimisation conduit à résoudre le système

$$\mathcal{L}^*\mathcal{L}v = \mathcal{L}^* f$$

Exemple. Considérons le même problème que précédemment

$$\begin{cases} u''(x) - u = -x & \text{sur } [0,1] \\ u(0) = u(1) = 0 \end{cases}$$

Si on choisit un seul point ($m = 1$) et la fonction de forme $N_1(x) = x(1-x)$, les solutions approchées s'écrivent

$$u(x) = u_1 N_1(x) = u_1 x(1-x)$$

Le coefficient u_1 est inconnu. La fonction $u(x)$ vérifie les conditions aux limites $u(0) = u(1) = 0$. Le résidu vaut

$$R(x) = u''(x) - u(x) + x = x + u_1(x^2 - x - 2)$$

Les fonctions de tests sont les dérivées partielles du résidu relativement aux coefficients u_i

$$v_1(x) = \frac{\partial R(x)}{\partial u_1} = x^2 - x - 2$$

Par conséquent, le calcul de l'intégrale

$$K(v_1) = \int_0^1 R(x).v_1(x)\ dx = 0$$

conduit à l'équation

$$-\frac{13}{12} + u_1\frac{47}{10} = 0$$

qui donne la valeur $u_1 = 65/282$. La solution approchée est donc la fonction

$$u(x) = 65x(1-x)/282$$

La *méthode de Galerkin* dans laquelle les fonctions

$$v_j(x) = \frac{\partial u(x)}{\partial u_j}$$

sont constituées par l'ensemble des variations des fonctions de u. Les fonctions de pondération ou de tests sont égales aux fonctions de forme. Dans certaines conditions, la méthode de Galerkin équivaut à minimiser la fonctionnelle d'énergie et devient une méthode variationnelle.

Exemple. Considérons le même problème que précédemment

$$\begin{cases} u''(x) - u = -x & \text{sur } [0,1] \\ u(0) = u(1) = 0 \end{cases}$$

La fonction de test est égale à

$$v_1(x) = \frac{\partial u(x)}{\partial u_1} = N_1(x) = x(x-1)$$

En calculant l'intégrale

$$K(v_1) = \int_0^1 R(x).v_1(x)\ dx = 0$$

on obtient l'équation

$$-\frac{1}{12} + u_1\frac{11}{30} = 0$$

qui admet comme solution $u_1 = 30/132$. La solution approchée est par conséquent

$$u(x) = 30x(1-x)/132$$

12.10 Méthode de Rayleigh-Ritz

Parmi les méthodes variationnelles, la *méthode de Rayleigh-Ritz* est la plus courante. Dans cette méthode, on considère l'expression

$$J(v) = \frac{1}{2}a(v,v) - \int_\Omega fvdx$$

et sa forme intégrée qui s'écrit

$$K(u) = \int_\Omega J(x,u,u',...,u^{(n)})dx$$

où u' désigne la dérivée en x. La fonction J vérifie l'équation d'Euler

$$\frac{\partial J}{\partial u} - \frac{d}{dx}\left(\frac{\partial J}{\partial u'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial J}{\partial u"}\right) + ... + (-1)^{n-1}\frac{d^{n-1}}{dx^{n-1}}\left(\frac{\partial J}{\partial u^{(n)}}\right) = 0$$

En choisissant un sous-espace de V de dimension m, et une base de fonctions $N_1, N_2, ..., N_m$, la fonction u s'écrit

$$u(x) = \sum_{j=1}^{m} u_j N_j(x)$$

où les quantités u_j sont inconnues. La méthode de Ritz consiste à déterminer les quantités $(u_1, u_2, ..., u_m)$ de sorte que $K(u_1, ..., u_m)$ soit extrémale, ce qui impose les m conditions suivantes

$$\frac{\partial K(u_1,...,u_m)}{\partial u_j} = 0$$

Exemple. Considérons le système

$$\begin{cases} u''(x) = -x^2 & \text{sur } [0,1] \\ u(0) = u(1) = 0 \end{cases}$$

Le calcul des variations conduit à la fonctionnelle

$$J(x,u,u') = \frac{1}{2}(u')^2 - x^2 u$$

L'équation d'Euler traduit l'équation de départ. La méthode de Rayleigh-Ritz cherche à minimiser l'intégrale

$$\int_0^1 \frac{\partial}{\partial u_j}\left(\frac{1}{2}(u')^2 - f(x)u\right)dx = 0$$

En décomposant u sur une base de fonctions de forme, le problème revient à résoudre

$$\sum_{j=1}^{m}\left(\int_0^1 N_j' N_j' dx.u_j - \int_0^1 f(x)N_j dx\right) = 0$$

soit sous forme matricielle

$$Au = b$$

A est la matrice formée des coefficients

$$A_{i,j} = \int_0^1 N_j' N_j' dx$$

b le vecteur de composantes

$$b_j = \int_0^1 f(x) N_j dx$$

et u le vecteur $(u_1, u_2, ..., u_m)$. En particulier, si on choisit $m = 2$, et les fonctions de forme $N_1(x) = x(1-x)$ et $N_2(x) = x^2(1-x)$, la matrice A s'écrit

$$A = \begin{pmatrix} 1/3 & 1/6 \\ 1/6 & 2/15 \end{pmatrix}$$

et le vecteur $b = (1/20, 1/30)$. La résolution conduit aux valeurs $u_1 = 1/6$ et $u_2 = 1/15$. La valeur approchée est donc $u(x) = x(1-x)/6 + x^2(1-x)/15$. Cette solution approche la solution exacte $u(x) = -x(x^3-1)/12$.

12.11 Exercices

1. On considère l'équation de la chaleur sur un domaine triangulaire constitué de l'axe des x réduit à l'intervalle [0,1] et de la première bissectrice reliant l'origine au point de coordonnées (1,1). On suppose que la température est nulle ($u = 0$) sur l'axe des x, que le flux est nul sur la bissectrice ($\partial u/\partial n = 0$) et qu'il a une valeur constante ($\partial u/\partial n = 2$) sur le troisième côté du triangle, constitué du segment de droite reliant le point (1,0) au point (1,1). On discrétise ce domaine en quatre éléments triangulaires et six nœuds. Les nœuds ont les coordonnées suivantes : nœud 1 (0, 0), nœud 2 (1/2, 0), nœud 3 (1/2, 1/2), nœud 4 (1, 0), nœud 5 (1, 1/2) et nœud 6 (1, 1). Calculer les matrices élémentaires. Assembler le système. Résoudre l'équation aux points nodaux.

2. On considère l'équation

$$\begin{cases} u''(x) = x & \text{sur } [0,1] \\ u(0) = u(1) = 0 \end{cases}$$

et la fonction de forme $N_1(x) = x(1-x)$. Résoudre cette équation par la méthode de collocation de points en $x = 1/2$. Même question pour la méthode des moindres carrés et la méthode de Galerkin.

3. On considère l'équation

$$\begin{cases} x^2u''(x) - 2xu'(x) + 2u(x) = 0 \quad \text{sur } [1,4] \\ u(0) = 0 \quad u(1) = 12 \end{cases}$$

Écrire la résolution de ce système pour des polynômes quadratiques. Même question pour les méthodes de résidus pondérés.

4. On considère l'équation d'une tige rigide

$$\rho S \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(SE\frac{\partial u}{\partial x}\right)$$

de module d'élasticité E, de densité volumique ρ et de section S. On suppose que la tige est de longueur l, de section $2S$ sur la moitié de sa longueur et que les nœuds ont deux degrés de liberté. Écrire les équations discrétisées et calculer les matrices élémentaires.

5. On considère l'équation d'une poutre

$$\rho\frac{\partial^2 u}{\partial t^2} + \frac{\partial^2}{\partial x^2}\left(EI\frac{\partial^2 u}{\partial x^2}\right) = f(x,t)$$

dans laquelle $u(x,t)$ est le déplacement transverse de la poutre, ρ la densité volumique de la poutre, EI son module de rigidité et $f(x,t)$ son chargement. On suppose que la poutre est de longueur l. Écrire la discrétisation de l'équation d'Euler en éléments finis et calculer les matrices élémentaires pour des éléments hermitiens.

13
Équations de physique

Les équations présentées dans ce chapitre sont des équations spécifiques aux sciences physiques. Elles posent la plupart du temps de nombreux problèmes de stabilité associés à l'apparition de phénomènes physiques nouveaux comme les phénomènes d'interface, les changements de phase, la propagation de flammes, etc. C'est pourquoi ces équations aux multiples interprétations ont nécessité le développement de leurs propres méthodes de résolution.

13.1 Équation de Navier-Stokes

L'évolution temporelle d'un fluide compressible, visqueux de densité $\rho(x,t)$ et de vitesse $\mathbf{u}(x,t)$ dans une région Ω de l'espace tridimensionnel $\mathbb{R}^3$, $x \in \Omega$ et $t \in [0,T]$ est donnée par le système d'équations de Navier-Stokes, comprenant l'équation de la conservation de la masse, en notant $\partial_t(\rho) = \partial\rho/\partial t$

$$\partial_t(\rho) + div(\rho\mathbf{u}) = \mathbf{0}$$

et l'équation de Navier-Stokes proprement dite

$$\partial_t(\rho\mathbf{u}) + div(\rho\mathbf{u} \otimes \mathbf{u}) - \mu_1\Delta\mathbf{u} - (\mu_1 + \mu_2)\nabla div(u) + a\nabla p = \rho\mathbf{f} + \mathbf{g}$$

où p est la pression, μ_1 et μ_2 sont les coefficients de viscosité qui vérifient $\mu_1 > 0$ et $2\mu_1/3 + \mu_2 \geq 0$, a est une constante positive, f représente les forces externes agissant sur le fluide et g est l'attraction universelle. Lorsque

le fluide est un gaz, la pression s'exprime par $p = \rho^\gamma$ où γ est la constante adiabatique $\gamma > 1$. Sous les conditions aux limites et initiales,

$$\begin{cases} \mathbf{u} = \mathbf{0} & (x,t) \in \partial\Omega \times [0,T] \\ \rho(0,x) = \rho_0(x) \text{ et } (\rho\mathbf{u})(0,x) = \mathbf{q}_0(x) & x \in \Omega \end{cases}$$

P.-L. Lions a montré que lorsque Ω est une région bornée suffisamment régulière, et sous certaines hypothèses dont $\gamma \geq 9/5$, les équations de Navier-Stokes ont des solutions faibles. Pour un fluide incompressible, de densité constante, les équations de Navier-Stokes s'écrivent

$$\begin{cases} \partial_t(\mathbf{u}) - \nu\Delta\mathbf{u} + (\mathbf{u}.\nabla)\mathbf{u} + \nabla p = \mathbf{f} & (x,t) \in \Omega \times [0,T] \\ \nabla.\mathbf{u} = 0 & \\ \mathbf{u} = \mathbf{0} & x \in \partial\ \Omega \end{cases}$$

L'incompressibilité revient à négliger les influences de la pression et de la température sur la masse volumique. Ces équations sont importantes, car la majorité des problèmes rencontrés dans des domaines autres que l'aérodynamique concernent les écoulements de fluides incompressibles. Bien qu'il soit impossible de faire l'hypothèse d'incompressibilité dans des phénomènes comme les transitions de phase liquide-vapeur ou la cavitation, on ne considérera ici que la résolution numérique des équations de Navier-Stokes incompressibles. En prenant le rotationnel, on montre que les équations se mettent sous la forme

$$\partial_t(\mathbf{rot}\ \mathbf{u}) = \mathbf{rot}(\mathbf{u} \wedge \mathbf{rot}\ \mathbf{u}) + \nu\Delta\mathbf{rot}\ \mathbf{u}$$

Dans le cas d'un écoulement plan, la vorticité $\mathbf{rot}(\mathbf{u})$ n'a qu'une composante

$$\omega = \frac{\partial u_2}{\partial x} - \frac{\partial u_1}{\partial y}$$

où u_1, u_2 sont les coordonnées cartésiennes du vecteur $\mathbf{u}$. L'équation s'écrit alors

$$\frac{\partial\omega}{\partial t} + \frac{\partial(u_1\omega)}{\partial x} + \frac{\partial(u_2\omega)}{\partial y} = \nu\Delta\omega$$

et dans le cas monodimensionnel, cette équation se réduit à sa plus simple expression

$$\frac{\partial\omega}{\partial t} + u\frac{\partial\omega}{\partial x} - \nu\frac{\partial^2\omega}{\partial x^2} = 0$$

Lorsque l'utilisation d'un schéma aux différences centrées conduit à des instabilités ou à des oscillations, on raffine le maillage ou on emploie un schéma décentré vers l'amont ou "upwind" pour prendre en compte l'aspect directionnel de l'écoulement. Pour $u > 0$ et $T > 0$, on a par exemple, en notant h le pas en espace Δx et k le pas en temps Δt,

$$\omega_{i,n+1} = \omega_{i,n} - \frac{ku}{h}(\omega_{i,n} - \omega_{i-1,n}) + \frac{\nu k}{h^2}(\omega_{i+1,n} - 2\omega_{i,n} + \omega_{i-1,n})$$

En définissant le nombre de Reynolds de maille par

$$\mathcal{R}_h = \frac{uh}{\nu}$$

on montre que le décentrement est très favorable aux grands nombres de Reynolds de maille, et que le schéma est stable sous la condition

$$\frac{\nu k}{h^2} \leq \frac{1}{2+\mathcal{R}_h}$$

Une autre grande classe de méthodes de résolution employée pour les équations de Navier-Stokes regroupe les méthodes du type prédiction-correction. Pour un système bidimensionnel, si on note u et v les vitesses supposées constantes, l'équation d'advection-diffusion s'écrit

$$\frac{\partial \omega}{\partial t} + u\frac{\partial \omega}{\partial x} + v\frac{\partial \omega}{\partial x} - \nu\left(\frac{\partial^2 \omega}{\partial x^2} + \frac{\partial^2 \omega}{\partial y^2}\right) = 0$$

Dans la méthode des directions alternées, en notant les discrétisations de l'espace en indice et celles du temps en exposant, le prédicteur est du type

$$\begin{aligned}\omega_{i,j}^{n+1/2} &= \omega_{i,j}^{n} - \frac{ku}{4h}\left(\omega_{i+1,j}^{n+1/2} - \omega_{i-1,j}^{n+1/2}\right) - \frac{kv}{4h}\left(\omega_{i,j+1}^{n} - \omega_{i,j-1}^{n}\right) + \\ &\quad \frac{\nu k}{2h^2}\left(\omega_{i+1,j}^{n+1/2} - 2\omega_{i,j}^{n+1/2} + \omega_{i-1,j}^{n+1/2} + \omega_{i,j+1}^{n} - 2\omega_{i,j}^{n} + \omega_{i,j-1}^{n}\right)\end{aligned}$$

et le correcteur s'écrit

$$\begin{aligned}\omega_{i,j}^{n+1/2} &= \omega_{i,j}^{n+1/2} - \frac{ku}{4h}\left(\omega_{i+1,j}^{n+1/2} - \omega_{i-1,j}^{n+1/2}\right) - \frac{kv}{4h}\left(\omega_{i,j+1}^{n+1} - \omega_{i,j-1}^{n+1}\right) + \\ &\quad \frac{\nu k}{2h^2}\left(\omega_{i+1,j}^{n+1/2} - 2\omega_{i,j}^{n+1/2} + \omega_{i-1,j}^{n+1/2} + \omega_{i,j+1}^{n+1} - 2\omega_{i,j}^{n+1} + \omega_{i,j-1}^{n+1}\right)\end{aligned}$$

On démontre que ce schéma est inconditionnellement stable, mais il nécessite un maillage fin. En effet, pour que la méthode soit réellement intéressante, il faut que la factorisation matricielle se fasse simplement et donc que les matrices soient à diagonale dominante pour conduire à une forme tridiagonale. On démontre que les sous-systèmes de prédiction-correction sont à diagonale dominante si le nombre de Reynolds par maille est plus petit que 2, donc que le maillage est fin.

En 1968, Chorin et Teman ont proposé indépendamment une méthode de projection, qui consiste à écrire pour l'équation des fluides incompressibles

$$\partial_t(\mathbf{u}) - \nu\Delta\mathbf{u} + (\mathbf{u}.\nabla)\mathbf{u} + \nabla p = \mathbf{f}$$

la décomposition de Hodge

$$\begin{aligned}\partial_t(\mathbf{u}) &= P(\nu\Delta\mathbf{u} - (\mathbf{u}.\nabla)\mathbf{u} + \mathbf{f}) \\ \nabla p &= (1-P)(\nu\Delta\mathbf{u} - (\mathbf{u}.\nabla)\mathbf{u} + \mathbf{f})\end{aligned}$$

où P est l'opérateur de projection. La discrétisation temporelle employée est par exemple

$$\begin{array}{l} \frac{\mathbf{u}^{n+1/2}-\mathbf{u}^n}{\Delta t} - \nu\Delta\left(\frac{\mathbf{u}^{n+1/2}+\mathbf{u}^n}{2}\right) + b(\mathbf{u}^n,\mathbf{u}^{n-1}) = \frac{\mathbf{f}^{n+1}-\mathbf{f}^n}{2} \\ -\nabla p^{n+1} = \frac{\mathbf{u}^{n+1}-\mathbf{u}^{n+1/2}}{\Delta t} \end{array}$$

avec

$$b(\mathbf{u}^n,\mathbf{u}^{n-1}) = \frac{3}{2}(\mathbf{u}^n.\nabla)\mathbf{u}^n - \frac{1}{2}(\mathbf{u}^{n-1}.\nabla)\mathbf{u}^{n-1}$$

Ce schéma est complété par une discrétisation des conditions initiales et des conditions aux limites. Des améliorations de cette méthode ont été proposées en 1986 par Van Kan.

13.2 Équation de Schrödinger

Dans sa forme la plus générale, l'équation de Schrödinger s'écrit

$$-i\hbar\partial_t\Psi = H\Psi$$

où H est le hamiltonien du système, Ψ est une fonction de carré intégrable appelée *fonction d'onde* et $\hbar = h/2\pi$, où h est la constante de Planck. En supposant que la fonction d'onde soit le produit d'une fonction temporelle et d'une fonction d'espace, on démontre que la résolution de l'équation de Schrödinger se ramène à l'équation aux valeurs propres

$$H\Psi = E\Psi$$

Dans la théorie quantique, on définit les composantes du moment cinétique orbital $\mathbf{L} = (L_x, L_y, L_z)$ par les opérateurs

$$\left\{ \begin{array}{l} L_x = -i\hbar\left(y\frac{\partial}{\partial z} - z\frac{\partial}{\partial y}\right) \\ L_y = -i\hbar\left(z\frac{\partial}{\partial x} - x\frac{\partial}{\partial z}\right) \\ L_z = -i\hbar\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right) \end{array} \right.$$

H est un opérateur hermitien qui commute avec les composantes du moment cinétique orbital et de son carré

$$[H, L_j] = HL_j - L_jH = 0 \text{ et } [H, L^2] = 0$$

Pour un système monoatomique de masse m et de potentiel V, l'équation de Schrödinger se simplifie en

$$\left[\frac{-\hbar^2}{2m}\Delta + V\right]\Psi = E\Psi$$

En coordonnées polaires (r,θ,φ), les composantes du moment cinétique sont indépendantes de r

$$\left\{\begin{array}{l} L_x = -i\hbar\left(\sin\varphi\frac{\partial}{\partial\theta} - \frac{\cos\varphi}{\tan\theta}\frac{\partial}{\partial\varphi}\right) \\ L_y = -i\hbar\left(-\cos\varphi\frac{\partial}{\partial\theta} + \frac{\sin\varphi}{\tan\theta}\frac{\partial}{\partial\varphi}\right) \\ L_z = -i\hbar\frac{\partial}{\partial\varphi} \end{array}\right.$$

Le laplacien s'exprime en fonction du carré du moment cinétique L^2

$$\Delta\Psi = \frac{1}{r}\frac{\partial^2}{\partial r^2}(r\Psi) - \frac{1}{r^2\hbar^2}L^2$$

où L^2 est donné par

$$L^2 = -\hbar^2\left(\frac{\partial^2}{\partial\theta^2} + \frac{1}{\tan\theta}\frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}\right)$$

Par conséquent, l'opérateur hamiltonien vaut

$$H\Psi = \frac{-\hbar^2}{2m}\frac{1}{r}\frac{\partial^2}{\partial r^2}(r\Psi) + \frac{1}{2mr^2}L^2\Psi + V(r)\Psi$$

Comme H, L^2, L_z commutent, on peut chercher $\Psi(r,\theta,\varphi)$ comme fonction propre de H, L^2 et de L_z. Les harmoniques sphériques $Y_\ell^m(\theta,\varphi)$ sont fonctions propres de L^2 et de L_z, avec pour valeurs propres $\ell(\ell+1)\hbar^2$ et $m\hbar$,

$$\left\{\begin{array}{l} L^2Y_\ell^m = \ell(\ell+1)\hbar^2Y_\ell^m \\ L_zY_\ell^m = m\hbar Y_\ell^m \end{array}\right.$$

On montre que les harmoniques sphériques sont données par la relation

$$Y_\ell^m(\theta,\varphi) = \varepsilon\sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-|m|)!}{(\ell+|m|)!}}P_\ell^{|m|}(\cos\theta)e^{im\varphi}$$

où $\varepsilon = (-1)^m$ si $m \geq 0$ et $\varepsilon = 1$ si $m < 0$. $P_\ell^m(u)$ est la fonction

$$P_\ell^m(u) = \sqrt{(1-u^2)^m}\frac{d^m}{du^m}P_\ell(u)$$

où $P_\ell(u)$ est le polynôme de Legendre défini par

$$P_\ell(u) = \frac{(-1)^\ell}{2^\ell\ell!}\frac{d^\ell}{du^\ell}(1-u^2)^\ell$$

Les fonctions d'onde $\Psi(r,\theta,\varphi) = u(r)Y_\ell^m(\theta,\varphi)$ sont solutions de l'équation de Schrödinger qui se réduit à l'équation radiale

$$\frac{-\hbar^2}{2m}\frac{1}{r}\frac{d^2}{dr^2}(ru) + \frac{\ell(\ell+1)\hbar^2}{2mr^2}u + V(r)u = Eu$$

Dans le cas général, pour un domaine de l'espace Ω quelconque, on ne connaît pas de solution analytique. On démontre que l'équation de Schrödinger qui s'écrit sous la forme mathématique

$$\begin{cases} i\partial_t u + \Delta u + f(u) = 0 & (x,t) \in \partial\Omega \times [0,T] \\ u(x,0) = u_0(x) & x \in \Omega \end{cases}$$

admet une solution unique sous des conditions de régularité de f, u_0 et Ω. De même, l'équation de Schrödinger non linéaire

$$\begin{cases} i\partial_t u + \Delta u + \alpha |u|^2 u = 0 & (x,t) \in \mathbb{R} \times [0,\infty[\\ u(x,0) = u_0(x) & x \in \mathbb{R} \end{cases}$$

admet une solution $u(x,t)$ complexe pour $\alpha \geq 0$. Dans le cas monodimensionnel, on démontre que le schéma

$$i\frac{u_{k,j+1}-u_{k,j}}{\Delta t} + \frac{1}{2\Delta x^2}((u_{k+1,j} - 2u_{k,j} + u_{k-1,j}) + (u_{k+1,j+1} \\ -2u_{k,j+1} + u_{k-1,j+1}) + \frac{\alpha}{4}\left(|u_{k,j}|^2 + |u_{k,j+1}|^2\right)(u_{k,j} + u_{k,j+1}) = 0$$

converge. On démontre aussi que si u est suffisamment régulière, les quantités

$$E(t) = \int_{-\infty}^{+\infty} |u(x,t)|^2 \, dx$$

et

$$F(t) = \int_{-\infty}^{+\infty} \left|\frac{\partial u(x,t)}{\partial x}\right|^2 dx - \frac{\alpha}{2}\int_{-\infty}^{+\infty} |u(x,t)|^4 \, dx$$

sont conservées $E(t) = E(0)$ et $F(t) = F(0)$.

13.3 Équation de Korteweg de Vries

L'équation de Korteweg de Vries (KdV)

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + a\frac{\partial^3 u}{\partial x^3} = 0$$

est équivalente à l'équation

$$\frac{\partial u}{\partial t} - 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0$$

par le changement de variable $X = a^{1/3}x$ et $U = -6a^{1/3}u$. Cette équation s'écrit de manière simplifiée

$$u_t - 6uu_x + u_{xxx} = 0$$

l'indice, éventuellement répété, indiquant la variable de dérivation. La solution de cette équation, appelée *soliton*, est une onde non dispersive, de la forme

$$u(x,t) = \beta sech^2(\frac{x-ct}{\ell}) = \beta sech^2\left(\sqrt{\frac{\beta}{2}}(x-2\beta t)\right)$$

avec $sech(x) = 1/\cosh(x)$, $c = 2\beta$ et $\ell = \sqrt{\beta/2}$. Si u est une solution de l'équation de KdV, alors une primitive v de u

$$v(x,t) = \int_{-\infty}^{x} u(x,t)dx$$

est solution de l'équation

$$v_t + 3v_x^2 + v_{xxx} = 0$$

Pour un paramètre a_1, l'équation de KdV admet une solution

$$u_1 = a_1 sech^2\left(\sqrt{\frac{a_1}{2}}(x-2a_1t)\right)$$

qui a pour intégrale

$$w_1 = \sqrt{2a_1}\tanh\left(\sqrt{\frac{a_1}{2}}(x-2a_1t)\right)$$

La transformation de Bäcklund de paramètre a_2

$$\begin{cases} w_x = a_2 - v_x - (v-w)^2/2 \\ w_t = -v_t + (v-w)(v_{xx} - w_{xx}) - 2(v_x + v_x w_x + w_x^2) \end{cases}$$

est complètement intégrable et si u est solution de l'équation de KdV, alors w est aussi solution de KdV. Ce système conduit à une nouvelle solution de l'équation de KdV

$$w_2 = \sqrt{2a_2}\coth\left(\sqrt{\frac{a_2}{2}}(x-2a_2t)\right)$$

de dérivée

$$u_2 = -a_2 csch\left(\sqrt{\frac{a_2}{2}}(x-2a_2t)\right)$$

Le principe de superposition non linéaire affirme que pour $a_2 > a_1$, la solution, appelée un 2-soliton, construite sur les solutions de KdV u_j, w_j, $j = 1, 2$ et donnée par l'équation

$$\begin{aligned} w_{12} &= w_0 + \frac{a_1 - a_2}{2}\frac{(u_1 - u_2)}{(w_1 - w_2)^2} \\ &= w_0 + (a_1 - a_2)\frac{a_1 sech^2 s_1 + a_2 csch^2 s_2}{(\sqrt{a_1}\tanh s_1 - \sqrt{a_2}\coth s_2)^2} \end{aligned}$$

où

$$s_i = \sqrt{\frac{a_i}{2}}(x - 2a_i t), \text{ pour } i = 1, 2$$

est encore solution de l'équation de KdV.

Les méthodes de traitement numérique de l'équation de KdV sont nombreuses. Pour le système périodique

$$\begin{cases} \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0 & x \in \mathbb{R},\ t > 0 \\ u(x+1,t) = u(x,t) & x \in \mathbb{R},\ t \geq 0 \\ u(x,0) = u_0(x) & x \in \mathbb{R} \end{cases}$$

on utilisera par exemple le schéma aux différences finies suivant

$$\frac{u_{i,j+1} - u_{i,j}}{\Delta t} + \frac{(u_{i+1,j} - u_{i-1,j})u_{i,j}}{\Delta x} + \frac{1}{2(\Delta x)^2}(u_{i+2,j} \\ -u_{i+1,j} - 2u_{i-1,j} - u_{i-2,j}) = 0$$

ou bien une méthode d'éléments finis, en utilisant la formulation variationnelle

$$\begin{cases} \int_0^1 \left(\frac{\partial u}{\partial t} v - \frac{1}{2}u^2 \frac{\partial v}{\partial x} + \frac{\partial u \partial^2 v}{\partial x \partial x^2} \right) dx = 0 \\ u(x,0) = u_0(x) \end{cases}$$

ou bien encore une méthode pseudo-spectrale. Cette dernière méthode consiste à approcher la solution par un développement de la forme

$$u^m(x,t) = \sum_{n=-m}^{m} u_n^m(t) e^{inx}$$

En prenant

$$u^m(x, t+k) = u^m(x,t) + k\frac{\partial u^m}{\partial t} + \frac{k^2}{2}\frac{\partial^2 u^m}{\partial t^2} + \frac{k^3}{6}\frac{\partial^3 u^m}{\partial t^3}$$

et en écrivant

$$\begin{aligned} \partial_t u^m &= -u^m \partial_x u^m - \partial_{xxx} u^m \\ \partial_{tt} u^m &= -\partial_t u^m \partial_x u^m - u^m \partial_x (\partial_t u^m) - \partial_{xxx} (\partial_t u^m) \\ \partial_{ttt} u^m &= \partial_{tt} u^m \partial_x u^m - 2\partial_t u^m - \partial_x (\partial_t u^m) - u^m \partial_x (\partial_{tt} u^m) - \partial_{xxx} (\partial_{tt} u^m) \end{aligned}$$

on calculera toutes les dérivées en utilisant

$$\partial_k u^m(x,t) = \sum_{n=-m}^{m} (in)^k u_n^m(t) e^{inx}$$

Notons enfin que l'équation de Korteweg de Vries se généralise à deux dimensions : c'est l'équation de *Kadomtsev-Petviashvili*

$$(u_t - 6uu_x + u_{xxx})_x + 3a^2 u_{yy} = 0$$

13.4 Équation de sine-Gordon

L'équation de sine-Gordon est de la forme

$$u_{xt} = \sin u$$

La transformation de Bäcklund de paramètre $a \neq 0$ est définie par

$$\begin{cases} v_x = u_x - 2a \sin \frac{1}{2}(u+v) \\ v_t = -u_t + \frac{2}{a} \sin \frac{1}{2}(u-v) \end{cases}$$

En dérivant la première équation en t et la seconde en x, on vérifie que u est solution de l'équation de sine-Gordon si et seulement si v est solution de la même équation. Par conséquent, la transformation de Bäcklund relie deux solutions. De plus, si on considère la solution triviale $v = 0$, le système d'équations de Bäcklund s'écrit

$$\begin{cases} u_x = 2a \sin(u/2) \\ u_t = \frac{2}{a} \sin(u/2) \end{cases}$$

La résolution de ces équations conduit à une nouvelle solution de l'équation de sine-Gordon, appelée *1-soliton*,

$$u(x,t) = 4 \arctan(\alpha \exp(ax + \frac{t}{a}))$$

où α est une constante d'intégration. Le principe de superposition non linéaire permet d'obtenir d'autres solutions. Soit u une solution de l'équation de sine-Gordon. La transformation de Bäcklund pour un paramètre a_1 conduit à une nouvelle solution u_1. Comme u_1 est solution de l'équation de sine-Gordon, on peut construire une transformation de Bäcklund relative à u_1 de paramètre a_2, qui conduit à une nouvelle solution u_{12}. De la même façon, on peut commencer par construire une transformation de Bäcklund relative à u de paramètre a_2, on obtient alors une solution u_2. Puis, on construit une nouvelle transformation de Bäcklund relativement à u_2 de paramètre a_1 qui conduit à une nouvelle solution u_{21}. Le principe de superposition non linéaire affirme que $u_{12} = u_{21}$ et que u_{12} est donné par

$$u_{12} = 4 \arctan \left[\frac{a_1 + a_2}{a_1 - a_2} \frac{\sinh((v_1 - v_2)/2)}{\cosh((v_1 + v_2)/2)} \right] + u_0$$

avec pour $i = 1, 2$

$$v_i = a_i x + t/a_i$$

Cette solution est appelée un 2-soliton. En itérant la procédure, on construit ainsi des n-solitons.

Le système périodique

$$\begin{cases} u_{tt} - u_{xx} = -sin(u) & x \in \mathbb{R},\ t > 0 \\ u(x+1,t) = u(x,t) & x \in \mathbb{R},\ t \geq 0 \\ u(x,0) = u_0(x) & x \in \mathbb{R} \\ u_t(x,0) = u_1(x) & x \in \mathbb{R} \end{cases}$$

conserve la quantité $E(t) = E(0)$

$$E(t) = \int_0^1 \left(\left(\frac{\partial u}{\partial t} \right)^2 + \left(\frac{\partial u}{\partial x} \right)^2 - 2cos(u) \right) dx$$

Pour la résolution numérique, on emploie des schémas qui conservent cette quantité, comme par exemple le schéma

$$\frac{1}{k^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) - \frac{1}{h^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) = s(u_{i,j})$$

avec $k = \Delta t$ et $h = \Delta x$ et

$$s(u_{i,j}) = \frac{cos(u_{i,j+1}) - cos(u_{i,j-1})}{u_{i,j+1} - u_{i,j-1}}$$

13.5 Équation de Klein-Gordon

L'équation de Klein-Gordon est définie par

$$u_{tt} - u_{xx} - au + bu\,|u|^{\alpha} = 0$$

où $a > 0$, $b \geq 0$, et $\alpha > 0$. Le système périodique

$$\begin{cases} u_{tt} - u_{xx} - au + bu\,|u|^{\alpha} = 0 & x \in \mathbb{R},\ t > 0 \\ u(x+1,t) = u(x,t) & x \in \mathbb{R},\ t \geq 0 \\ u(x,0) = u_0(x) & x \in \mathbb{R} \\ u_t(x,0) = u_1(x) & x \in \mathbb{R} \end{cases}$$

conserve la quantité $E(t) = E(0)$,

$$E(t) = \int_0^1 \left(\left(\frac{\partial u}{\partial t} \right)^2 + \left(\frac{\partial u}{\partial x} \right)^2 - au + \frac{2b}{\alpha+2}\,|u|^{\alpha+2} \right) dx$$

Pour la résolution numérique, on emploie des schémas qui conservent cette quantité. Par exemple, pour $\alpha = 2$, en posant $k = \Delta t$ et $h = \Delta x$, on prendra

$$\begin{array}{l} \frac{1}{k^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) - \frac{1}{h^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) \\ -\frac{a}{2}(u_{i,j+1} + u_{i,j-1}) + s(u_{i,j}) = 0 \end{array}$$

avec

$$s(u_{i,j}) = \frac{b}{4}(u_{i,j+1}^3 + u_{i,j-1}u_{i,j+1}^2 + u_{i,j-1}^2 u_{i,j+1} + u_{i,j-1}^3)$$

13.6 Équation de Benjamin-Bona-Mahony

L'équation de Benjamin-Ono est l'équation

$$u_t + uu_x + H(u_{xx}) = 0$$

où H est la transformée de Hilbert, définie par

$$Hf(x,t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{f(u,t)}{u-x} du$$

et P est la partie principale de Cauchy.
L'équation de Benjamin-Bona-Mahony s'écrit

$$u_t + u_x + uu_x - u_{xxt} = 0$$

Pour le traitement numérique de cette équation, on utilisera par exemple le schéma

$$\frac{1}{k}(u_{i,j+1} - u_{i,j}) - \frac{1}{2h} u_{i,j}(u_{i+1,j} - u_{i-1,j} + u_{i+1,j+1} - u_{i-1,j+1}) + \frac{1}{kh^2}(u_{i+1,j+1} - 2u_{i,j} + u_{i-1,j+1} - u_{i+1,j} + 2u_{i,j} - u_{i-1,j}) = 0$$

13.7 Exercices

1. En utilisant la transformation de Bäcklund

$$\begin{cases} v_x = -\frac{1}{2c} uv \\ v_t = \frac{v}{4c}(u^2 - 2u_x v) \end{cases}$$

montrer que u et v vérifient l'équation de Burgers

$$u_t + uu_x = cu_{xx}$$

et l'équation de la diffusion

$$v_t = cu_{xx}$$

2. En utilisant la transformation de Bäcklund

$$\begin{cases} v_x = u - v^2 \\ v_t = -u_{xx} + 2(uv_x + u_x v) \end{cases}$$

montrer que u est solution de l'équation de KdV

$$u_t - 6uu_x + u_{xxx} = 0$$

et v est solution de l'équation de KdV modifiée (mKdV)

$$v_t - 6v^2 v_x + v_{xxx} = 0$$

3. Montrer que l'équation de Boussinesq

$$u_{tt} - u_{xx} + (3u^2)_{xx} - u_{xxxx} = 0$$

admet une solution de la forme

$$u(x,t) = asech^2(b(x-ct)+d)$$

Préciser les relations liant les constantes a, b, c et d et vérifier que l'onde se propage dans les deux sens.

4. Montrer que l'équation de KdV

$$2u_t + 3uu_x + u_{xxx}/3 = 0$$

conserve les fonctions

$$M(t) = \int_{-\infty}^{\infty} udx$$

$$E(t) = \int_{-\infty}^{\infty} (u^3 - \frac{1}{3}u_x^2)dx$$

et

$$J(t) = \int_{-\infty}^{\infty} u^2 dx$$

5. Vérifier que l'équation de Kadomtsev-Petviashvili

$$(u_t - 6uu_x + u_{xxx})_x + 3u_{yy} = 0$$

admet la solution

$$u(x,y,t) = \frac{-1}{2}k^2 sech^2(\frac{1}{2}(kx + \ell y - \omega t))$$

où $\omega = k^3 + 3\ell^2/k$.

6. Vérifier que l'équation de Schrödinger non linéaire (SNL)

$$iu_t + u_{xx} + |u|^2 u = 0$$

admet une solution de la forme

$$u(x,t) = \left[(1-4(1+2it))/(1+2x^2+4t^2)\right] e^{it}$$

Vérifier que l'expression $v(x,t)$ suivante est aussi solution de SNL

$$v(x,t) = a\left[1 + 2m(mcos\theta + insin\theta)/w(x,t)\right] e^{ia^2 t}$$

où a et m sont des réels quelconques, $n^2 = (1+m^2)$, $\theta = 2nma^2 t$ et

$$w(x,t) = ncosh(ma\sqrt{2}x) + cos\theta$$

Annexe A
Polynômes orthogonaux

A.1 Polynômes de Legendre

On appelle *polynôme de Legendre* de degré n, le polynôme défini par la relation de récurrence

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x)$$

initialisée par

$$P_0(x) = 1 \quad P_1(x) = x$$

Les premiers polynômes sont

$$\begin{aligned}
P_0(x) &= 1 \\
P_1(x) &= x \\
P_2(x) &= (3x^2 - 1)/2 \\
P_3(x) &= (5x^3 - 3x)/2 \\
P_4(x) &= (35x^4 - 30x^2 + 3)/8 \\
P_5(x) &= (63x^5 - 70x^3 + 15x)/8 \\
P_6(x) &= (231x^6 - 315x^4 + 105x^2 - 5)/16
\end{aligned}$$

(1) Les polynômes de Legendre sont solutions de l'équation différentielle

$$(1-x^2)y'' - 2xy' + n(n+1)y = 0$$

(2) Les polynômes de Legendre satisfont la relation de récurrence

$$(1-x^2)P_n'(x) = -nxP_n(x) + nP_{n-1}(x)$$

(3) Formule de Rodrigues

$$P_n(x) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n}(1-x^2)^n$$

(4) Majorations

$$\forall x \in [-1,+1]\,, \qquad |P_n(x)| \leq 1$$

$$\forall x \in [-1,+1]\,, \qquad |P_n'(x)| \leq \frac{n(n+1)}{2}$$

$$\forall x \in [-1,+1]\,, \qquad |P_n(x)| \leq \frac{1}{\sqrt{8\pi n(1-x^2)}}$$

$$\frac{1-P_n^2(x)}{(2n-1)(n+1)} \leq P_n^2(x) - P_{n-1}(x)P_{n+1}(x) \leq \frac{2n+1}{3n(n+1)}$$

(5) Les polynômes de Legendre sont des polynômes orthogonaux relativement à la fonction de poids $\omega(x) = 1$ sur l'intervalle $[-1,+1]$.

$$\int_{-1}^{+1} P_n(x)P_m(x)dx = \frac{2}{2n+1}\delta_{n,m}$$

En particulier, $P_n(1) = 1$ et

$$\|P_n\|_2 = \left(\int_{-1}^{+1} P_n^2(x)dx\right)^{1/2} = \sqrt{\frac{2}{2n+1}}$$

(6) Les polynômes de Legendre vérifient la formule

$$\int_{-1}^{+1} P_n(x)\frac{dx}{\sqrt{1-x}} = \frac{2^{3/2}}{2n+1}$$

A.2 Polynômes de Laguerre

On appelle *polynôme de Laguerre* d'ordre n, le polynôme défini par la relation de récurrence

$$(n+1)L_{n+1}(x) = (2n+1-x)L_n(x) - nL_{n-1}(x)$$

et les conditions d'initialisation

$$L_0(x) = 1 \quad L_1(x) = 1-x$$

Pour $\alpha > -1$. On appelle *polynôme de Laguerre généralisé* d'ordre n et on note $L_n^{(\alpha)}(x)$ le polynôme défini par la relation de récurrence

$$(n+1)L_{n+1}^{(\alpha)}(x) = (2n+\alpha+1-x)L_n^{(\alpha)}(x) - (n+\alpha)L_{n-1}^{(\alpha)}(x)$$

initialisée par

$$L_0^{(\alpha)}(x) = 1 \quad L_1^{(\alpha)}(x) = 1 + \alpha - x$$

Le polynôme de Laguerre proprement dit correspond au cas $\alpha = 0$. Les premiers polynômes sont

$$L_0(x) = 1$$

$$L_1(x) = 1 - x$$

$$L_2(x) = \tfrac{1}{2}x^2 - 2x + 1$$

$$L_3(x) = \tfrac{-1}{6}x^3 + \tfrac{3}{2}x^2 - 3x + 1$$

$$L_4(x) = \tfrac{1}{24}x^4 - \tfrac{2}{3}x^3 + 3x^2 - 4x + 1$$

$$L_5(x) = \tfrac{-1}{120}x^5 + \tfrac{5}{24}x^4 - \tfrac{5}{3}x^3 + 5x^2 - 5x + 1$$

$$L_6(x) = \tfrac{1}{720}x^6 - \tfrac{1}{20}x^5 + \tfrac{5}{8}x^4 - \tfrac{10}{3}x^3 + \tfrac{15}{2}x^2 - 6x + 1$$

(1) Les polynômes de Laguerre sont solutions de l'équation différentielle

$$xy'' + (\alpha + 1 - x)y' + ny = 0$$

(2) Les polynômes de Laguerre satisfont les relations de récurrence, pour α entier

$$xL_n'(x) = nL_n(x) - nL_{n-1}(x)$$

$$x\frac{dL_n^{(\alpha)}(x)}{dx} = nL_n^{(\alpha)}(x) - (n+\alpha)L_{n-1}^{(\alpha)}(x)$$

$$L_{n+1}^{(\alpha-1)}(x) = L_n^{(\alpha)}(x) - L_{n-1}^{(\alpha)}(x)$$

$$xL_n^{(\alpha+1)}(x) = (x-n)L_n^{(\alpha)}(x) - (n+\alpha)L_{n-1}^{(\alpha)}(x)$$

$$xL_n^{(\alpha+1)}(x) = (n+\alpha+1)L_n^{(\alpha)}(x) - (n+1)L_{n+1}^{(\alpha)}(x)$$

$$(n+\alpha)L_n^{(\alpha-1)}(x) = (n+1)L_{n+1}^{(\alpha)}(x) - (n+1-x)L_n^{(\alpha)}(x)$$

(3) Formule de Rodrigues

$$L_n^{(\alpha)}(x) = \frac{e^x}{n!x^\alpha}\frac{d^n}{dx^n}\left(x^{n+\alpha}e^{-x}\right)$$

(4) Majorations

$$\forall x \geq 0, \qquad |L_n(x)| \leq e^{x/2}$$

$$\forall x \geq 0, \forall \alpha \geq 0 \qquad \left|L_n^{(\alpha)}(x)\right| \leq \frac{\Gamma(\alpha+n+1)}{n!\Gamma(\alpha+1)}e^{x/2}$$

(5) Les polynômes de Laguerre sont des polynômes orthogonaux relativement à la fonction de poids $\omega(x) = x^\alpha e^{-x}$ définie sur l'intervalle $]0, \infty[$

$$\int_0^\infty L_n^{(\alpha)}(x) L_m^{(\alpha)}(x) x^\alpha e^{-x} dx = \frac{\Gamma(\alpha + n + 1)}{n!} \delta_{n,m}$$

(6) L'intégrale du produit de deux polynômes de Laguerre vérifie

$$\int_0^x L_n(t) L_m(x - t) dt = L_{m+n}(x) - L_{m+n+1}(x)$$

(7) Si $Re(\alpha) > -1$ et $Re(\beta) > 0$, on a

$$\Gamma(\alpha+\beta+n+1) \int_0^x (x-t)^{\beta-1} t^\alpha L_n^{(\alpha)}(t) dt = \Gamma(\alpha+n+1)\Gamma(\beta) x^{\alpha+\beta} L_n^{(\alpha+\beta)}(x)$$

et

$$\int_x^\infty e^{-t} L_n^{(\alpha)}(t) dt = e^{-x} \left(L_n^{(\alpha)}(x) - L_{n-1}^{(\alpha)}(x) \right)$$

A.3 Polynômes de Tchebychev

Les *polynômes de Tchebychev* (de première espèce) d'ordre n, sont définis par la relation de récurrence

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

et les conditions d'initialisation

$$T_0(x) = 1 \quad T_1(x) = x$$

Les premiers polynômes sont

$$\begin{aligned}
T_0(x) &= 1 \\
T_1(x) &= x \\
T_2(x) &= 2x^2 - 1 \\
T_3(x) &= 4x^3 - 3x \\
T_4(x) &= 8x^4 - 8x^2 + 1 \\
T_5(x) &= 16x^5 - 20x^3 + 5x \\
T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1
\end{aligned}$$

(1) Le polynôme T_n peut être défini par la relation

$$T_n(\cos\theta) = \cos(n\theta)$$

ou bien encore par la relation

$$T_n(x) = \frac{1}{2} \left(\left(x + \sqrt{x^2 + 1} \right)^n + \left(x - \sqrt{x^2 - 1} \right)^n \right)$$

$T_n(x)$ est un polynôme de degré n dont le coefficient de plus haut degré est 2^{n-1} vérifiant

$$T_n(1) = 1 \quad \forall n$$

Ces polynômes vérifient la relation

$$T_n(-x) = (-1)^n T_n(x)$$

(2) Les polynômes de Tchebychev sont solutions de l'équation différentielle

$$(1-x^2)y'' - xy' + n^2 y = 0$$

(3) Les polynômes de Tchebychev satisfont la relation de récurrence

$$(1-x^2)T_n'(x) = -nxT_n(x) + nT_{n-1}(x)$$

(4) Pour $i = 0, ..., n$ la relation du produit de deux polynômes

$$2T_i(x)T_n(x) = T_{n+i}(x) + T_{n-i}(x)$$

(5) Majorations

$$\forall x \in [-1,1]\,, \qquad |T_n(x)| \leq 1$$

$$\forall x \in [-1,1]\,, \qquad \left|\frac{dT_n(x)}{dx}\right| \leq n^2$$

(6) Les polynômes de Tchebychev sont des polynômes orthogonaux relativement à la fonction de poids définie sur l'intervalle $[-1,1]$

$$\omega(x) = \frac{1}{\sqrt{1-x^2}}$$

$$\int_{-1}^{+1} T_n(x)T_m(x)\frac{dx}{\sqrt{1-x^2}} = \frac{\pi}{2}\delta_{n,m} \qquad n \neq 0$$

$$\int_{-1}^{+1} T_0^2(x)\frac{dx}{\sqrt{1-x^2}} = \pi$$

(7) Si ν désigne la partie entière de $n/2$, les monômes s'expriment en fonction des polynômes de Tchebychev

$$x^n = \frac{1}{2^{n-1}}\left[T_n + C_n^1 T_{n-2} + \cdots + C_n^{\nu-1}T_{n+2-2\nu} + C_n^{\nu}T_{n-2\nu}\left(\frac{3-(-1)^n}{4}\right)\right]$$

(8) Entre les abscisses

$$x_k^+ = \cos\left(\frac{2k\pi}{n}\right)$$

pour lesquelles $T_n(x_k^+) = +1$ et les abscisses

$$x_k^- = \cos\left(\frac{(2k+1)\pi}{n}\right)$$

pour lesquelles $T_n(x_k^-) = -1$, le polynôme de Tchebychev de degré n admet exactement n racines réelles données par

$$\cos\left(\frac{(2k+1)\pi}{2n}\right) \qquad k = 0, 1, \ldots, n-1$$

A.4 Polynômes d'Hermite

Les *polynômes d'Hermite d'ordre n* sont les polynômes définis par la relation de récurrence

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$$

et les conditions d'initialisation

$$H_0(x) = 1 \quad H_1(x) = 2x$$

Les premiers polynômes sont

$$\begin{aligned}
H_0(x) &= 1 \\
H_1(x) &= 2x \\
H_2(x) &= 4x^2 - 2 \\
H_3(x) &= 8x^3 - 12x \\
H_4(x) &= 16x^4 - 48x^2 + 12 \\
H_5(x) &= 32x^5 - 160x^3 + 120x \\
H_6(x) &= 64x^6 - 480x^4 + 720x^2 - 120
\end{aligned}$$

(1) Les polynômes d'Hermite satisfont l'équation différentielle

$$y'' - 2xy' + 2ny = 0$$

(2) Les polynômes d'Hermite vérifient la relation de récurrence

$$H_n'(x) = 2nH_{n-1}(x)$$

(3) Formule de Rodrigues

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2})$$

(4) Majorations

$$|H_n(x)| < e^{x^2/2} k 2^{n/2} \sqrt{n!} \qquad avec \qquad k \simeq 1,086435...$$

(5) Les polynômes d'Hermite sont des polynômes orthogonaux relativement à la fonction de poids $\omega(x) = e^{-x^2}$

$$\int_{-\infty}^{+\infty} H_n(x) H_m(x) e^{-x^2} dx = \sqrt{\pi} 2^n n! \delta_{n,m}$$

(6) Formules d'intégration

$$\int_0^x H_n(t)e^{-t^2}dt = H_{n-1}(0) - e^{-x^2}H_{n-1}(x)$$

$$\int_0^x H_n(t)dt = \frac{1}{2(n+1)}\left(H_{n+1}(x) - H_{n+1}(0)\right)$$

$$\int_{-\infty}^{+\infty} H_{2n}(tx)e^{-t^2}dt = \sqrt{\pi}\frac{(2n)!}{n!}(x^2-1)^n$$

$$\int_{-\infty}^{+\infty} te^{-t^2}H_{2n+1}(tx)dt = \sqrt{\pi}\frac{(2n+1)!}{n!}x(x^2-1)^n$$

A.5 Polynômes de Gegenbauer

Les *polynômes de Gegenbauer* $G_n^{(\alpha)}$ de degré n, sont les polynômes définis par la relation de récurrence pour $\alpha > -1/2$

$$(n+1)G_{n+1}^{(\alpha)}(x) = 2(n+\alpha)G_n^{(\alpha)}(x) - (n+2\alpha-1)G_{n-1}^{(\alpha)}(x)$$

et les conditions d'initialisation

$$G_0^{(\alpha)}(x) = 1 \quad G_1^{(\alpha)}(x) = 2\alpha x \quad \text{si} \quad \alpha \neq 0 \quad G_1^{(0)}(x) = 2x$$

Les premiers polynômes sont pour $\alpha = 1$

$$\begin{aligned}
G_0(x) &= 1\\
G_1(x) &= 2x\\
G_2(x) &= 4x^2 - 1\\
G_3(x) &= 8x^3 - 4x\\
G_4(x) &= 16x^4 - 12x^2 + 1\\
G_5(x) &= 32x^5 - 32x^3 + 6x\\
G_6(x) &= 64x^6 - 80x^4 + 24x^2 - 1
\end{aligned}$$

(1) Les polynômes de Gegenbauer sont solutions de l'équation différentielle

$$(1-x^2)y'' - (2\alpha+1)xy' + n(n+2\alpha)y = 0$$

(2) Les polynômes de Gegenbauer vérifient les relations de récurrence

$$(1-x^2)\frac{dG_n^{(\alpha)}}{dx}(x) = -nxG_n^{(\alpha)}(x) + (n+2\alpha-1)G_{n-1}^{(\alpha)}(x)$$

$$(n+\alpha)G_{n+1}^{(\alpha-1)}(x) = (\alpha-1)\left(G_{n+1}^{(\alpha)}(x) - G_{n-1}^{(\alpha)}(x)\right)$$

(3) Formule de Rodrigues

$$G_n^{(\alpha)}(x) = \frac{(-1)^n}{2^n n!}\frac{\Gamma(\alpha+1/2)\Gamma(n+2\alpha)}{\Gamma(2\alpha)\Gamma(\alpha+n+1/2)}\frac{d^n}{dx^n}\left((1-x^2)^{\alpha+n-\frac{1}{2}}\right)$$

(4) Les polynômes de Gegenbauer sont des polynômes orthogonaux sur l'intervalle $]-1,+1[$ relativement à la fonction de poids $\omega(x) = (1-x^2)^{\alpha-\frac{1}{2}}$, on a si $\alpha \neq 0$

$$\int_{-1}^{+1} G_n^{(\alpha)}(x)G_m^{(\alpha)}(x)(1-x^2)^{\alpha-\frac{1}{2}}dx = \frac{\pi 2^{1-2\alpha}}{n!(n+\alpha)!}\frac{\Gamma(n+2\alpha)}{\Gamma(\alpha)^2}\delta_{n,m}$$

et si $\alpha = 0$

$$\int_{-1}^{+1} G_n^{(0)}(x)G_m^{(0)}(x)(1-x^2)^{-\frac{1}{2}}dx = \frac{2\pi}{n^2}\delta_{n,m}$$

(5) Formule d'intégration

$$\frac{n(2\alpha+n)}{2\alpha}\int_0^x G_n^{(\alpha)}(t)(1-t^2)^{\alpha-\frac{1}{2}}dx = G_{n-1}^{(\alpha+1)}(0) - (1-x^2)^{\alpha+\frac{1}{2}}G_{n-1}^{(\alpha+1)}(x)$$

A.6 Polynômes de Jacobi

Les *polynômes de Jacobi* de degré n, notés $J_n^{(\alpha,\beta)}(x)$ ou $J_n(x)$ lorsqu'il n'y a pas ambiguïté, sont les polynômes définis par la relation de récurrence pour $\alpha > -1$ et $\beta > -1$

$$a_n J_{n+1}(x) = (b_n + xc_n)J_n(x) - d_n J_{n-1}(x)$$

initialisée par

$$J_0(x) = 1 \quad J_1(x) = (\alpha-\beta)/2 + (1+(\alpha+\beta)/2)x$$

avec les coefficients suivants

$$\begin{aligned} a_n &= 2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta) \\ b_n &= (2n+\alpha+\beta+1)(\alpha^2-\beta^2) \\ c_n &= (2n+\alpha+\beta) \\ d_n &= 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2) \end{aligned}$$

Les premiers polynômes sont pour $\alpha = 1$ et $\beta = 0$

$$\begin{aligned}
J_0(x) &= 1\\
J_1(x) &= (3x+1)/2\\
J_2(x) &= (5x^2+2x-1)/2\\
J_3(x) &= \frac{35}{8}x^3 + 15x^2 - \frac{15}{8}x - \frac{3}{8}
\end{aligned}$$

$$J_4(x) = \frac{63}{8}x^4 + \frac{7}{2}x^3 - \frac{21}{4}x^2 - \frac{3}{2}x + \frac{3}{8}$$

$$J_5(x) = \frac{231}{16}x^5 + \frac{105}{16}x^4 - \frac{105}{8}x^3 - \frac{35}{8}x^2 + \frac{35}{16}x + \frac{5}{16}$$

(1) Les polynômes de Jacobi satisfont l'équation différentielle

$$(1-x^2)y'' + (\beta - \alpha - (\alpha+\beta+2)x)y' + n(n+\alpha+\beta+1)y = 0$$

(2) En posant $\lambda = 2n+\alpha+\beta$, les polynômes de Jacobi vérifient les relations

$$\lambda(1-x^2)J_n'(x) = n(\alpha-\beta-\lambda x)J_n(x) + 2(n+\alpha)(n+\beta)J_{n-1}(x)$$

$$J_n^{(\alpha,\beta-1)}(x) - J_n^{(\alpha-1,\beta)}(x) = J_{n-1}^{(\alpha,\beta)}(x)$$

$$\lambda J_n^{(\alpha,\beta-1)}(x) = (n+\alpha+\beta)J_n^{(\alpha,\beta)}(x) + (n+\alpha)J_{n-1}^{(\alpha,\beta)}(x)$$

$$(1-x)J_n^{(\alpha+1,\beta)}(x) + (1+x)J_n^{(\alpha,\beta+1)}(x) = 2J_n^{(\alpha,\beta)}(x)$$

(3) Formule de Rodrigues

$$J_n(x) = \frac{(-1)^n}{2^n n!}\frac{1}{(1-x)^\alpha(1+x)^\beta}\frac{d^n}{dx^n}\left((1-x^2)^n(1-x)^\alpha(1+x)^\beta\right)$$

(4) Les polynômes de Jacobi sont des polynômes orthogonaux sur l'intervalle $]-1,1[$ relativement à la fonction de poids

$$\omega(x) = (1-x)^\alpha(1+x)^\beta$$

$$\int_{-1}^{+1} J_n(x)J_m(x)(1-x)^\alpha(1+x)^\beta dx = u_n\delta_{n,m}$$

avec

$$u_n = \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1}\frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{n!\Gamma(n+\alpha+\beta+1)}$$

(5) Formule d'intégration

$$2n\int_0^x J_n^{(\alpha,\beta)}(t)(1-t)^\alpha(1+t)^\beta dt = J_{n-1}^{(\alpha+1,\beta+1)}(0) - h(x)J_{n-1}^{(\alpha+1,\beta+1)}(x)$$

où $h(x) = (1-x)^{1+\alpha}(1+x)^{1+\beta}$

Bibliographie

[1] J. Abdeljaoued, H. Lombardi, *Méthodes matricielles. Introduction à la complexité algébrique*. Mathématiques et Applications, vol. 42, Springer, 2004.

[2] R. Adams, *Sobolev Spaces*, Academic Press, 1975.

[3] R. P. Agarwal, *Boundary Value Problems for High Order Differential Equations*, World Scientific, 1986.

[4] J. Ahlberg, *The Theory of Splines and Their Applications*, Academic Press, 1967.

[5] J. Akin, *Application and Implementation of Finite Elements Methods*, Academic Press, 1982.

[6] S. Alinhac, P. Gérard, *Opérateurs pseudo-différentiels et théorème de Nash-Moser*, Éditions du CNRS, 1996.

[7] W. Ames, W. Rheinboldt, *Numerical Methods for Partial Differential Equations*, Academic Press, 1992.

[8] G. Anger, *Inverse Problems in Differential Equations*, Plenum Press, 1990.

[9] D. V. Anosov, V. I. Arnold, *Dynamical Systems*, Springer, 1988.

[10] K. Arbenz, A. Wohlhauser, *Analyse numérique,* Presses Polytechniques romandes, 1980.

[11] A. Arcangeli, M. Artola, J. M. Blondel, J. Grenet, *Problèmes d'analyse numérique, agrégation années 1969-1978*, Masson, 1980.

[12] V. I. Arnold, *Ordinary Differential Equations*, Springer, 1992.

[13] F. M. Arscott, *Periodic Differential Equations. An introduction to Mathieu, Lamé and Allied Functions*, Mac Millan, 1964.

[14] U. M. Ascher, R. M. Mattey, R. D. Russel, *The numerical solution of boundary value problems for ordinary equations*, Prentice Hall, 1987.

[15] K. E. Atkinson, *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind*, Philadelphia, S.I.A.M., 1976.

[16] O. Axelsson, V. Baker, *Finite Element Solution of Boundary Problems*, Academic Press, 1984.

[17] O. Axelson, *Iterative Solution Method*, Cambridge University Press, 1996.

[18] A. Aziz, *Lectures in Differential Equations*, Van Nostrand, 1969.

[19] G. Bader, P. Deuflhard, *Numerische Mathematik*, vol. 41, 1983.

[20] A. Baker, *Finite Element Computational Fluid Mechanics*, Hemisphere Publishing Corp. 1983.

[21] G. Baker, *Essentials of Pade Approximants*, Academic Press, 1975.

[22] G. Baker, *Pade Approximant*, Addison Wesley, 2 vol., 1981.

[23] N. Bakhvalov, *Méthodes numériques*, Moscou, Mir, 1976.

[24] J. Baranger, *Introduction à l'analyse numérique*, Hermann, 1977.

[25] R. Bartels, *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*, Keufmann, 1987.

[26] K. Bathe, E. Wilson, *Numerical Methods in Finite Element Analysis*, Prentice Hall, 1976.

[27] A. Bamberger, *La Méthode des éléments finis*, Polycopié de Paris-VI, 1982.

[28] A. Bamberger, *Analyse, optimisation et filtrage numérique, Compléments, analyse numérique de l'équation de la chaleur*, École Polytechnique, 1990.

[29] J. Baranger, *Analyse numérique*, Hermann, 1991.

[30] H. Bastin, *Éléments d'analyse numérique*, Presses Universitaires de Bruxelles, 1972.

[31] R. Bellman, *Stability Theory of Differential Equations*, Dover, 1969.

[32] R. Bellman, *Methods in Approximation*, Reidel Publishing Corp., 1986.

[33] A. Beltzer, *Variational and Finite Elements Methods*, Springer, 1990.

[34] J. Bergh, J. Lofstrom, *Interpolation Spaces*, Springer, 1976.

[35] J. S. Berezin, N. P. Zhidkov, *Computing methods*, Pergamon Press (traduit du russe), 1973.

[36] M. Bernadou, *Méthodes d'éléments finis pour les problèmes de coques minces*, Masson, 1994.

[37] C. Bernardi, Y. Maday, *Approximations spectrales de problèmes aux limites elliptiques*, Springer Verlag, Mathématiques et Applications vol. 10, 1992.

[38] G. Birkhoff, *Ordinary Differential Equations*, John Wiley, 1989.

[39] C. Blanc, *Equations aux dérivées partielles : un cours pour ingénieurs*, Birkhäuser, 1976.

[40] E. Blum, *Numerical Analysis and Computation Theory and Practice*, Addison Wesley, 1972.

[41] G. Bluman, S. Kumei, *Symmetries and Differential Equations*, Springer, 1989.

[42] J.-M. Bony, J.-Y. Chemin, C. Gérard, G. Lebeau, *Équations aux dérivées partielles*, Majeure de mathématiques, École Polytechnique, 1997.

[43] J. F. Botha, *Fundamental Concepts in the Numerical Solution of Differential Equations*, John Wiley, 1983.

[44] M. Braun, *Differential equations and their applications*, Springer, 1975.

[45] C. Brebbia, J. Connor, *Fundamentals of Finite Element Technique for Structural Engineers*, Butterworths, 1973.

[46] S. Brenner, *The Mathematical Theory of Finite Elements Methods*, Springer, 1994.

[47] R. P. Brent, *Algorithms for Minimization without Derivatives*, Prentice-Hall, 1973.

[48] C. Brezinski, *Accélération de la convergence en analyse numérique*, Springer, 1977.

[49] C. Brezinski, *Padé Type Approximation and General Orthogonal Polynomials*, Birkäuser, 1980.

[50] H. Brezis, *Analyse fonctionnelle*, Paris, Masson, 1987.

[51] F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element methods*, Springer Series in Comp. Math, 15, 1991.

[52] W. L. Briggs, *A Multigrid Tutorial*, Philadelphia, S.I.A.M., 1987.

[53] Ya. Brudnyi, N. Y. Krugljak, *Interpolation Functions and Interpolation Spaces,* North Holland, 1991.

[54] J. Bull, *Finite Element Analysis of Thin-walled Structures*, Elsevier, 1988.

[55] J. R. Bunch, D. J. Rose (eds), *Sparse Matrix Computations*, Academic Press, 1976.

[56] R. Burden, D. Faires, *Numerical Analysis*, Prindle, Weber and Schmidt, 1985.

[57] D. Burnett, *Finite Element Analysis, From Concepts to Applications*, Addison Wesley, 1987.

[58] T. A. Burton, *Stability and Periodic Solutions of Ordinary and Functional Differential Equations*, Academic Press, 1985.

[59] J. C. Butcher, *The numerical analysis of ordinary differential equations, Runge-Kutta and general linear methods*, Wiley, 1987.

[60] H. Cabannes, *Pade Approximants Method and its Applications to Mechanics*, Springer, 1976.

[61] C. Canuto, M.Y. Hussiaini, A. Quarteroni, T.A. Zang, *Spectral Methods in Fluid Dynamics*, Springer Verlag, 1988.

[62] B. Carnahan, H. A. Luther, J. O. Wilkes, *Applied Numerical Methods*, Wiley, 1969.

[63] J. L. Chabert *et al.*, *Histoire d'algorithmes*, Paris, Belin, 1994.

[64] B. Carnahan, H. A. Luther, J. O. Wilkes, *Applied Numerical Methods*, John Wiley, 1969.

[65] G. Carrier, C. Pearson, *Partial Differential Equations*, Academic Press, 1988.

[66] J. R. Cash, A. H. Karp, *ACM Transactions on Mathematical Software*, vol. 16, 1990, 201-222.

[67] J. Chazarin, A. Piriou, *Introduction à la théorie des équations aux dérivées partielles*, Gauthiers Villars, 1981.

[68] E. W. Cheney, *Introduction to approximation theory*, Chelsea, reprint, 1982.

[69] C. Chester, *Techniques in Partial Differential Equations*, Mc Graw Hill, 1971.

[70] T. J. Chung, W. J. Minkowycz, E. M. Sparrow, *Finite Elements in Fluids*, Hemisphere Publishing Corp. 1992.

[71] P. G. Ciarlet, *Introduction à l'analyse matricielle et à l'optimisation*, Paris, Masson, 1990.

[72] P. G. Ciarlet, *Exercices d'analyse matricielle*, Paris, Masson, 1990.

[73] P. G. Ciarlet, *Les équations de Von Karman*, Springer Verlag, 1980.

[74] P. G. Ciarlet, *Numerical Analysis of the Finite Element Method*, Presses Universitaires de Montréal, 1976.

[75] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, 1978.

[76] P. G. Ciarlet, *Mathematical Elasticity*, North-Holland, 1988.

[77] P. G. Ciarlet, J. L. Lions, *Handbook of Numerical Analysis*, North-Holland, 1990.

[78] P. G. Ciarlet, B. Miara, J. M. Thomas, *Exercices d'analyse matricielle et d'optimisation*, Paris, Masson, 1991.

[79] P. Clarkson, *Applications of Analytric and Geometric Methods to Nonlinear Differential Equations*, Kluwer, 1993.

[80] E. Coddington, N. Levinson, *Theory of Ordinary Differential Equations*, Krieger, 1986.

[81] T. F.Coleman, C. Van Loan, *Handbook for Matrix Computations*, Philadelphia, S.I.A.M.,1988.

[82] S. Colombo, *Les Équations aux dérivées partielles en physique et en mécanique des milieux continus*, Masson, 1976.

[83] P. Constatin, C. Foias, *Navier-Stokes Equations*, University Chicago Press, 1988.

[84] H. O. Cordes, *Spectral Theory of Linear Differential Operators and Comparison Algebras*, Cambrige University Press, 1987.

[85] M. Crouzeix, A. L. Mignot, *Analyse numérique des équations différentielles*, Paris, Masson, 1989.

[86] M. Crouzeix, A. L. Mignot, *Exercices d'analyse numérique des équations différentielles*, Paris, Masson, 1989.

[87] C. Cuvelier, A. Segal, A. van Steenhoven, *Finite Element Methods and the Navier-Stokes Equation*, Reidel Publishing Corp. 1986.

[88] A. Cuyt , L. Wuytack, *Nonlinear Methods in Numerical Analysis*, North-Holland, 1987.

[89] G. Dahlquist, A. Bjorck, *Numerical Methods*, Prentice Hall, 1974.

[90] R. Dautray, J. L. Lions, *Analyse mathématique et calcul numérique*, Masson, 10 vol., 1984.

[91] A. Davies, *The Finite Element Method, A first Approach*, Oxford University Press, 1980.

[92] H. T. Davis, *Introduction to Nonlinear Differential and Integral Equations*, Dover, 1962.

[93] Ph. J. Davis, Ph. Rabinowitz, *Methods of Numerical Integration*, Academic Press, 2nd ed., 1984.

[94] L. Debnath, *Nonlinear Partial Differential Equations for Scientists and Engineers*, Birkhäuser, 2nd edition, 2005.

[95] C. De Boor, *A Pratical Guide to Splines*, Springer Verlag, 1978.

[96] K. Dekker, J. Verwer, *Stability of Runge-Kutta methods for shiff non linear differential equations*, CWI Monographs, North Holland, 1984.

[97] L. M. Delves, J. L. Mohamed, *Computational Methods for Integral Equations*, Cambridge University Press, 1985.

[98] J.-P. Demailly, *Analyse numérique et équations différentielles*, Presses Universitaires de Grenoble, 1991.

[99] J. E. Dennis, R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, 1983.

[100] G. Dhatt, G. Touzot, *Une présentation de la méthode des éléments finis*, Maloine, 1981.

[101] J. J. Dongarra *et al.*, *LINPACK User's Guide*, Philadelphia, S.I.A.M., 1979.

[102] G. D. Doolen, *Lattice Gaz Methods for Partial Differential Equations*, Addison Wesley, 1990.

[103] J. R. Dormand, P. J. Prince, "A family of embedded Runge-Kutta formulae", *J. Comp. Appl. Math.*, vol. 6, 1980, pp. 19-26.

[104] A. Draux, *Polynômes orthogonaux et approximants de Padé*, Technip, 1987.

[105] I. Duff, *Sparse Matrix and Their Uses*, Academic Press, 1981.

[106] D. Edelen, *Transformation Methods for Nonlinear Differential Equations*, World Scientific, 1992.

[107] I. U. V. Egorov, M. A. Shubin, *Partial Differential Equations*, Springer, 1992.

[108] I. Ekeland, R. Temam, *Analyse convexe et problèmes variationnels*, Dunod-Gauthier Villars, Paris, 1974.

[109] J. Elschner, *Singular Ordinay Differential Operators and Pseudo-differential Equations*, Springer, 1985.

[110] D. Euvrard, *Résolution numérique des équations aux dérivées partielles*, Masson, 1988.

[111] S. Farlow, *Partial Differential Equations for Scientits and Engineers*, John Wiley, 1982.

[112] S. O. Fatunla, *Numerical Methods for Initial Value Problems in Ordinary Differential Equations*, Academic Press, 1988.

[113] P. Faurre, *Analyse numérique, notes d'optimisation*, École Polytechnique, 1988.

[114] G. Folland, *Introduction to Patial Differential Equations*, Princeton University Press, 1976.

[115] G. E. Forsythe, W. R. Wasov, *Finite Difference Methods for Partial Differential Equations*, John Wiley, 1960.

[116] G. E. Forsythe, M. A. Malcolm, C. B. Moler, *Computer Methods for Mathematical Computations*, Prentice Hall, 1977.

[117] G. E. Forsythe, C. B. Moler, *Computer Solution of Linear Algebraic Systems*, New-York, Prentice Hall, 1967.

[118] L. Fox, *Numerical Solution of Ordinary and Partial Differential Equations*, Addison Wesley, 1962.

[119] L. Fox, *Chebychev Polynomials in Numerical Analysis*, Oxford University Press, 1968.

[120] L. Fox, *Numerical Solution of Ordinary Differential Equations*, Chapman and Hall, 1987.

[121] L. Fox, D. F. Mayers, *Computing methods for scientists and engineers*, Clarendon Press, 1968.

[122] I. Fried, *Numerical Solution of Differential Equations*, Academic Press, 1979.

[123] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice Hall, 1964.

[124] S. Fucik, A. Kufner, *Nonlinear Differential Equations*, Elsevier, 1980.

[125] R. H. Gallagher, *Introduction aux éléments finis*, Pluralis édition, 1966.

[126] P. Garabedjan, *Partial Differential Equations*, Wiley, 1964.

[127] N. Gastinel, *Analyse numérique linéaire*, Hermann, 1966.

[128] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, 1971.

[129] C. F. Gerald, *Applied Numerical Analysis*, Addison-Wesley, 1970.

[130] R. P. Gilbert, R. J. Weinhacht, *Function Theoretic Methods in Differential Equations*, Pitman, 1976.

[131] J. Gilewicz, *Approximants de Padé*, Springer, 1978.

[132] P. E. Gill, W. Murray, M. H. Wright, *Numerical Linear Algebra and Optimization*, 2 vol, Addison-Wesley, 1991.

[133] V. Girault, P.-A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Springer, 1979.

[134] I. Gladwell, R. Wait, *A Survey of Numerical Methods for Partial Differential Equations*, Oxford University Press, 1979.

[135] E. Godlewski, P.-A. Raviart, *Hyperbolic Systems of Conservation Laws*, Ellipses, 1991.

[136] E. Godlewski, P.-A. Raviart, J.E. Marsden (eds), *Numerical approximation of hyperbolic systems of conservation laws,* Springer, 1996.

[137] G. H. Golub, G. Meurant, *Résolution numérique des grands systèmes linéaire*s, Eyrolles, 1982.

[138] G. H. Golub, C. F. Van Loan, *Matrix Computations*, Baltimore, John Hopkins University Press, 2nd ed., 1989.

[139] V. I. Gorbachuk, *Boundary Value Problems for Operator Differential Equations*, Kluwer, 1991.

[140] D. Gottlieb, S. Orszag, *Numerical Analysis of Spectral Methods, Theory and Applications*, SIAM, 1977.

[141] P. Gould, *Finite Element Analysis of Shells of Revolution*, Pitman, 1985.

[142] C. D. Green, *Integral Equations Methods*, New-York, Barnes & Noble, 1969.

[143] D. Greenspar, V. Casulli, *Numerical Analysis for Applied Mathematics, Science and Engineering*, Addison Wesley, 1988.

[144] M. Gregus, *Third Order Linear Differential Equations*, Reidel Pub. Co., 1987.

[145] D. F. Griffiths, G. A. Watson, *Numerical Analysis*, Longman Scientific and Technical, 1986.

[146] M. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic Press, 1989.

[147] K. Gustafson, *Introduction to Partial Differential Equations and Hilbert Space Methods*, John Wiley, 1980.

[148] W. Hackbush, *Multigrid Methods and Applications*, Springer, 1985.

[149] W. Hackbush, U. Trottenberg, *Multigrid Methods*, Lecture Notes in Mathematics, Springer, 1982.

[150] E. Hairer, G. Wanner, *Solving Ordinary Differential Equations, 1. Non stiff problems*, Springer, 1987.

[151] J. Hale, *Functional Differential Equations*, Springer, 1971.

[152] R. W. Hamming, *Numerical Methods for Engineers and Scientists*, New-York, Dover, (1962), reprint 1986.

[153] S. I. Hariharan, T. H. Moulten, *Numerical Methods for Partial Differential Equations*, Longman Scientific and Technical, 1986.

[154] J. F. Hart *et al.*, *Computer Approximations*, Wiley, 1968.

[155] J. P. Hennart, *Numerical Analysis,* Proceeding of the third IIMAS, Springer, 1982.

[156] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, 1962.

[157] P. Henrici, *Applied and Computational Complex Analysis*, Wiley, 1974.

[158] F. B. Hildebrand, *Introduction to Numerical Analysis*, Mc Graw Hill, 1974.

[159] E. Hinton, D. Owen, *Finite Element Software for Plates and Shells*, Pineridge Press, 1984.

[160] A. S. B. Holland, B. N. Sahney, *The general problem of approximation and spline functions*, Krieger, 1979.

[161] M. Holt, *Numerical Method in Fluid Dynamics*, Springer Verlag, 1977.

[162] L. Hörmander, *Linear Partial Differential Operators*, Springer, 1963.

[163] A. S. Householder, *The Numerical Treatment of a Single Nonlinear Equation*, Mc Graw Hill, 1970.

[164] T. Hughes, *The Finite Element Method*, Prentice Hall, 1987.

[165] J. F. Imbert, *Analyse de structures par éléments finis*, Cepaduès, Toulouse, 1979.

[166] E. Isaacson, H. B. Keller, *Analysis of Numerical Methods*, New-York, John Wiley, 1966.

[167] L. G. Ixaru, *Numerical Methods for Differential Equations by the Finite Element Method*, Editura Academicil, 1984.

[168] D. A. H. Jacobs (ed.), *The State of the Art in Numerical Analysis*, Academic Press, 1977.

[169] M. K. Jain, *Numerical Solution of Differential Equations*, Wiley Eastern, 1984.

[170] D. Jespersen, *Multigrid Methods for Partial Differential Equations*, Washington, Mathematical Association of America, 1984.

[171] F. John, *Lectures on Advanced Numerical Analysis*, Gordon and Breach, 1967.

[172] F. John, *Partial Differential Equations*, Springer, 1975.

[173] C. Johnson, *Numerical Solution of Partial Differential Equation by the Finite Element Method*, Cambridge University Press, 1987.

[174] L. W. Johnson, R. D. Riess, *Numerical Analysis*, Addison-Wesley, 2nd ed. 1982.

[175] P. Joly, *Mise en œuvre de la méthode des éléments finis*, Paris, Ellipses, 1990.

[176] W. Joppich, S. Mijalkovic, *Multigrid Methods for Process Simulation*, Springer, 1993.

[177] D. W. Jordan, *Nonlinear Ordinary Differential Equations*, Oxford University Press, 1987.

[178] D. Kahaner, C. Moler, S. Nash, *Numerical Methods and Software*, New York, Prentice Hall, 1989.

[179] R. P. Kanwal, *Linear Integral Equations*, Academic Press, 1971.

[180] H. Kardestuncer, *Unification of Finite Elements Methods*, North-Holland, 1984.

[181] S. Karlin, *Studies in Spline Functions and Approximation Theory*, Academic Press, 1976.

[182] T. Kato, *Perturbation Theory for Linear Operators*, Springer, 1976.

[183] O. Kavian, *Introduction à la théorie des points critiques et applications aux problèmes elliptiques*, Springer Verlag, Mathématiques et Applications vol. 13, 1993.

[184] H. B. Keller, *Numerical Methods for Two Point Boundary Value Problems*, Waltham, Blaisdell, 1968.

[185] G. Kirov, *Approximation with Quasi-Splines*, Adam Hilger, 1992.

[186] D. E. Knuth, "*Fundamental Algorithms*", in *The Art of Computer Programming*, vol. 1, Addison Wesley, 1968.

[187] A. Korganoff *et al.*, *Méthodes de calcul numérique, Tome 1, Algèbre non linéaire, Tome 2, Éléments de théorie des matrices carrées et rectangles en analyse numérique*, Dunod, 1961 et 1967.

[188] M. Kracht, *Methods of Complex Analysis in Partial Differential Equations with Applications*, John Wiley, 1988.

[189] N. Krasovski, *Stability of Motion*, Stanford University Press, 1963.

[190] H. P. Kuenzi, H. G. Tzschach, C. A. Zehnder, *Numerical Methods of Mathematical Optimization*, Addison Wesley, 1971.

[191] J. Kurzweil, *Ordinary Differential Equations, Introduction to the theory of ordinary differential equations in real domain*, Amsterdam, Elsevier, 1986.

[192] O. Ladyzhenokaia, *Équations aux dérivées partielles de type elliptique*, Dunod, 1968.

[193] V. Lakshmikantham, D. Bainov, P. Simeonov, *Theory of Impulsive Differential Equations*, World Scientific, 1989.

[194] J. D. Lambert, *Computational Methods in Ordinary Differential Equations*, John Wiley, 1973.

[195] L. Lapidus, W. Schiesser, *Numerical Methods for Differential Systems. Recent Developments in Algorithms, Software and Applications*, Academic Press, 1976.

[196] L. Lapidus, J. Steinfeld, *Numerical Solution of Ordinary Differential Equations*, Academic Press, 1971.

[197] L. Lapidus, G. Pinder, *Numerical Solution of Partial Differential Equations in Science and Engineering*, John Wiley, 1982.

[198] P. Lascaux, R. Theodor, *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, 2 vol., Masson, 1987.

[199] I. Lasiecka, *Differential and Algebraic Riccati Equations with Application to Boundary Point Control Problems*, Springer, 1991.

[200] P. J. Laurent, *Approximation et optimisation*, Hermann, 1972.

[201] A. Law, C. Wang, *Approximation, Optimization and Computing Theory and Applications*, North-Holland, 1990.

[202] C. L. Lawson, R. Hanson, *Solving Least Squares Problems*, Prentice-Hall, 1974.

[203] J. Legras, *Méthodes et techniques de l'analyse numérique*, Dunod, 1971.

[204] D. Leguillon, E. Sanchez-Palencia, *Computation of Singular Solutions in Elliptic Problems and Elasticity*, Masson, 1987.

[205] A. Le Pourhiet, *Résolution numérique des équations aux dérivées partielles, une première approche*, Cepaduès Éditions, Toulouse, 1988.

[206] P. Le Tallec, *Numerical Analysis of Viscoelastic Problem*s, Masson, RMA, 1990.

[207] A. Leung, *Systems of Nonlinear Partial Differential Equations with Applications to Biology and Engineering*, Kluwer, 1989.

[208] T. Li, *Global Classical Solutions for Quasilinear Hyperbolic Systems*, John Wiley, 1993.

[209] W. Lick, *Difference Equations from Differential Equations*, Springer, 1989.

[210] W. Light, *Advances in Numerical Analysis*, Clarendon, 1991.

[211] P. Linz, *Theoretical Numerical Analysis, An Introduction to Advanced Techniques*, Wiley, 1979.

[212] P. Linz, *Analytical and Numerical Methods for Volterra Equations*, Philadelphia, S.I.A.M., 1985.

[213] J.-L. Lions, É. Mangenes, *Problèmes aux limites non homogènes et applications*, 2 vol. Dunod, 1968.

[214] J.-L. Lions, *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Dunod, 1968.

[215] J.-L. Lions, *Cours d'analyse numérique*, Hermann, 1973.

[216] Y. L. Luke, *Mathematical Functions and Their Approximations*, Academic Press, 1975.

[217] G. I. Marchuk, *Methods of Numerical Mathematics*, Springer Verlag, 1975.

[218] J. Marti, *Introduction to Sobolev Spaces and Finite Element Solution of Elliptic Boundary Value Problems*, Academic Press, 1986.

[219] R. H. Martin, *Nonlinear Operators and Differential Equations in Banach Spaces*, Wiley, 1976.

[220] H. Martin, G. Carey, *Introduction to Element Analysis, Theory and Application*, Mc Graw Hill, 1973.

[221] S. F. McCormick (ed.), *Multigrid Methods, Theory, Applications, and Supercomputing*, New-York, Marcel Dekker, 1988.

[222] G. Meinardus, *Approximation of Functions*, Springer, 1967.

[223] T. Meis, *Numerical Solution of Partial Differential Equations*, Springer, 1981.

[224] B. Mercier, *Analyse numérique des méthodes spectrales,* Springer, 1989.

[225] Y. Meyer, *Ondelettes*, Hermann, 1990.

[226] Y. Meyer, *Les Ondelettes, Algorithmes et applications*, Armand Colin, 1992.

[227] C. Miranda, *Partial Differential Equations of Elliptic Type*, Springer, 1970.

[228] W. L. Mirankar, *Numerical Methods for Stiff Equations*, Reidel, 1981.

[229] A. Mitchell, *The Finite Element Method in Partial Differential Equations*, John Wiley, 1977.

[230] A.R. Mitchell, D.F. Griffiths, *The Finite Difference Method in Partial Differential Equations*, Wiley, 1980.

[231] T. Myoshi, *Foundations of the Numerical Analysis of Plasticity*, North-Holland, 1985.

[232] S. Mizohata, *The Theory of Partial Differential Equations*, University Press, 1973.

[233] G. A. Mohr, *Finite Elements for Solids, Fluids and Optimization*, Oxford University Press, 1992.

[234] G. Murphy, *Ordinary Differential Equations and Their Solutions*, Van Nostrand, 1960.

[235] J.-C. Nedelec, *Notions sur les techniques d'éléments finis*, Ellipses, Mathématiques et Applications vol. 7, 1991.

[236] J.-P. Nougier, *Méthodes de calcul numérique*, Masson, 1983.

[237] D. Norrie, G. DeVries, *The Finite Element Method*, Academic Press, 1973.

[238] J. Noye, *Computational Techniques for Differential Equations*, North-Holland, 1984.

[239] G. Nuernberger, *Approximation by Spline Functions*, Springer, 1989.

[240] P. J. Olver, *Applications of Lie Groups to Differential Equations*, Springer, 1986.

[241] R. E. O'Malley, *Singular Perturbation Methods for Ordinary Differential Equations*, Springer, 1991.

[242] J. Ortega, *Numerical Analysis, A Second Course*, Academic Press 1972.

[243] J. Ortega, W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several variables*, Academic Press, 1970.

[244] L. Ovsiannikov, W. Ames, *Group Analysis of Differential Equations*, Academic Press, 1982.

[245] D. J. Paddon, H. Holstein, *Multigrid Methods for Integral and Differential Equations*, Oxford University Press, 1985.

[246] A. Pankov, *Bounded and Almost Periodic Solutions of Nonlinear Operator Differential Equations*, Kluwer, 1990.

[247] B. Parlett, *The symmetric eigenvalue problem*, Prentice Hall, 1980.

[248] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, 1991.

[249] D. Pepper, J. Heinrich, *The Finite Element Method*, Hemisphere Publishing Corp. 1992.

[250] L. Perko, *Differential Equations and Dynamical Systems*, Springer, 1991.

[251] P. Petrushev, V. A. Popov, *Rational Approximation of Real Functions*, Cambridge University Press, 1987.

[252] R. Peyret, T. D. Taylor, *Computational Methods of Fluid Flows*, Springer, 1983.

[253] G. Phillips, P. Taylor, *Theory and Applications of Numerical Analysis*, Academic Press, 1973.

[254] O. Pironneau, *Méthode des éléments finis pour les fluides*, Masson, 1988.

[255] E. Polak, *Computational Methods in Optimization*, Academic Press, 1971.

[256] M. Powell, *Approximation Theory and Methods*, Cambridge University Press, 1981.

[257] P. Prenter, *Spline and Variational Methods*, Wiley, 1975.

[258] W. H. Press, S. A.Terkolsky , W. T. Vetterling, B. P. Flannery, *The Art of Scientific Computing*, Cambridge University Press, 1986.

[259] S. Prossdorf, *Numerical Analysis for Integral and Related Operator Equations*, Birkhäuser, 1991.

[260] M. Protter, H. Weinberger, *Maximum Principles in Differential Equations*, Prentice Hall, 1967.

[261] A. Quateroni, A. Valli, R. Graham, J. Stoer, *Numerical Approximation of Partial Differential Equations*, Springer, 1994.

[262] P. Rabier, J.-M. Thomas, *Exercices d'analyse numérique des équations aux dérivées partielles*, Paris, Masson, 1985.

[263] A. Ralston, P. Rabinowitz, *A First Course in Numerical Analysis*, Mc Graw Hill, 1978.

[264] S. Rao, *The Finite Element Method in Engineering*, Pergamon Press, 1989.

[265] J. M. Rassias, *Counter Examples in Differential Equations and Related Topics*, World Scientific, 1991.

[266] J. Rauch, *Partial differential equations*, Springer, Graduate texts in mathematics, vol. 128, 1991.

[267] P.-A. Raviart, *Les Méthodes d'éléments finis en mécanique des fluides*, Eyrolles, 1981.

[268] P.-A. Raviart, J.-M. Thomas, *Introduction à l'analyse numérique des équations aux dérivées partielles*, Paris, Masson, 1983.

[269] H. J. Reinhardt, *Analysis of Approximation Methods for Differential and Integral Equations*, Springer, 1985.

[270] W. Rheinboldt, *Numerical Analysis of Parametrized Nonlinear Equations*, Wiley, 1986.

[271] J. R. Rice, *The approximation of functions*, Addison Wesley, 2 vol., 1964-68, *Approximation des fonctions*, (traduction française), Dunod, 1969.

[272] J.R. Rice, *Numerical Methods, Software, and Analysis*, McGraw Hill, 1983.

[273] R. D. Richtmeyer, K. W. Morton, *Difference Methods for Initial Value Problems*, John Wiley, 1967.

[274] P. J. Roache, *Computational Fluid Dynamics*, Hermosa Publ., Albuquerque, 1972.

[275] J. Robinson, *Integrated Theory of Finite Elements Methods*, Wiley, 1973.

[276] K. Rockey, *Éléments finis*, traduit de l'anglais par Claude Gomez, Paris, Eyrolles, 1979.

[277] E. Rosinger, *Nonlinear Partial Differential Equations*, North-Holland, 1990.

[278] S. Ross, *Introduction to Ordinary Differential Equations*, John Wiley, 1980.

[279] I. Rubinstein, *Partial Differential Equations in Classical Mathematical Physics*, Cambridge University Press, 1993.

[280] U. Ruede, *Mathematical and Computational Techniques for Multilevel Adaptative Methods*, SIAM, Philadelphia, 1993.

[281] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing Compagny, 1996.

[282] A. Sard, *A Book of Splines*, John Wiley, 1971.

[283] M. Schechter, *Modern Methods in Partial Differential Equations*, Mc GrawHill, 1977.

[284] M. Schultz, *Spline Analysis*, Prentice Hall, 1973.

[285] L. Schumaker, *Spline Functions*, John Wiley, 1981.

[286] H. R. Schwarz, *Numerical Analysis of Symmetric Matrices*, Prentice Hall, 1973.

[287] H. R. Schwarz, *Finite Element Methods*, Academic Press, 1988.

[288] G. Sewell, *The Numerical Solution of Ordinary and Partial Differential Equations*, Academic Press, 1988.

[289] S. Shu, *Boundary Value Problems of Linear Partial Differential Equations for Engineers and Scientists*, World Scientific, 1987.

[290] K. S. Sibirski, *Introduction to Algebraic Theory of Invariants of Differential Equations*, Manchester University Press, 1988.

[291] M. Sibony, J. L. Mardon, *Analyse numérique*, 2 tomes, Hermann, 1982.

[292] M. Sibony, *Itérations et approximations*, Hermann, 1988.

[293] S. R. Simanca, *Pseudo-differential Operators*, John Wiley, 1990.

[294] J. Singer, *Elements of Numerical Analysis*, Academic Press, 1964.

[295] S. Singh, *Approximation Theory and Spline Functions*, Reidel, 1984.

[296] B. T. Smith *et al.*, *EISPACK Guide*, Lecture Notes in Computer Science, vol. 6, Springer Verlag, 1976.

[297] G. D. Smith, *Numerical Solution of Partial Differential Equations*, Clarendon Press, 1984.

[298] F. Smithies, *Integral Equations*, Cambridge University Press, 1958.

[299] S. Sobolev, *Partial Differential Equations of Mathematical Physics*, Pergamon Press, 1964.

[300] G. Stampacchia, *Équations elliptiques du second ordre à coefficients discontinus*, Presses Universitaires de Montréal, 1966.

[301] E. Stein, W. Wendland, *Finite Element and Boundary Techniques from Mathematical and Engineering Point of View*, Springer, 1988.

[302] H. Stephani, *Differential Equations, Their Solution Using Symmetries*, Cambridge University press, 1989.

[303] H. J. Stetter, *Analysis of discretization methods for ordinary differential equations*, Springer, 1973.

[304] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, 1973.

[305] E. Stiefel, *An Introduction to Numerical Mathematics*, Academic Press, 1965.

[306] J. Stoer, R. Burlirsch, *Introduction to Numerical Analysis*, Springer Verlag, 1980.

[307] F. Strange, G. Fix, *An Analysis of the Element Method*, Prentice Hall, 1973.

[308] A. H. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice Hall, 1971.

[309] A. H. Stroud, *Numerical Quadrature and Solution of Ordinary Differential Equations*, Springer, 1974.

[310] M. Struve, *Variational Methods*, Springer, 1990.

[311] J. Szabados, P. Vertesi, *Interpolation of Functions*, World Scientific, 1990.

[312] Z. Szmydt, *Fourier Transformation and Linear Differential Equations*, Reidel, 1977.

[313] M. Taylor, J. E. Marsden (ed.), *Partial differential equations*, 3 vol, Springer, 1997.

[314] R. Temam, *Problèmes mathématiques en plasticité*, Paris, Gauthiers-Villars, 1983.

[315] R. Temam, *Analyse numérique*, Paris, Presses Universitaires de France, 1970.

[316] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis*, North-Holland, 1977.

[317] R. P. Tewarson, *Sparse Matrices*, Academic Press, 1973.

[318] R. Theodor, *Initiation à l'analyse numérique*, CNAM, Masson, 1989.

[319] F. Thomasset, *Implementation of Finite Elements Methods for Navier-Stokes Equations,* Springer, 1981.

[320] J. Todd, *Basic Numerical Mathematics*, 2 vol., Birkhäuser, 1977.

[321] E. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics,* Springer, 1999.

[322] E. Tournier, *Computer Algebra and Differential Equations*, Academic Press, 1988.

[323] F. Treves, *Basic Linear Partial Differential Equations*, Academic Press, 1975.

[324] H. Triebel, *Interpolation Theory, Function Spaces, Ddifferential Operators*, North-Holland, 1978.

[325] S. Vandewolle, *Parallel Multigrid Waveform Relaxation for Parabolic Problems*, Teubner, Stuttgart, 1993.

[326] R. S. Varga, *Matrix Iteration Analysis*, Prentice Hall, 1962.

[327] R. S. Varga, *Functional Analysis and Approximation Theory in Numerical Analysis*, SIAM, 1971.

[328] F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, Springer, 1990.

[329] R. Vichnevetsky, *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*, SIAM, 1982.

[330] J. Villadsen, M. L. Michelsen, *Solution of Differential Equation Models by Polynomial Approximation*, Prentice Hall, 1978.

[331] A. M. Vinogradov, *Symmetries of Partial Differential Equations*, Kluwer, 1989.

[332] R. Voigt, *Spectral Methods for Partial Differential Equations*, SIAM, 1984.

[333] G. Watson, *Approximation Theory and Numerical Methods*, Cambridge University Press, 1981.

[334] J. Weidmann, *Spectral Theory of Ordinary Differential Operators*, Springer, 1987.

[335] B. Wendroff, *Theoretical Numerical Analysis*, Academic Press, 1966.

[336] P. Wesseling, *An Introduction to Multigrid Methods*, John Wiley, 1992.

[337] J. R. Westlake, *A Handbook of Numerical Matrix Inversion and Solution of Linear Equations*, Wiley, 1968.

[338] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965.

[339] M. W. Wong, *An Introduction to Pseudo-differential Operators*, World Scientific, 1991.

[340] V. A. Yakubovitch, *Linear Differential Equations with Periodic Coefficients*, John Wiley, 1975.

[341] H. Yoshiyuki, *Functional Differential Equations with Finite Delay*, Springer, 1991.

[342] K. Yosida, *Functional Analysis*, Springer Verlag, 1965.

[343] K. Yosida, *Equations différentielles et intégrales*, Paris, Dunod, 1971.

[344] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Pres, 1971.

[345] D. M. Young, R. T. Gregory, *A Survey of Numerical Mathematics*, 2 vols, , New-York, Dover, reprinted 1988.

[346] M. Zamansky, *Approximation des fonctions*, Hermann, 1985.

[347] O. C. Zienkiewicz, *La Méthode des éléments finis*, Pluralis éditions, 1973.

[348] D. Zill, *Differential Equations with Boundary Value Problems*, PWS-Kent Pub., 1989.

[349] D. Zwillinger, *Handbook of Differential Equations*, Academic Press, 1989.

Index

Achevé d'imprimer sur les presses de l'Imprimerie BARNÉOUD
B.P. 44 - 53960 BONCHAMP-LÈS-LAVAL
Dépôt légal : Juin 2005 - N° d'imprimeur : 505.097
Imprimé en France