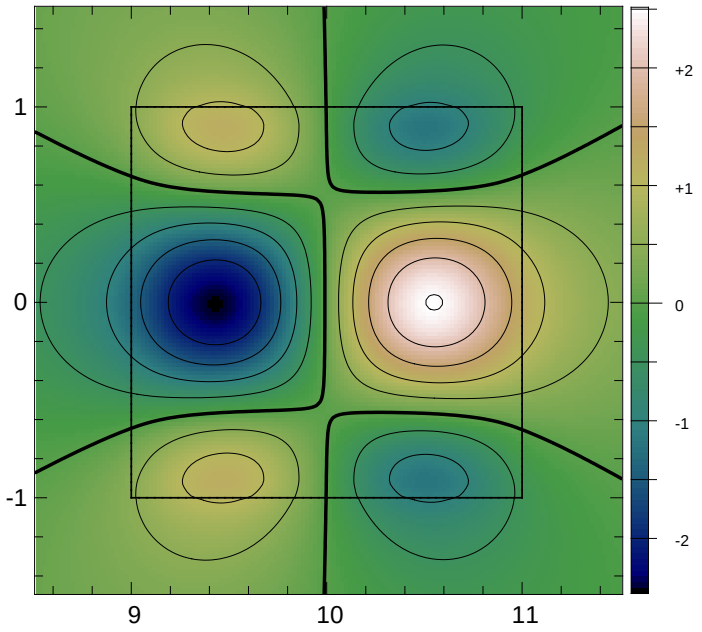
DEA Astrophysique & Méthodes associées — 2002/2003
Ecole doctorale d'Ile-de-France
Universités Paris 7 et 11 — Observatoire de Paris

MÉTHODES NUMÉRIQUES

Élements d'un premier parcours

Jean-Marc Huré
LUTh/Observatoire de Paris-Meudon, et Université Paris 7 Denis Diderot
(Jean-Marc.Hure@obspm.fr)

Didier Pelat
LUTh/Observatoire de Paris-Meudon
(Didier.Pelat@obspm.fr)



J.-M. Huré
Maître de Conférences à l'Université Paris 7,

LUTh/Observatoire de Paris-Meudon, Section de Meudon,
Place Jules Janssen, F-92195 Meudon
et
Université Paris 7 Denis Diderot, Place Jussieu, F-75251 Paris Cedex 05

e-mail: Jean-Marc.Hure@obspm.fr
tél: (33) 01 45 07 75 14
fax: (33) 01 45 07 74 69

Didier Pelat
Astronome à l'Observatoire de Paris-Meudon

LUTh/Observatoire de Paris-Meudon, Section de Meudon,
Place Jules Janssen, F-92195 Meudon

e-mail: Didier.Pelat@obspm.fr
tél: (33) 01 45 07 74 37
fax: (33) 01 45 07 74 69

Illustration de couverture: composante verticale du champ électrostatique généré par un tore axi-symmétrique de section carrée, contenant selon sa section une double distribution sinusoïdale de charge. Le champ est calculé par intégration du noyau de Poisson avec traitement des singularités.

MÉTHODES NUMÉRIQUES

Élements d'un premier parcours

Version 2, Novembre 2002

Jean-Marc Huré

LUTh/Observatoire de Paris-Meudon, et Université Paris 7 Denis Diderot
(Jean-Marc.Hure@obspm.fr)

Didier Pelat

LUTh/Observatoire de Paris-Meudon
(Didier.Pelat@obspm.fr)

Mise en garde

Ce cours est dispensé depuis l'année universitaire 2001 aux étudiants des DEA d'Astrophysique de l'Ecole Doctorale d'Ile-de-France dans le cadre de l'option méthodologique intitulée "Méthodes numériques". Il aborde les thèmes suivants (chapitres 2 à 7 inclus): quelques généralités sur le calcul, la dérivation, les polynômes, l'interpolation et les ajustements, les quadratures, la recherche des zéros de fonctions et de systèmes non-linéaires et les équations aux dérivées ordinaires. Chaque chapitre contient des exercices et des problèmes qui permettront d'illustrer ou d'approfondir les notions rencontrées. Dans le cadre du DEA, ce cours est complété par un enseignement pratique couvrant une soixantaine d'heures et durant lequel les étudiants sont chargés de mettre en place un outil numérique fiable résolvant un problème astrophysique particulier. On trouvera au chapitre 8 quelques exemples de projets proposés.

Le contenu de ce manuscript ne prétend pas être exhaustif et couvrir l'ensemble des techniques relevant des méthodes numériques. Bien au contraire, il reste très succint et ne donne que quelques "élements d'un premier parcours". Il s'adresse essentiellement aux étudiants des deuxième et troisième cycles universitaires des disciplines scientifiques souhaitant se familiariser avec le vocabulaire et les méthodes de base, et ne nécessite aucune connaissance préalable. Pour ceux qui souhaitent aller plus loin, il existe de nombreux ouvrages spécialisés

Le calcul numérique est un domaine de recherche à lui seul. Une abondante littérature lui est dédiée (voir le chapitre 2 et les réfèrences bibliographiques): des livres, des revues périodiques et des bibliothèques de programmes. Comme dans de nombreux domaines, la pratique joue un rôle essentiel dans la maîtrise du sujet. On ne peut raisonnablement pas espérer assimiler les méthodes sans les manipuler sur des exemples concrets et connus.

Le calcul numérique est aussi un art difficile. On peut s'y perdre et oublier qu'il n'est pour nous qu'un outil. Beaucoup ont tendance à vouloir tout réinventer et perdent beaucoup de temps à ré-écrire, souvent mal, les méthodes standards. Si cette démarche est légitime voire nécessaire pour un apprentissage, on ne doit pas en faire une habitude, au risque de devenir Numéricien; c'est un choix. Pour le Physicien, il est essentiel de savoir estimer la place réelle du calcul numérique dans la modélisation et de trouver un juste équilibre. Ainsi, il ne faudra pas hésiter à utiliser des outils déjà existants, mis au point par des professionnels. Afin de bien manipuler les méthodes et les routines que vous pourrez trouver ça-et-là et que vous implémenterez dans votre programme et celles, plus spécifiques, que vous produirez vous-même, il est fondamental de connaître un minimum de concepts. C'est l'objectif de ce cours introductif. Ceci pourra d'une part vous aider à faire des choix judicieux dans la consultation d'arbres de décision et d'autre part, vous permettre de mieux faire interagir les méthodes entre-elles. En tous cas, vous ne pourrez qu'améliorer la qualité de votre programme, donc de votre modèle physique et aussi mieux le comprendre.

J.-M. Huré et D. Pelat

Table des matières

Chapitre 1

Un peu d'histoire ...

D'après les historiens, le calcul numérique remonte au moins au troisième millénaire avant notre ère. Il est à l'origine favorisé par le besoin d'effectuer des mesures dans différents domaines de la vie courante, notamment en agriculture, commerce, architecture, géographie et navigation ainsi qu'en astronomie. Il semble que les Babyloniens (qui peuplaient l'actuelle Syrie/Iraq) sont parmi les premiers à réaliser des calculs algébriques et géométriques alliant complexité et haute précision. Surtout, ils donnent une importance et un sens au placement relatif des chiffres constituant un nombre, c'est-à-dire à introduire la notion de *base* de dénombrement, en l'occurrence, la base *sexagésimale* que nous avons fini par adopter dans certains domaines. Ils se distinguent ainsi d'autres civilisations, même bien plus récentes, qui développent des méthodes plus lourdes, en introduisant une pléthore de symboles. Il y a environ 3500 ans, les populations de la vallée de l'Indus (regions de l'Inde et du Pakistan) introduisent les notions de zéro et emploient les nombres négatifs. Il adaptent également le système de comptage Babylonien au système *décimal* qui est le nôtre aujourd'hui. Ces premiers outils de calcul sont largement développé par la suite par les Grecs, puis transmis en Europe par l'intermédiaire des civilisations mulsumanes peuplant le bassin méditérannéen.

Le calcul numérique tel que nous le concevons pratiquement aujourd'hui connaît son premier veritable essor à partir du XVIIeme siècle avec les progrès fulgurants des Mathématiques et de la Physique, plus ou moins liés aux observations et aux calculs astronomiques. Plusieurs machines de calcul sont en effet construites, comme la "Pascaline" inventée par B. Pascal en 1643, la DENO ("Difference Engine Number One"; voir la figure 1.2) de C. Babbage en 1834 mais qui fonctionnait mal, ou encore le tabulateur de H. Hollerith spécialement conçu pour recenser la population américaine, vers 1890. Il s'agit bien-entendu de machines mécaniques imposantes et d'utilisation assez limitée. Le manque de moyens de calcul performants limite en fait l'expansion et la validation de certaines théories du début du XXeme siècle. Ce fut le cas en particulier de la théorie de la Relativité Générale due à A. Einstein.

La Seconde Guerre Mondiale et les progrès technologiques qu'elle engendre va permettre au calcul numérique d'amorcer un second envol. Les anglais mettent au point le premier ordinateur en 1939, Colossus, dont la mission est de décripter les messages codés envoyés par l'émetteur Enigma de l'Allemagne nazie. Cette machine introduit les concepts révolutionnaires émis par A. Tur-

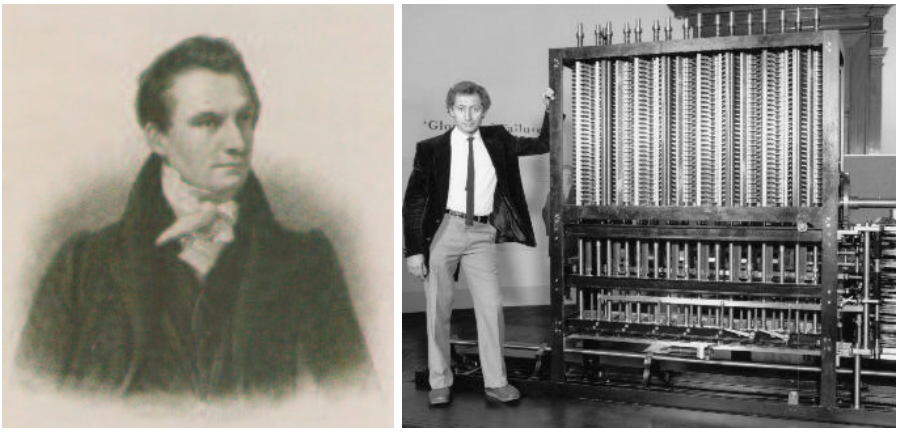


Figure 1.1: Charles Babbage (*à gauche*), l'une des grandes figures de l'histoire de l'ordinateur. Il fut l'inventeur de la DENO ("Difference Engine Number One") dont une copie est exposée dans un musée de Londres (*à droite*).

ing dans les années 1936 concernant l'automatisation des calculs. Les calculateurs sont désormais entièrement électroniques. Autre machine qui fait date dans l'histoire, le ENIAC ("Electronic Numerical Integrator And Computer) construit en 1946. Malheureusement, ce type de machine ne dispose pas de mémoire interne et doit être en permanence reprogrammée.

A la fin des années 1940, un certain J. von Neumann (voir la figure 1.3) repense l'architecture des ordinateurs et introduit, entre autres, les mémoires permettant de sauvegarder les programmes, et les concepts de *hardware* (matériel) et de *software* (logiciel). La première machine de calcul incluant les concepts de von Neumann (et ceux de Turing) est ainsi produite par la firme américaine IBM; elle s'appelle MARK I et pèse 5 tonnes. Les premières applications concernent tous les domaines scientiques et techniques. Le FORTRAN I, un langage de programmation destiné aux scientifiques, est conçu dès 1954 ... mais il lui manque un vrai compilateur.

Vers la fin des années 1960, l'apparition progressive des transitors et de leur assemblage massif sur des surfaces de plus en plus réduites augmente considérablement les performance des machines et permet des simulations numériques de réalisme croissant. Cet effort de miniaturisation est d'ailleurs imposé par la course à la conquête de l'espace. Apparaissent ainsi en 1970 les fameux microprocesseurs mis au point par les firmes INTEL et MOTOROLA qui équipent la majeure partie des sondes spatiales de l'époque. Le calcul numérique devient rapidement une science à part entière. Les années 70 marquent aussi le tournant pour les langages de programmation: certains sont définitivement produits à des fins scientifiques, alors que d'autres seront pensés pour la gestion, comme le COBOL. Au début des années 1980, l'ordinateur le plus puissant du monde s'appelle CRAY I (voir la figure 1.3). Sa forme est spécialement choisie pour optimiser la rapidité des calculs. C'est aussi le début de l'informatique familiale avec la mise sur le marché des PERSONAL COMPUTERS d'IBM.

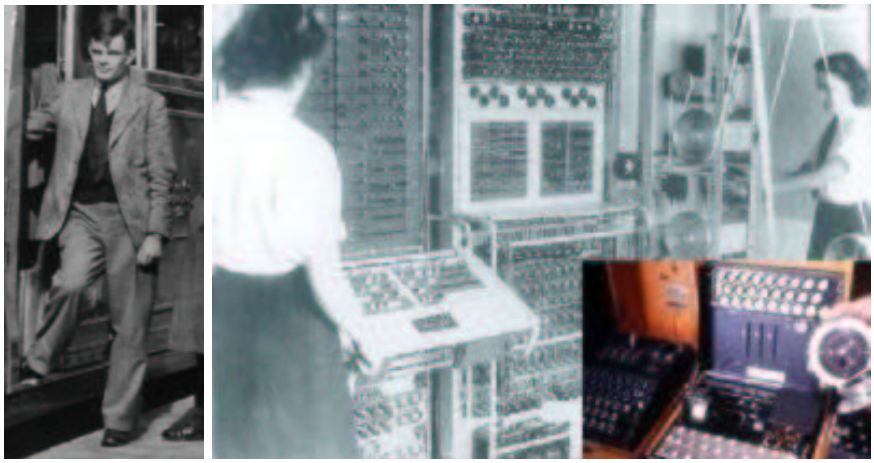


Figure 1.2: Alan Turing (*à gauche*) donnera aux premiers ordinateurs les moyens de "penser" et de travailler de manière autonome. Sa théorie sera l'une des pièces maitresses de Colossus (*à droite*), le premier calculateur mis au point par les anglais à l'aube de la Seconde Guerre Mondiale pour décrypter les messages secrets émis par la machine Enigma (*en médaillon*) des Nazis.

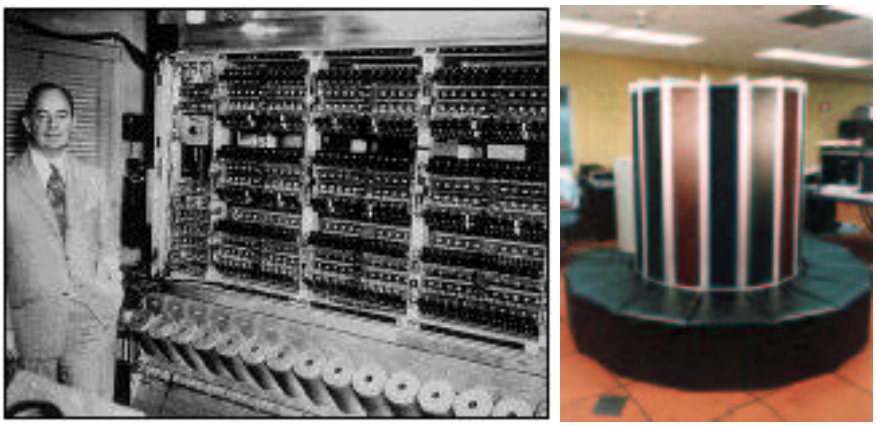


Figure 1.3: John L. von Neumann (*à gauche*), enfant prodige de l"'informatique" de l'après-guerre. Le Cray I (*à droite*) construit dans les années 1980 dispose d'une puissance de calcul de 160 millions d'opérations à la seconde. Sa conception en forme de cylindre devaient assurer une transmission optimale de l'information.

En une quinzaine d'années, la rapidité des calculateurs a été multipliée par plus de 10000. La vitesse d'exécution des opérations élémentaires se compte maintenant en dizaines de millions de millions d'opérations à la seconde (ou dizaines de *téra-flops*, à comparer à la centaine de *méga-flops* du CRAY I). Les capacités de stockage ont gagné 7 ordres de grandeur au moins. Aujourd'hui, toutes ces performances doublent tous les ans. Pour le monde scientifique, celui de la Recherche Fondamentale et de l'Industrie, les calculateurs et le développement de techniques de programmation spécifiques (comme la *programmation parallèle*) sont devenus des outils incontournables à la connaissance et ouvrent de nouveaux horizons pour la modélisation et la compréhension des phénomènes complexes et la mise au point de nouvelles technologies.

Chapitre 2

Généralités

2.1 Objectifs

On regroupe sous le terme générique de "méthodes numériques", toutes les techniques de calcul qui permettent de résoudre de manière exacte ou, le plus souvent, de manière approchée un problème donné. Le concept de calcul est assez vaste et doit être pris au sens large. Il peut s'agir de déterminer l'inconnue d'une équation, de calculer la valeur d'une fonction en un point ou sur un intervalle, d'intégrer une fonction, d'inverser une matrice, etc. Bien que la mise en équation d'un problème et sa résolution passent naturellement par les Mathématiques, les problématiques sous-jacentes concernent des disciplines aussi variées que la Physique, l'Astrophysique, la Biologie, la Médecine, l'Economie, etc. Il existe ainsi une grande variété de problèmes possibles avec pour chacun d'eux, des méthodes très spécifiques. De fait, le nombre total de méthodes numériques dont nous disposons à l'heure actuelle est vraisemblablement gigantesque.

Une méthode numérique met en oeuvre une certaine procédure, une suite d'opérations, généralement en très grand nombre, que l'on transcrira ensuite dans un langage de programmation. Bien qu'une méthode numérique puisse s'effectuer mentalement (du moins avec un crayon et un papier) comme inverser une matrice 2×2, résoudre $\tan x - 1 = 0$, ou calculer $\sqrt{2}$, elle nécessite dans la majorité des cas un ordinateur qui a l'avantage de la rapidité (mais pas de la précision). Il convient à ce niveau de bien différencier la partie méthode numérique, souvent indépendante du calculateur et du langage, et la partie programmation qui met en oeuvre d'une part l'algorithme et d'autre part une suite d'instructions écrites dans un langage de programmation. Bien-sûr, une méthode numérique pourra dépendre de l'architecture d'un ordinateur et du langage utilisé. Toutefois, l'un des soucis majeurs de l'utilisateur et du programmeur est d'assurer à son programme une certaine *portabilité*, c'est-à-dire de pouvoir l'exécuter sur des machines différentes sans avoir besoin d'adapations (trop) spécifiques.

Les méthodes numériques sont indispensables à la réalisation de programmes de calculs ou *codes de calcul*. En particulier, pour les astrophysiciens qui ne bénéficient pas d'un laboratoire permettant de valider leurs théories à partir d'expériences renouvelables à loisir et contrôlables, ces outils sont le seul moyen de simuler ou de modéliser les phénomènes que nous observons, de les interpréter et de les comprendre. Rappelons que les méthodes numériques sont en effet

présentent dans toutes les disciplines de l'Astrophysique moderne: la cosmologie, l'instrumentation, le traitement de données, la planétologie, la physique solaire, la physique des galaxies, la physique extragalactique, etc. S'il est vrai qu'il existe une très grande diversité de méthodes numériques avec lesquelles ont peut, en pratique, quasiment tout faire, certains problèmes (par exemple en traitement du signal, en mécanique céleste ou en mécanique des fluides) ont nécessité la mise au point de méthodes très spécifiques.

2.2 Philosophie

L'état d'esprit dans lequel travaille le "numéricien" est davantage celui que l'on connaît en Physique, avec ses multiples approximations (parfois difficilement justifiables) et hypothèses simplificatrices, plutôt que celui que l'on doit développer en Mathématiques. Une certaine rigueur est toutefois indispensable. Ainsi serons-nous le plus souvent amenés à changer assez significativement le problème initial, à le simplifier dans une certaine mesure, à le rendre linéaire, plus académique aussi, condition *sine qua non* au traitement numériquement. Une étape majeure sera alors de critiquer les solutions obtenues et de les interpreter par rapport au problème initialement posé. Une bonne illustration est le pendule pesant de longueur l dont l'équation du mouvement dans un champ de pesanteur g est

$$l^2\ddot{\theta} - lg\sin\theta = 0 \tag{2.1}$$

L'harmonicité tant recherché n'est pas réelle; elle n'est que le pur produit d'une série d'approximations, notamment celle des "petits angles" qui permet de poser $\sin\theta \sim \theta$. L'équation précédente devient alors

$$l^2\ddot{\theta} - lg\theta = 0 \tag{2.2}$$

et l'on conçoit facilement que ses solutions soient différentes.

Numériquement par contre, il sera possible de décrire assez naturellement son mouvement an-harmonique quelque soit l'amplitude initiale de son mouvement, soit par un développement limité de la fonction sinus à l'ordre 2 ou plus, soit par integration directe de l'équation différentielle (2.1). Mais dans les deux cas, la solution obtenue, aussi précise soit-elle, ne sera qu'approchée, soit en raison de la troncature exercée sur le développement de Taylor, soit en raison des erreurs incontournables faites en intégrant l'équation différentielle ci-dessus.

2.3 Précision et temps de calcul

Deux aspects fondamentaux sous-tendent l'utilisation ou la mise en place de toute méthode numérique: la précision souhaitée (ou disponible) d'une part, et le temps de calcul d'autre part. Disposer de beaucoup de précision est confortable voire indispensable lorsque l'on étudie certains phénomèmes, comme les systèmes chaotiques dont les propriétés affichent une forte sensibilité aux conditions initiales et à la précision. Mais lorsque l'on est trop "gourmand" sur cet aspect, le temps de calcul se trouve accru, parfois de manière prohibitive. Il n'est pas inutile de garder à l'esprit la *relation d'incertitude de Numerics*

$$\Delta t \times \epsilon \gtrsim \hbar \tag{2.3}$$

où Δt exprime un temps de calcul (par exemple 13 min), ϵ la précision relative de la méthode (par exemple 10^{-4}; voir ci-dessous) et $\hbar$ est une constante pour

la méthode considérée. Cette relation rappelle l'impossibilité d'allier précision et rapidité.

2.3.1 Précision

La précision est limitée par les machines. Inévitablement, elle se trouve détériorée au fil des calculs. Sans entrez dans le détail, il faut savoir qu'un ordinateur fonctionne en effet avec un nombre limité de chiffres significatifs, au maximum 14, 15 voire 16 pour les plus performants, ces chiffres étant rangés en mémoire dans des boites virtuelles ou *bits*, après transcription en *base 2*. Certains bits sont réservés pour affecter un *signe* (+ ou −) au nombre et d'autres pour l'*exposant*. En ce sens, les machines classiques ne pourront généralement pas distinguer `3.1415926535897931` de `3.1415926535897932`, à moins d'un traitement spécifique de la *mantisse* à la charge de l'utilisateur. Dans la mémoire, seul les premiers chiffres d'un nombre seront enregistrés (voir le tableau 2.1), les autres seront définitivement perdus. La plupart des langages permettent de choisir le nombre de décimales au niveau de la procédure déclarative, comme les anciennes instructions `REAL*4` ou `REAL*8` du langage FORTRAN 77, dont les équivalents sont respectivement `REAL(KIND=1)` et `REAL(KIND=2)` en FORTRAN 90.

bits	minimum (positif, non nul)	maximum	précision
32	`2.938736E-39`	`1.701412E+38`	$\sim 10^{-7}$
48	`2.9387358771E-39`	`1.7014118346E+38`	$\sim 10^{-12}$
64	`5.562684646268003E-309`	`8.988465674311580E+308`	$\sim 10^{-16}$

Table 2.1: Mimimum et maximum réels adressables et précision relative possible en fonction du nombre de bits de la machine.

2.3.2 Temps de calcul

Le temps de calcul est limité par la capacité des ordinateurs, par la durée de vie du matériel, par des facteurs extérieurs (comme les coupures d'électricité), et par l'utilisation que l'on souhaite classiquement faire des programmes de simulations. La plupart du temps, il s'agit d'une utilisation intensive où l'on balaye l'espace des paramètres d'un modèle. Ceci impose de rechercher toujours les méthodes les plus rapides. Il est de plus en plus fréquent de travailler sur des problèmes complexes dont la résolution nécessite des heures, des jours, parfois même des semaines de calculs sans interruption. Le choix s'orientera alors vers une méthode numérique rapide. Mais la rapidité se fera forcément au détriment de la précision (voir Eq.(2.3)). Quand les calculs sont intrinsèquement rapides et que le temps de calcul n'est pas un facteur déterminant, on aura tout loisir de (et peut-être intérêt à) choisir une méthode plus lente mais plus précise. Notons également un fait souvent oublié: les accès aux périphériques sont très couteux en temps, en particulier les accès aux disques durs et les impressions à l'écran.

Pour une méthode numérique donnée, on peut toujours estimer le temps de calcul en comptant le nombre total $\mathcal{N}$ d'opérations qu'effectue la machine. On peut ensuite éventuellement multiplier ce nombre par le temps de calcul τ_{e} relatif à une opération élémentaire, mais cela n'a de valeur qu'à titre comparatif

(car ce temps τ_e varie d'une machine à l'autre et dépend aussi de la nature des opérations). Par exemple, pour calculer la force d'interaction gravitationnelle subie par une particule de masse m_i sous l'influence d'une particule de masse m_j située à la distance $\vec{r}_{ij} = x_{ij}\vec{u}_x + y_{ij}\vec{u}_y$, soit

$$\vec{F} = -\frac{m_i m_j}{r_{ij}^3}\vec{r}_{ij} \tag{2.4}$$

il faut effectuer $\mathcal{N} = 6 \times \frac{1}{2}(N-1)N$ opérations élémentaires s'il y a N particules au total, soit un temps de calcul égal à $3(N-1)N\tau_e$. En effet, le schéma associé à la relation (2.4) pourra être transcris sous la forme

$$\begin{cases} \mathtt{a} \leftarrow -\mathtt{m(i)/r(i,j)} * \mathtt{m(j)/r(i,j)/r(i,j)} \\ \mathtt{Fx(i,j)} \leftarrow \mathtt{a} * \mathtt{x(i,j)} \\ \mathtt{Fy(i,j)} \leftarrow \mathtt{a} * \mathtt{y(i,j)} \end{cases}$$

et met en jeu 3 multiplications et 3 divisions. On dira alors que l'*ordre de la méthode* est $\sim (N-1) \times N$, soit $\sim N^2$ lorsque N est grand. Nous voyons que sur une machine parallèle qui pourra traiter simultanément l'interaction de chaque particule (soit une machine dotée de N processeurs indépendants), l'ordre est très inférieur: il varie comme $(N-1)$. C'est donc un gain de temps colossal pour N grand.

2.4 Les erreurs

Outre les erreurs de programmation (celles que le compilateur détectera et que vous corrigerez, et celles que vous ne trouverez jamais), il existe trois sources d'erreur qu'il convient d'avoir présent à l'esprit: l'erreur de schéma, l'erreur de représentation et l'erreur par perte d'information.

2.4.1 Le schéma

L'erreur de schéma est générée lorsque l'on remplace une relation "exacte" par une autre, plus simple ou plus facilement manipulable. C'est par exemple le cas d'une série de Taylor tronquée (on parle ici plus précisément d'*erreur de troncature*), comme

$$e^x \approx 1 + x^2, \tag{2.5}$$

ou le cas de l'approximation d'une dérivée par une *différence finie* (comme nous le verrons au chapitre suivant) comme

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \tag{2.6}$$

ou encore de l'estimation d'une *quadrature*

$$\int_a^b f(x)dx \approx f(a)(b-a) \tag{2.7}$$

L'erreur de schéma est généralement l'erreur dominante. Elle est en principe incontournable car toutes les méthodes numériques sont basées sur des schémas. Il est important qu'elle soit connue ou appréciable. Notez que tous les schémas ne génèrent pas les mêmes erreurs. On choisira donc, quand c'est possible, les schémas les plus précis.

2.4.2 La représentation machine

L'erreur de représentation est liée aux modes de représentation, de stockage et de calcul des ordinateurs. Par exemple, supposons que l'on veuille effectuer l'opération $\frac{1}{3} - \left(\frac{2}{3} - \frac{1}{3}\right)$ sur une machine fonctionnant avec 4 chiffres significatifs. En interne, la machine effectuera d'abord $\frac{2}{3} - \frac{1}{3}$, ce qui pourra donner

$$\texttt{0.6667} - \texttt{0.3333} = \texttt{0.3334}$$

Ici, $\frac{2}{3}$ a été remplacé par `0.6667`; c'est une erreur de représentation (erreur d'arrondi). Reste à effectuer $\frac{1}{3} - 0.3334$, d'où le résultat faux (ou vrai à 10^{-4} près) : `+0.0001`. Certaines machines peuvent d'ailleurs fournir `0.6666` pour $\frac{2}{3} - \frac{1}{6}$ et finalement donner, par coïncidence, un résultat exact.

L'erreur de représentation est généralement négligeable devant les autres, sauf dans certains cas particuliers où, sans précaution, l'on manipule des nombres très différents ou très proches.

2.4.3 La perte d'information

La perte d'information est une consésquence de la représentation machine. Elle se produit lorsque l'on soustrait deux nombres très proches; le résultat est alors un nombre comportant peu de chiffres significatifs. Soit à effectuer $\frac{71}{500} - \frac{1}{7}$, toujours sur une machine travaillant avec 4 chiffres significatifs en mantisse. En pratique, la machine donne

$$\texttt{0.1420} - \texttt{0.1429} = -\texttt{0.0009}$$

contre $-\frac{3}{3500} \simeq -8.5714 \times 10^{-4}$. Ce résultat, compte-tenu de l'arrondi, est correct. Toutefois, il ne comporte plus qu'un seul chiffre significatif alors que les deux nombres initiaux en possédaient quatre.

L'erreur par perte d'information peut être dévastatrice (par exemple losque l'on veut calculer numériquement une dérivée). Souvent, on ne la soupçonne pas. Elle survient aussi lorsque l'on effectue beaucoup d'additions et soustractions à la queue leu-leu mettant en jeu des termes de même amplitude (c'est la cas par exemple du calcul de séries alternées).

2.5 Conditionnement et sensibilité

2.5.1 Propagation des erreurs

Dans toute méthode numérique, il y a au moins un paramètre permettant de régler (c'est-à-dire d'imposer) ou de contrôler le bon déroulement du processus et la précision du résultat. Il s'agit généralement d'un critère faisant appel à l'*erreur absolue* ou à l'*erreur relative* sur une quantité y par rapport à une quantité de référence y^*. Ces erreurs sont définies respectivement par

$$\Delta(y) = y - y^* \tag{2.8}$$

et par

$$\epsilon(y) = \frac{y - y^*}{y^*}, \qquad y^* \neq 0 \tag{2.9}$$

Un fait est incontournable: les erreurs, quelque soit leur origine, se propagent et s'amplifient. On peut le comprendre sur un exemple simple. Si une quantité x

est connue avec une précision relative $\epsilon(x) \ll 1$, alors son image par une fonction f est aussi entachée d'une erreur

$$f(x + \epsilon(x)x) \approx f(x) + \epsilon(x)xf'(x) \tag{2.10}$$

et l'erreur relative sur f vaut

$$\epsilon(f) = \frac{f(x + \epsilon(x)x) - f(x)}{f(x)} \approx \epsilon(x)x\frac{f'(x)}{f(x)} \tag{2.11}$$

On voit donc que si $f'(x)$ est important et $f(x)$ faible, à la fois $\Delta(f)$ et $\epsilon(f)$ peuvent être grands. Notez que l'on peut aussi écrire la relation précédente sous la forme

$$\epsilon(f) = \epsilon(x)\eta \tag{2.12}$$

où η est le *nombre de conditionnement* (voir ci-dessous).

Ajoutons qu'il faudrait en toute rigueur être capable de mettre une barre d'erreur sur les résultats issus d'une méthode ou plus généralement d'une simulation, en tenant compte de toutes les sources d'erreur possibles. Il est navrant de constater que la notion de barre d'erreur ne semble concerner que les physiciens expérimentateurs. C'est un tort, car une simulation numérique est une expérience comme une autre...

2.5.2 Nombre de conditionnement

Une méthode numérique travaille généralement avec un certain nombre de paramètres, les *paramètres d'entrée*, et renvoie des résultats, que l'on peut assimiler à des *paramètres de sortie*. Un objectif (souvent difficile à atteindre) que recherchera le Physicien est de produire des résultats qui ne dépendent pas (ou peu) de la méthode choisie. L'aspect qui guidera son choix sur une méthode plutôt qu'un autre est la *stabilité*, c'est-à-dire une certaine garantie que la procédure n'aura pas la facheuse propriété d'amplifier et/ou de générer sans limites les erreurs.

La sensibilté (ou la stabilité) peut être quantifée grâce au *nombre de conditionnement* (ou nombre de condition). Prenons le cas où d'une méthode qui, pour une quantité x donnée fournit une quantité y. On définit ce nombre pour deux couples (x, y) et (x', y') par

$$\eta = \frac{\frac{y'-y}{y}}{\frac{x'-x}{x}} \equiv \eta(x) \tag{2.13}$$

Ce nombre rend compte du *pouvoir amplificateur* d'une méthode. Ainsi une méthode numérique est très sensible (ou instable) si $\eta \gg 1$, peu sensible (stable) si $\eta \lesssim 1$ et est insensible si $\eta \ll 1$.

Notez que si x et x' son très proches, alors

$$\eta \sim \frac{xf'(x)}{f} \tag{2.14}$$

La sensibilté peut donc être importante lorsque $f \to 0$ et/ou $f' \to \infty$ et/ou $x \to \infty$.

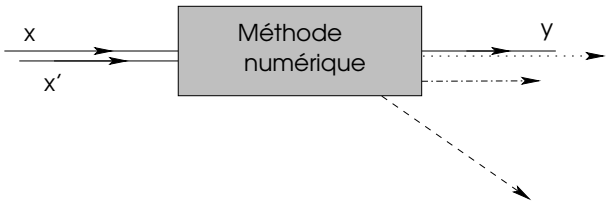


Figure 2.1: La sensibilté d'une méthode se mesure avec le nombre de conditionnement.

2.6 Du modèle physique au modèle numérique

Lorsque l'on met un système physique ou un modèle en équations et que l'on aborde l'étape numérique, plusieurs questions simples mais indispensables doivent être posées . . . et résolues. Parmi ces questions/réponses, qui peuvent plus ou moins conditionner le choix d'une méthode particulière, on trouve: mon voisin de palier aurait-il déjà travaillé sur un problème similaire ? les inconnues de problème physique sont-elles bien identifiées ? quel est le domaine de variation des variables (temps, espace, . . .) ? les variables physiques sont-elles aussi les bonnes variables numériques ? quelle est précision souhaitée sur la solution cherchée ? quelle est l'échelle de temps d'éxécution de mon programme ? la précision est-elle déterminante ? la temps de calcul est-il un facteur critique ? etc. Deux aspects généralement absents des manuels (ou implicitement évoqués) mais qui constituent une étape-clé dans le passage du modèle physique au modèle numérique sont l'adimensionnement des équations et le choix des variables numériques.

2.6.1 Adimensionnement des équations

L'adimensionnement des équations est une étape qui permet de réduire la *dynamique* des variables numériques. Il est en effet préférable de travailler sur un domaine numérique raisonnablement restreint et de manipuler des nombre de l'ordre de l'unité. Toutefois, une dynamique trop faible peut engendrer une perte d'information. Cette procédure vise donc à transformer les équations du problème en des équations souvent plus "lisibles" d'un point de vue mathématique (et donc numérique) où les nouvelles variables (les variables adimensionnées) apparaîssent alors comme des corrections. Comme sous-produit de l'adimensionnement, on obtiendra naturellement des équations les *échelles caractéristiques* du problème (par exemple une ou des échelles de longueur, des échelles de vitesses, des échelles de temps, etc.).

2.6.2 Choix des variables

Concernant le choix des variables numériques maintenant, il faut rappeler que, dans beaucoup de problèmes, les variables possèdent des dynamiques complèment différentes. Même adimensionnées, des variables peuvent beaucoup varier. Ces variables obéissent forcément à une certaine loi en fonction des paramètres ou d'autres variables et il peut être astucieux d'inclure ces lois dans les variables numériques. Cela rendra la partie numérique beaucoup stable.

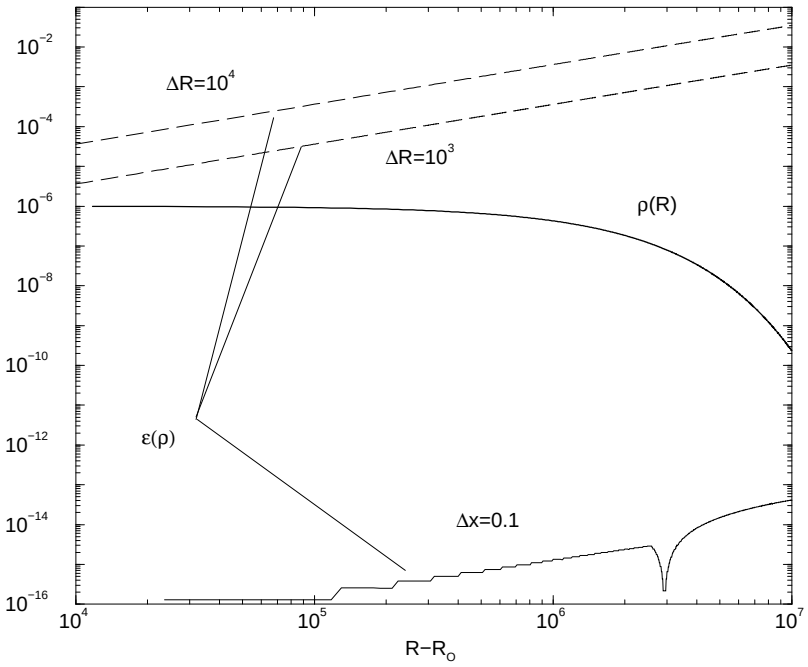


Figure 2.2: Illustration des effets bénéfiques d'adimensionnement et de choix des variables numériques. Solution analytique $\rho(R)$ de l'équation (2.16) (*gras*) et erreurs relatives sur la solution en intégrant la relation (2.16) par la méthode d'Euler, en variables (R, ρ) pour deux pas d'intégrations différents (*pointillés*) et par la même méthode mais en variables (x, f). Dans ce dernier cas, l'erreur (proche de la précision machine) est essentiellement due à des erreurs de représentation et à leur propagation.

2.6.3 Exemple

Prenons un exemple concret qui résumera l'ensemble des notions evoquées ci-dessus. Considerons l'équation qui régit l'équilibre hydrostatique à une dimension d'une atmosphère statique et sans masse entourant un corps solide de rayon $R_\oplus$. En prenant en compte le changement de gravité avec l'altitude R, cette équation s'écrit

$$\frac{\gamma c_s^2}{\rho}\frac{d\rho}{dR} = -g_\oplus \frac{R_\oplus^2}{R^2} \tag{2.15}$$

et supposons que l'on connaisse la densité ρ à la base de l'atmosphère $R = R_\oplus$, soit $\rho(R_\oplus)$. On peut évidemment choisir ρ et R comme variables numériques, mais ce choix n'est pas très judicieux. On doit d'abord adimensionner l'équation (2.16) en posant par exemple

$$\begin{cases} r = \dfrac{R}{R_\oplus} \\ \tilde{\rho} = \dfrac{\rho}{\rho(R_\oplus)} \end{cases}$$

d'où

$$\frac{\gamma c_s^2}{\tilde{\rho}}\frac{d\tilde{\rho}}{dr} = -g_\oplus \frac{1}{r^2} \tag{2.16}$$

C'est un bon début. Mais cela ne change pas le fait que ρ varie exponentiellement avec l'inverse de l'altitude et qu'il reste des termes de grande amplitude c_s^2 et $g_\oplus$, sans oublier r et surtour $\tilde{\rho}$ qui peut varier de plusieurs ordres de grandeurs sur le domaine d'intégration choise. Afin de réduire au maximum les risques d'erreurs numériques, il faut réduire la dynamique de ces variables et de l'équation par un choix judicieux de variables avec lesquelles le programme travaillera finalement. Pour cela, on peut poser

$$\begin{cases} f = \ln \rho \\ x = \dfrac{L}{r} \end{cases}$$

où $L\gamma c_s^2 R_\oplus^2 = GM_\oplus$ (où $g_\oplus = GM_\oplus / R_\oplus^2$). L'équation (2.16) devient alors

$$\frac{df}{dx} = 1 \tag{2.17}$$

Cette forme présente deux avantages: d'une part, elle est d'une extrême simplicité et d'autre part, elle met en évidence l'existence d'une échelle caractéristique de la densité, la longueur L. La figure 2.2 montre l'erreur produite sur la solution exacte $\rho(R)$ en variables (R, ρ) et en variables (x, f); le résultat est plutôt convainquant.

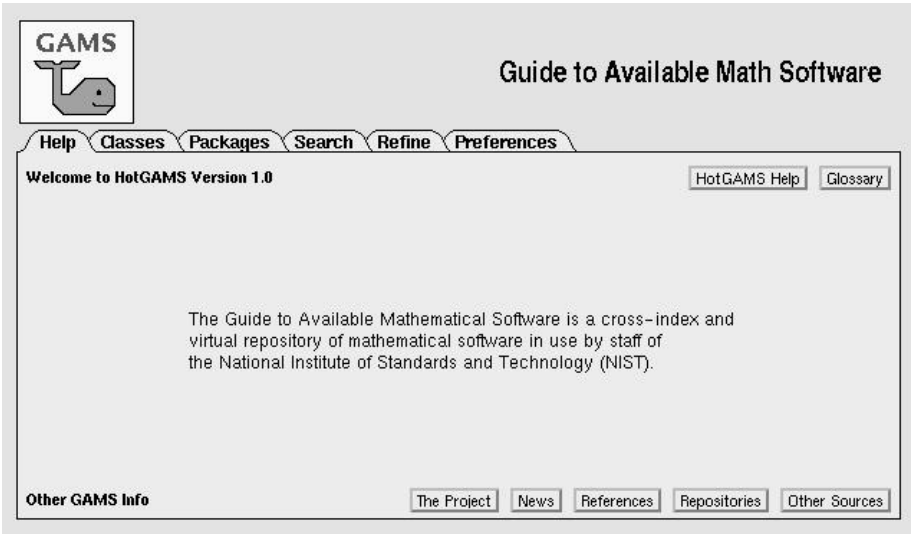


Figure 2.3: Page d'accueil du site Internet du GAMS, permettant l'accès à un large catalogue de routines en majorité disponibles en ligne.

2.7 Bibliothèques numériques

Grâce aux capacités de stockage dont nous disposons à l'heure actuelle, il a été possible de construire de véritables bibliothèques virtuelles réunissant un grand nombre de programmes ou *routines* de base que l'on peut maintenant facilement consulter via Internet. Ces routines sont généralement écrites en FORTRAN 77, FORTRAN 90 et C. Dans la majorité des cas, elles sont écrites et maintenues par

classe	contenu
A	Arithmetic, error analysis
B	Number theory
C	Elementary and special functions (search also class L5)
D	Linear Algebra
E	Interpolation
F	Solution of nonlinear equations
G	Optimization (search also classes K, L8)
H	Differentiation, integration
I	Differential and integral equations
J	Integral transforms
K	Approximation (search also class L8)
L	Statistics, probability
M	Simulation, stochastic modeling (search also classes L6 and L10)
N	Data handling (search also class L2)
O	Symbolic computation
P	Computational geometry (search also classes G and Q)
Q	Graphics (search also class L3)
R	Service routines
S	Software development tools
Z	Other

Table 2.2: Les 20 entrées principales de l'arbre de décision du "Guide to Available Mathematical Software (GAMS) project".

des numériciens et mathématiciens. L'utilisation de ces routines est à la responsabilité de l'utilisateur. Il sera nécessaire d'être vigilant et de garder un esprit critique quant aux résultats que vous obtiendrez avec ces routines. Toutefois, leur utilisation intensive par l'ensemble de la communauté scientifique offre une réelle garantie de fiabilité. En bref, mefiez vous plutôt des routines que vous écrirez vous-même et qui n'aurons été testées que, au mieux, par vous même !

Un site Internet donnant accès à ces routines est le serveur GAMS ("Guide to Available Mathematical Software") du NIST ("National Institue for Standard Technology") que l'on peut consulter à l'adresse `http://gams.nist.gov/`. Le GAMS est un projet de service public dirigé par R.F. Boisvert (on trouvera en appendice un extrait des motivations du projet). Il répertorie quelques 8800 modules de calcul répartis à travers 111 bibliothèques ou *packages*. Certaines bibliothèques de programmes sont malheureusement payantes (c'est-à-dire qu'il vous faudra acheter les routines précompilées, souvent contre quelques centaines d'euros). D'autres (une majorité) appartiennent au domaine public et peuvent être téléchargées en quelques secondes, avec un guide succint d'implémentation. L'une des force de ce serveur est qu'il offre en ligne, grâce à son arbre de décision, un moyen rapide de sélectionner la ou les routines que l'on cherche. Les 20 entrées principales de cet arbre sont données dans le tableau 2.2. L'utilisation de ces routines numériques nécessite un minimum d'investissement et de précaution, c'est pourquoi on prendra toujours soin d'effectuer un certain nombre de tests préalables, sur des exemples connus, avant de se lancer dans l'inconnu ...

Il existe aussi les (très nombreux) sites associés aux laboratoires de recherche

en Mathématiques Appliquées, Analyse Numérique, Physique, etc. qui souvent rendent publiques leurs programmes.

2.8 Exercices et problèmes

■ Calculez $\sum_{i=1}^{N} i^3$ et $\sum_{i=N}^{1} i^3$ pour $N = 1000$ et 10^6. Comparez.

■ Calculez $\sum_{i=1}^{N} \frac{1}{i(i+1)}$ pour $N = 99$, 999 et 9999 de 1 à N, puis à rebours. Dans les deux cas, comparez à la valeur exacte.

■ Développez e^x en série entières à l'ordre N et calculez la série. Comparez à la valeur exacte.

■ Comment calculez précisément, sur une machine dotée de 4 chiffres significatifs, les deux racines de l'équation $x^2 + 100x + 1 = 0$?

■ Quel est le nombre de conditionnement de la méthode qui consiste à remplacer $\sin x$ par $x - \frac{x^3}{3!}$, au voisinage de 0 ? de $\frac{\pi}{2}$. Même question avec le développement de Taylor de la fonction $\tan x$.

Chapitre 3

Dérivation

3.1 Rappels

Calculer en un point la dérivée f' d'une fonction f de la variable x dont on connaît une expression analytique ne pose aucune difficulté particulière si l'on admet quelques règles mathématiques simples. On peut même envisager l'utilisation d'un logiciel ou de routines de *calcul symbolique* ou formel. Plus délicat est l'estimation de la dérivée d'une fonction non-analytique, par exemple une fonction définie par un ensemble de N points de mesure $\{(x_i, y_i = f(x_i))\}_N$, qu'il s'agisse de la dérivée en l'un des points x_i ou bien de la dérivée quelque part entre deux points de l'échantillon.

Rappelons qu'une dérivée est une information locale, définie par passage à la limite du taux d'accroissement de f en x_0, soit

$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h) - f(x_0)}{h} \tag{3.1}$$

Une estimation précise de cette quantité est généralement assez délicate et difficle à partir d'informations de nature discrète. Cela nécessite un échantillonage extrêmement poussé de la fonction, comme cela est illustré à la figure 3.1. Un tel échantillonage n'est pas toujours disponible.

3.2 Différences finies

3.2.1 Développement de Taylor

Le développement de Taylor est un outil fondamental de l'analyse et du calcul numérique. Pour une fonction f, il s'écrit au voisinage de x_0

$$f(x) = f(x_0) + \sum_1^\infty a_i (x - x_0)^i \tag{3.2}$$

$$= a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots + a_N(x - x_0)^N + \ldots \tag{3.3}$$

où les coefficients a_k de la série sont donnés par

$$a_k = \frac{f^{(k)}(x_0)}{k!}, \qquad k = 0, \infty \tag{3.4}$$

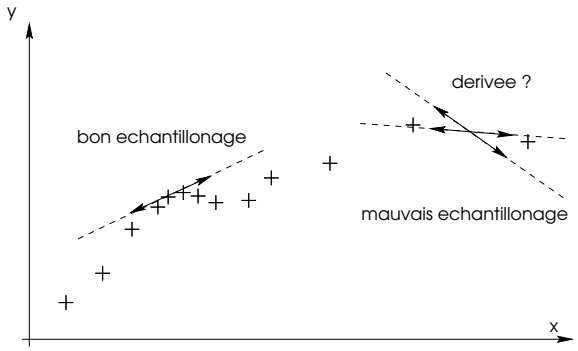


Figure 3.1: Dans le calcul numérique des dérivées, l'échantillonage de la fonction joue un rôle majeur.

et $f^{(k)}$ désigne la dérivée d'ordre k de f. Lorsque la série est tronquée à l'ordre N inclus, l'ordre de grandeur de l'erreur faite sur le développement est donnée par le terme $\frac{f^{(N+1)}(x_0)}{(N+1)!}(x-x_0)^{N+1}$ et l'on écrit alors

$$f(x) = a_0 + a_1(x-x_0) + \cdots + a_N(x-x_0)^N + \mathcal{O}\left((x-x_0)^{N+1}\right) \tag{3.5}$$

3.2.2 Différences excentrées

En utilisant le résultat ci-dessus, on peut essayer de transcrire l'expression (3.1). A l'ordre 1 inclus et en posant $x - x_0 = h$, le taux d'accroissement au point considéré s'écrit

$$\frac{f(x_0+h)-f(x_0)}{h} \approx \frac{f^{(1)}(x_0)}{1!} + \sum_2^\infty a_i h^{i-1} \tag{3.6}$$

Tronquée à l'ordre 1, cette relation donne

$$\frac{f(x_0+h)-f(x_0)}{h} \approx f'(x_0) + \mathcal{O}(h) \tag{3.7}$$

soit une représentation de la dérivée première de la fonction f, précise à l'ordre h. Il y a donc confusion (volontaire) entre la dérivée de la fonction et son taux d'accroissement. La relation (3.7) est une *différence finie à 2 points*, dite *en avant* ou FDF en anglais (pour "forward difference formula"), car elle fait appel à deux évaluations de la fonction, l'une au point considéré et l'autre un peu plus loin, en $x = x_0 + h$. On peut construire une différence finie à 2 points *en arrière* ou BDF en anglais (pour "backward difference formula"), en changeant h en $-h$ dans les relations précédentes, soit

$$\frac{f(x_0-h)-f(x_0)}{h} \approx \frac{f^{(1)}(x_0)}{1!} + \sum_2^\infty a_i (-h)^{i-1} \tag{3.8}$$

$$\approx f'(x_0) + \mathcal{O}(h) \tag{3.9}$$

Notez que l'ordre est le même, et que ces relations donnent des demi-dérivées. De fait, une même formule ne permet pas de calculer la dérivée pour le premier point de l'échantillon et pour le dernier point. Il y a un *effet de bord*.

D'un point de vue numérique, on écrira plutôt ces schémas

$$f'_i = \frac{f_{i\pm1} - f_i}{\pm h} \tag{3.10}$$

pour un point x_i quelconque de l'échantillon.

3.2.3 Différences centrées

En combinant les schémas excentrés avant et arrière ci-dessus, on parvient facilement à éliminer le terme d'ordre h. Reste alors une relation donnant la dérivée première à l'ordre supérieur, soit

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0 - h)}{2h} + \mathcal{O}(h^2) \tag{3.11}$$

que l'on écrira au point x_i

$$f'_i = \frac{f_{i+1} - f_{i-1}}{2h} \tag{3.12}$$

Ce schéma est appelé *différence finie centré à trois points*. Il met en jeu les 2 points immédiatement à gauche et à droite du point où l'on calcule la dérivée (ainsi que le point en question mais avec un poids nul). Une conbinaison différente des deux schémas excentrés précédents donne accès à un schéma d'ordre h^2 pour la dérivée seconde de f, soit

$$f''(x_0) = \frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h)}{h^2} + \mathcal{O}(h^2) \tag{3.13}$$

que l'on écrira

$$f''_i = \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2} \tag{3.14}$$

Ces deux schémas sont très courants dans les problèmes aux dérivées ordinaires et aux dérivées partielles. Il est important de remarquer qu'ils imposent un *échantillonnage régulier* des points définissant f.

3.2.4 Schémas d'ordres élevés

La relation (3.11) introduit deux effets de bords, l'un à gauche pour et l'autre à droite. On peut bien-sûr utiliser les différences en avant et en arrière données par la relation (3.10) mais elles sont d'ordre inférieur. Pour que l'ordre soit le même partout à l'intérieur de l'intervalle comme sur les bords, on peut construire des *schémas excentrés* d'ordre h^2. Ils sont donnés au point x_i par

$$f'_i = \pm\frac{-3f_i + 4f_{i\pm1} - f_{i\pm2}}{2h} \tag{3.15}$$

où le signe + vaut pour la différence en avant (pour $i = 1$) et le signe − pour la différence en arrière (en $i = N$). Des schémas d'ordres plus élevés existent, avec des effets de bords encore plus étendus, comme cette formule

$$f'_i = \frac{-f_{i+2} + 8f_{i-1} - 8f_{i-1} + f_{i-2}}{12h} \tag{3.16}$$

précise à l'ordre 4.

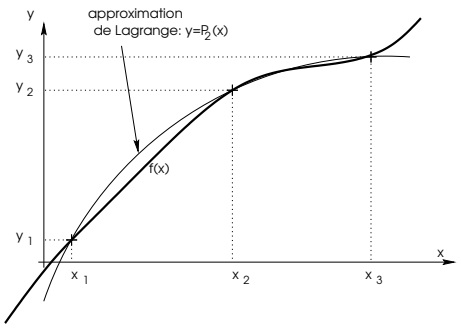


Figure 3.2: L'approximation polynômiale permet de calculer les dérivées successive d'une fonction en tout point d'un intervalle.

3.2.5 Echantillonage quelconque

On peut bien-sûr étendre ces schémas au cas d'une fonction irrégulièrement échantillonée. Par exemple, pour trois points consécutifs x_{i-1}, x_i et x_{i+1}, tels que $h = x_i - x_{i-1}$ et $h' = x_{i+1} - x_i$ avec $h \neq h'$, on montre que les dérivées première et seconde peuvent s'exprimer par les schémas

$$f'_i = \frac{hf_{i+1} + (h - h')f_i - h'f_{i-1}}{2hh'} \tag{3.17}$$

et

$$f''_i = \frac{(h + h')(hf_{i+1} + h'f_{i-1}) + (h - h')^2 f_i}{2h^2 h'^2} \tag{3.18}$$

respectivement.

3.3 Méthode génerale

L'une des difficultés liée à l'utilisation des schémas ci-dessus est que l'on ne peut pas calculer la dérivée en dehors des points de l'échantillon. Un méthode plus générale s'impose. Classiquement, on remplace, au moins localement, la fonction f par une fonction analytique simple que l'on dérivera par la suite. Il s'agit le plus souvent d'un polynôme $P_N(x)$ de degré N. Pour construire ce polynôme, il faudra $N+1$ points, au moins (dans certains cas, on pourra préférer réaliser un *ajustement*). On peut partir des polynômes de Lagrange et de Newton, particulièrement bien adaptés à l'interpolation, et on prendra garde de ne pas utiliser des polynômes de dégrés trop élévés au risque de voir apparaître le *phénomène de Runge* (des "oscillations; voir le chapitre 4). L'estimation la plus simple de la dérivée sera obtenue en assimilant la fonction à une droite, soit $f(x) \approx P_1(x)$, construite à partir de 2 points de l'échantillon, ce qui redonne les schémas excentrés d'ordre h établis plus haut.

L'utilisation des approximants (polynomiaux) permet également le calcul des dérivées seconde, troisième, etc. de f de manière complètement analytique. C'est très pratique, bien qu'il n'y ait pas de garantie de précision sur des dérivées d'ordre élevé.

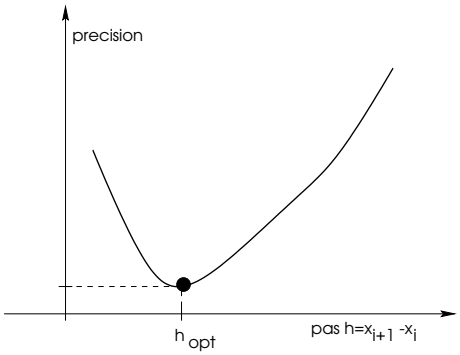


Figure 3.3: La précision sur la dérivée numérique d'une fonction dépend fortement du pas. En général, il existe un pas optimum h_{opt} qui correspond à la précision la plus grande possible copmpte-tenu des erreurs de troncature des schémas et des erreurs d'arrondis.

Enfin, mentionnons qu'il existe des méthodes qui emploient non seulement des informations sur la fonction mais aussi sur sa ou ses dérivées en x_i, comme les méthodes de Bernouilli, de Bessel, de Aitken ou de Everett. Ces méthodes sont aussi utilisées dans les interpolations et les quadratures. Ici, elles permettraient d'estimer la dérivée en $x \neq x_i$ avec une précision bien supérieure à celle obtenue par les seules différences finies sur f. Dans l'objectif d'utiliser ces méthodes que nous n'exposerons pas ici, il donc est recommander de générer une table plus complète, du type $\{(x_i, f(x_i), f'(x_i), \dots)\}_N$.

3.4 Pas optimum

La précision du calcul de dérivée se trouve limitée par deux effets cumulatifs: l'un est lié au schéma utilisé et l'autre vient des erreurs de représentation de machines. De ce fait, et contrairement à ce que l'on pourrait penser, faire tendre le pas h vers zéro ne donne pas une précision infinie. En d'autres termes, il existe un choix optimum h_{opt} pour h qui correspond à une précision-seuil qu'il est impossible de franchir. Cet effet est illustré à la figure 3.3. Par exemple, pour le schéma (3.11), l'erreur de troncature varie comme h^2, alors que l'erreur de représentation varie comme $1/h$. On peut dans ce cas montrer que le pas optimum est approximativement donné par

$$h_{\text{opt}} \approx \left(\frac{E}{M}\right)^{1/3} \tag{3.19}$$

où E correspond aux erreurs d'arrondis dans le calcul de f (c'est souvent un peu plus que la précision machine) et où M correspond à l'erreur de troncature (c'est en fait le majorant de la dérivée troisième sur l'intervalle). Pour les fonctions simples, h_{opt} pourra être très petit et approcher la précision machine. Un exemple est donné à la figure 3.4.

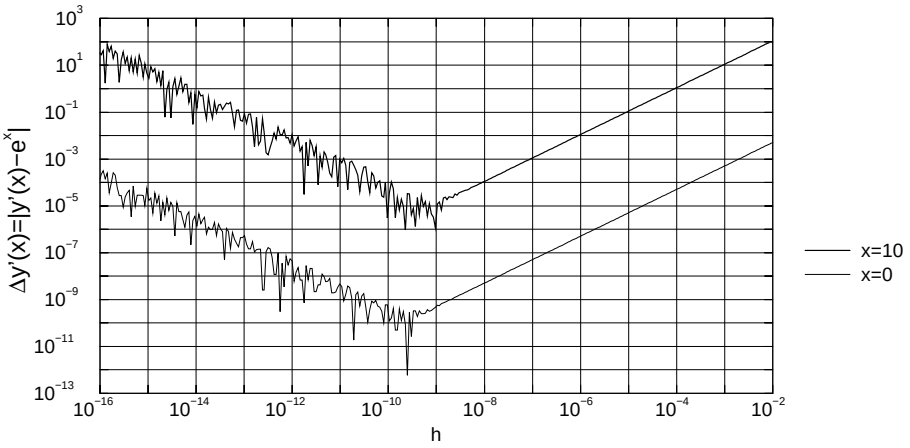


Figure 3.4: Exemple illustrant l'existence d'un pas optimum h_{opt} pour le calcul de la dérivée de la fonction e^x en $x = 0$ et $x = 10$. Ici, $h_{\text{opt}} \sim 10^{-9}$.

3.5 Extrapolation de Richardson

Une méthode puissante pour accroître l'ordre d'un schéma de dérivation est l'*extrapolation de Richardson*. Par cette méthode, l'amélioration de la précision est obtenue en considérant, quand c'est possible, la dérivée pour différents pas

$$h_0 \equiv h, \qquad h_1 \equiv \frac{h_0}{2}, \qquad h_2 \equiv \frac{h_0}{2^2}, \qquad \ldots, \qquad h_n \equiv \frac{h_0}{2^n}$$

et en combinant les dérivées obtenues pour chacun de ces pas. Ainsi, si $f'[h_k]$ désigne la dérivée en un point obtenue pour le pas $h_k = \frac{h_0}{2^k}$, alors une meilleure approximation est donnée par

$$f' \approx \frac{2^k f'[h_{k+1}] - f'[h_k]}{2^k - 1} \tag{3.20}$$

En pratique, on gagne au moins un ordre, comme le montre l'exemple de la figure 3.5.

3.6 Exercices et problèmes

■ Trouvez les schémas de différences finies centrées, en avant et en arrière pour les dérivées seconde et troisième.

■ Retrouvez les relations (3.17) et (3.18). Quel est leur ordre ?

■ Retrouvez la relation (3.16) en utilisant uniquement le schéma de dérivation centrée.

■ Etablissez le schéma (3.11) à l'aide de l'approximation polynomiale de Lagrange.

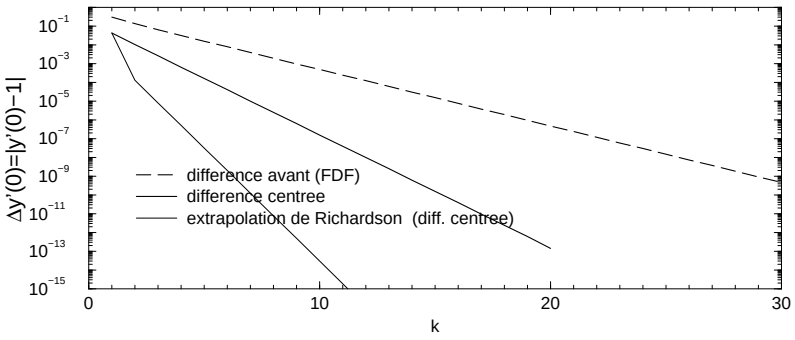


Figure 3.5: Calcul de la dérivée de e^x en $x = 0$ pour différents pas $h_k = 2^{-k}$.

■ Genérez un ensemble de N points $\{(x_i, f_i)\}_N$ pour une fonction f connue et calculez la dérivée en chaque point x_i de l'échantillon pour différentes largeurs d'intervalle et comparez à la valeur vraie. Quel intervalle conduit à la meilleure estimation ?

■ Tracez l'erreur commise sur la dérivée d'une fonction connue, par exemple $f(x) = \sin x$, en fonction de la largeur h de l'intervalle. Déduisez la largeur la plus pertinente. Comparez les résultats avec les dérivées estimées par un ensemble de p points $\{(x_i, f(x_i))\}_p$. Reprennez ces questions avec un schéma d'ordre plus élevé.

■ Quelle est l'ordre de grandeur du pas optimum permettant de calculer au mieux la dérivée première par le schéma centré à 5 points ? et la dérivée seconde par le shéma centré à 3 points ?

■ Dans quelle(s) condition(s) les schémas centrés à 3 et à 5 points donnent-ils des résultats similaires ?

■ Démontrez la relation (3.20)

Chapitre 4

Polynômes, interpolations et ajustements

Les polynômes constituent une famille de fonctions tout à fait remarquable en Mathématiques. Ils sont aussi un outil essentiel du calcul et de l'analyse numérique, notamment dans l'évaluation ou l'approximation des fonctions, dans les problèmes d'interpolation et d'extrapolation, dans la résolution d'équations différentielles, etc. Rappelons qu'un polynôme de degré N, noté P_N, est de la forme

$$P_N(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_{N-1} x^{N-1} + a_N x^N \tag{4.1}$$

où $a_N \neq 0$. Il a au plus N racines distinctes r_k, réelles ou imaginaires, telles que $P_N(r_k) = 0$. Ses dérivées successives sont aussi des polynômes, avec

$$P'_N(x) = a_1 x + 2a_2 x + 3a_3 x^2 + \cdots + (N-1)a_{N-1} x^{N-2} + N a_N x^{N-1} \tag{4.2}$$

$$P''_N(x) = a_1 + 2a_2 + 6a_3 x + \ldots + (N-1)(N-2)a_{N-1} x^{N-3} + N(N-1)a_N x^{N-2} \tag{4.3}$$

$$\ldots$$

$$P_N^{(N)}(x) = a_N N! \tag{4.4}$$

et

$$P_N^{(N+j)}(x) = 0, \qquad j \geq 1, \qquad \text{pour tout } x. \tag{4.5}$$

4.1 Formes polynômiales remarquables

On peut écrire un polynôme de plusieurs façons, selon l'utilisation que l'on en fait. Parmi les formes les plus intéressantes, on trouve les formes de Taylor, de Lagrange et de Newton.

4.1.1 Forme de Taylor

Elle est construite à partir de la forme (4.1), soit

$$\begin{aligned} \mathcal{E}_N(x) &= P_N(x - x_0) \\ &= a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots + a_N(x - x_0)^N \end{aligned} \tag{4.6}$$

Pour toute fonction $f(x)$, on peut en principe contruire le Polynôme de Taylor associé $\mathcal{E}_N(x)$ tel que $f(x) \approx P_N(x)$. Il suffit de connaître les dérivées de f jusqu'à l'ordre N, soient $f'(x)$, $f''(x)$, ..., $f^{(N-1)}$ et $f^{(N)}$. Les coefficients a_k de la série sont alors donnés par

$$a_k = \frac{f^{(k)}(x_0)}{k!}, \qquad k = 0, N \tag{4.7}$$

et l'erreur faite sur l'approximation est de l'ordre de $\frac{f^{(N+1)}(x_0)}{(N+1)!}(x - x_0)^{N+1}$.

4.1.2 Forme de Lagrange

Elle s'écrit

$$\mathcal{L}_N(x) = \sum_{k=0}^{N} y_k L_{N,k}(x) \tag{4.8}$$

où $L_{N,k}(x)$ sont les coefficients de Lagrange définis par

$$L_{N,k}(x) = \frac{(x - x_0)\dots(x - x_{k-1})(x - x_{k+1})\dots(x - x_N)}{(x_k - x_0)\dots(x_k - x_{k-1})(x_k - x_{k+1})\dots(x_k - x_N)} \tag{4.9}$$

et $y_k = P_N(x_k)$. Notez que ces coefficients sont en fait des polynômes.

4.1.3 Forme de Newton

Elle s'écrit

$$\begin{aligned}\mathcal{N}_N(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \dots \\ + a_N(x - x_0)(x - x_1)\dots(x - x_{N-1})\end{aligned} \tag{4.10}$$

où les x_k sont les *centres*. En particulier, on voit que

$$\mathcal{N}_N(x) = \mathcal{N}_{N-1}(x) + a_N(x - x_0)(x - x_1)\dots(x - x_{N-1}) \tag{4.11}$$

4.2 Quelques polynômes remarquables

Les polynômes de Legendre, de Chebyshev et de Hermite dont nous donnons ci-dessous une forme génératrice forment une famille de polynômes orthogonaux. Ils jouissent d'un certain nombre de propriétés interessantes que l'on retrouvera dans des ouvrages spécialisés.

4.2.1 Polynômes de Chebyshev

Ils sont définis par la relation de récurrence

$$\mathcal{T}_k(x) = 2x\mathcal{T}_{k-1}(x) - \mathcal{T}_{k-2}(x) \tag{4.12}$$

avec $\mathcal{T}_0(x) = 1$ et $\mathcal{T}_1(x) = x$. En particulier, les polynômes sont pairs pour k pair et impairs pour k impair. Sur l'intervalle $[-1, 1]$ sur lequel ils sont essentiellement utilisés (voir la figure 4.1), ils sont de norme finie avec $|\mathcal{T}_k(x)| \leq 1$. Les k racines de $\mathcal{T}_k$ (encore appelées *noeuds*) sont donnés par $r_j = \cos\frac{\pi}{2N}(2j + 1)$, $j = 0, k - 1$.

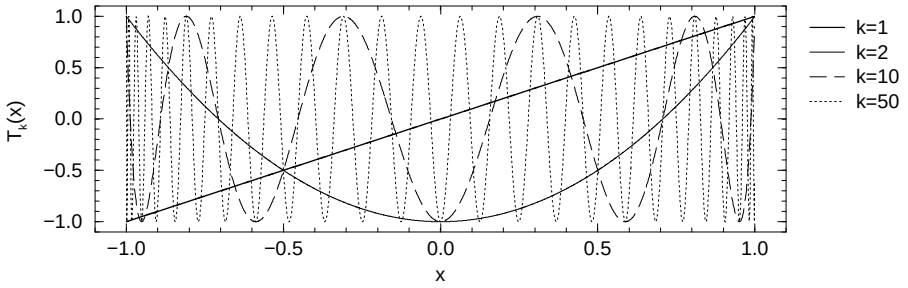


Figure 4.1: Polynôme de Chebyshev de première espèce $\mathcal{T}_k(x)$ sur $[-1, 1]$ pour $k = 1, 2, 10$ et 50.

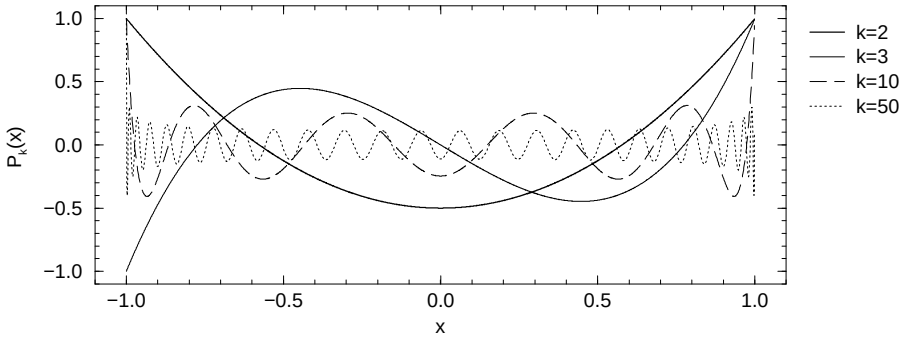


Figure 4.2: Polynôme de Legendre $\mathcal{P}_k(x)$ sur $[-1, 1]$ pour $k = 2, 3, 10$ et 50.

4.2.2 Polynômes de Legendre

Ils sont définis par la relation de récurrence

$$(k+1)\mathcal{P}_{k+1}(x) = (2k+1)x\mathcal{P}_k(x) - k\mathcal{P}_{k-1}(x) \tag{4.13}$$

avec $\mathcal{P}_0(x) = 1$ et $\mathcal{P}_1(x) = x$. Leur parité est celle k. Sur l'intervalle $[-1, 1]$ (voir la figure 4.2), ils satisfont $|\mathcal{P}_k(x)| \leq 1$.

4.2.3 Polynômes de Hermite

Ils sont définis par la relation de récurrence

$$\mathcal{H}_{k+1}(x) = 2x\mathcal{H}_k(x) - 2k\mathcal{H}_{k-1}(x) \tag{4.14}$$

avec $\mathcal{H}_0(x) = 1$ et $\mathcal{H}_1(x) = 2x$ (voir la figure 4.3). Il ont la même parité que k.

4.3 Evaluation d'un polynôme

Calculer la valeur $P_N(x)$ d'un polynôme en un point x n'est pas aussi immédiat qu'on pourrait le penser, en particulier lorsque les coefficients a_k sont très

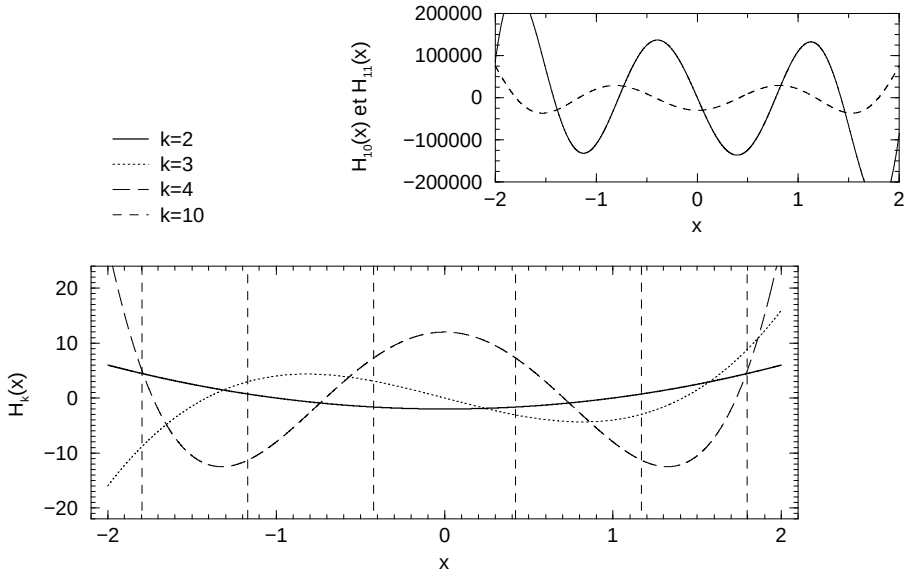


Figure 4.3: Polynôme de Hermite $\mathcal{H}_k(x)$ sur $[-2, 2]$ pour $k = 2, 3, 4, 10$ et 11.

différents les uns des autres, ou quand l'ordre du polynôme est élevé (disons $N \gtrsim 5$). Comme pour l'évaluation de n'importe qu'elle fonction simple, on s'attend naïvement à ce $P_N(x)$ soit connu à la précision de la machine; mais les erreurs numériques (représentation et annulation soustractive) peuvent être importantes sans précautions particulières, notamment à cause de l'opération d'élévation à la puissance. Plutôt que d'appliquer *stricto sensu* le schéma suggéré par la relation (4.1), c'est-à-dire

$$\mathtt{Y} \leftarrow \mathtt{a_0} + \mathtt{a_1} * \mathtt{x} + \mathtt{a_2} * \mathtt{x} * *2 + \cdots + \mathtt{a_N} * \mathtt{x} * *\mathtt{N}, \tag{4.15}$$

il est préférable de faire appel, si possible, à la forme *imbriquée*, ou encore *forme de Horner* où $P_N(x)$ est ré-écrit de la façon suivante

$$P_N(x) = a_0 + x(a_1 + x(a_2 + x(a_3 + \cdots + x(a_{N-1} + a_N x)))) \tag{4.16}$$

Sous cette nouvelle forme, l'algorithme devient *récursif* avec

$$\begin{cases} \mathtt{Y} \leftarrow \mathtt{x} \\ \mathtt{Y} \leftarrow \mathtt{a_{N-k}} + \mathtt{a_{N-k+1}}\mathtt{Y}, \qquad \mathtt{k} = 1, \mathtt{N}, \end{cases}$$

et il donc très facile à programmer. A la fin du cycle, la valeur du polynôme est contenue dans la variable intermédiaire `Y`. Alors que le schéma (4.15) compte environ $\frac{N^2}{2}$ opérations (essentiellement des multiplications), celui-ci n'en compte plus que $2N$ environ (multiplications et additions). Il est donc plus rapide, mais surtout, il diminue les risques d'erreurs.

On voit que les formes de Taylor et de Newton sont particulièrement bien adaptées au calcul numérique de ce type. Pour les polynômes qui peuvent être directement définis par une relation de récurrence, comme les polynômes de Legendre, de Chebyshev, de Hermite ou autre, la méthode de Horner ne presente pas d'intérêt, car on utilisera évidemment le schéma récursif associé.

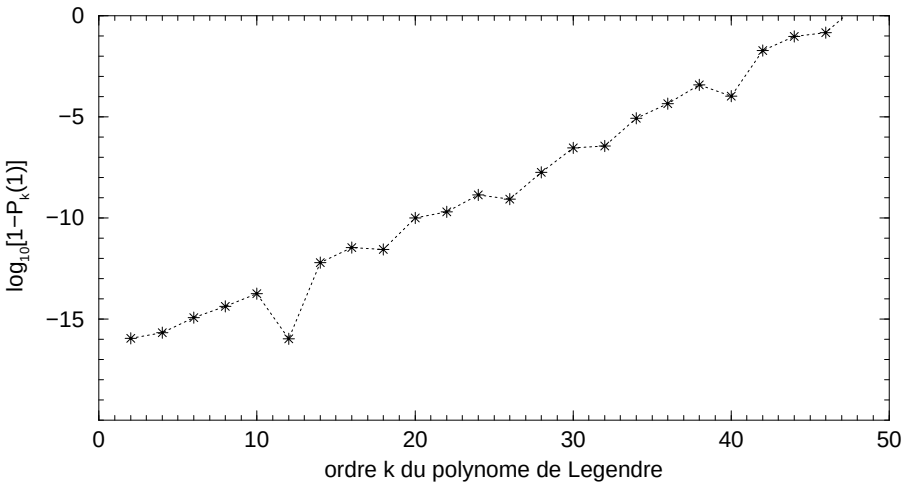


Figure 4.4: Exemple d'erreur commise sur le calcul de $\mathcal{P}_{2k}(1)$ lorsque l'on utilise un développement en série des polynôme de Legendre plutôt que la relation de récurrence (qui donne $P_{2k}(1) = 1$ à la précision de la machine, pour tout k).

4.4 Interpolations et extrapolation

On désigne par *interpolation*, l'opération qui consite à estimer la valeur d'une fonction f en un point d'abscisse $x = c$ de l'intervalle $[x_1, x_N]$, sachant que f est uniquement connue sous forme tabulée $\{(x_i, f_i)\}_N$. Lorsque $c < x_1$ ou $c > x_N$, on parle d'*extrapolation*.

4.4.1 Interpolation linéaire

L'interpolation la plus simple est l'interpolation linéaire qui s'écrit à une dimension

$$f(c) = pf_j + (1-p)f_{j+1}, \qquad 1 \leq j \leq N-1 \tag{4.17}$$

où $[x_j, x_{j+1}]$ est le sous-intervalle contenant c et

$$p = \frac{x_{j+1} - x}{x_{j+1} - x_j} \tag{4.18}$$

Cette méthode est imparable. Non seulement elle est rapide et ne pré-suppose pas un espacement régulier des points, mais elle assure que $f(c)$ est toujours situé dans le rectangle défini par les points diamétralement opposés (x_j, f_j) et (x_{j+1}, f_{j+1}). Son inconvénient majeur, surtout si l'échantillonnage est "diffus", est illustré à la figure 4.5: elle produit une fonction d'apparence non lissée, plutôt en ligne brisée. Ceci peut être très génant (en particulier, la dérivée première est discontinue).

4.4.2 Approximations polynomiales

On remédie au problème ci-dessus grâce à un polynôme d'ordre $N > 1$. Comme la fonction est tabulée en N points, il existe toujours un polynôme de degré

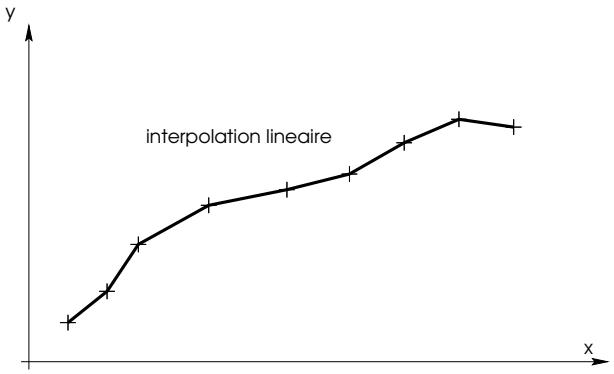


Figure 4.5: L'interpolation linéaire produit toujours une fonction en ligne brisée (sauf lorsque les points sont exactement alignés!).

$N-1$ qui passe par ces N points. En pratique, on cherche N coefficients a_k, tels que $P_{N-1}(x_i) = f(x_i)$, puis l'on estime $f(c)$ *via* P_{N-1}. On obtient alors un système linéaire de N équations où les N inconnus sont les coefficients a_k de la décomposition que l'on peut en principe résoudre par les méthodes standards. Il s'agira d'inverser la matrice de *Vandermonde* V associée

$$V = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{N-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{N-1} \\ 1 & x_2 & x_3^2 & \dots & x_3^{N-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_N & x_N^2 & \dots & x_N^{N-1} \end{pmatrix} \tag{4.19}$$

Une méthode beaucoup plus rapide, car plus directe consiste à utiliser les polynômes de Lagrange (voir la figure 4.6) dont la construction est très ingénieuse. Nous voyons en effet que pour $x = x_k$, tous les coefficients de Lagrange sont nuls sauf $L_{N,k}(x_k)$ qui vaut 1. Autrement dit, en imposant $\mathcal{L}_N(x) = f(x)$ en tous les points de collocation, le problème de l'interpolation se résumera au calcul de ces coefficients donnés par la formule (4.8) et à l'application de la relation (4.9) avec $x = c$. On réalise alors une *approximation de Lagrange*. Notez que lorsque les points sont régulièrement espacés, le calcul des coefficients de Lagrange se simplifie: leur dénominateur vaut $\pm h^{N-1}$, où h est l'espacement entre les points. Cette méthode est donc très efficace, à condition que les abscisses soient fixes.

Autre méthode efficace, celle qui consiste à construire les polynômes de Newton. Comme le montre la relation (4.10), $\mathcal{N}_{N-1}(x_0) = a_0 = f(x_0)$. Par

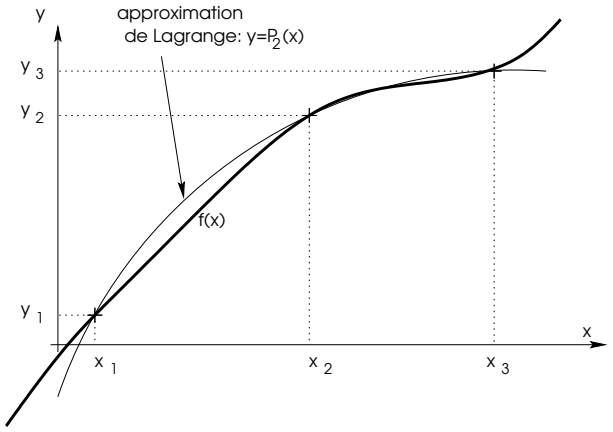


Figure 4.6: L'approximation de Lagrange consiste à faire passer un polynôme de degré $N-1$ par N points tabulés $(x_i, f(x_i))_N$.

conséquent $\frac{\mathcal{N}_{N-1}(x_0)-a_0}{x-x_0}$ est un polynôme de Newton. Il s'écrit

$$\begin{aligned}\frac{\mathcal{N}_{N-1}(x_0)-a_0}{x-x_0} &= a_1 + a_2(x-x_1) + \cdots + a_N(x-x_1)\ldots(x-x_{N-1}) \\ &\equiv \mathcal{N}^{(I)}_{N-2}(x) \end{aligned} \tag{4.20}$$

En particulier, en $x = x_1$, il fournit $a_1 = f(x_1)$. A son tour, $\frac{\mathcal{N}^{(I)}_{N-2}(x_1)-a_1}{x-x_1}$ est un polynôme de Newton qui permet de calculer a_2, etc. En poursuivant ce procédé que l'on appelle processus des *différences divisées*, on trouve tous les coefficients du polynôme de Newton, donc le polynôme interpolant. On aura donc $f(c) = \mathcal{N}_{N-1}(c)$. On réalise alors une *approximation de Newton.*

4.4.3 Phénomème de Runge

Si l'on peut toujours faire passer un polynôme par un ensemble de N points, la fonction interpolante n'a pas le rôle de lissage que l'on pourrait lui prêter intuitivement ou que l'on souhaiterait. Comme l'indique la figure 4.8, le résultat peut donner des oscillations, ou phénomène de *Runge*, en particulier lorsque N est grand. Autrement dit, si l'erreur d'interpolation est nulle sur les points de collocation par construction, elle peut être grande entre eux. L'ordre n'est pas seul en cause: le phénomène peut disparaître avec un espacement irrégulier.

Mentionnons à ce propos que l'interpolation est un processus qui génère de l'information là où elle n'est en principe pas disponible. Il faudra donc garder à l'esprit que toute opération qui pourra être effectuée à partir d'une valeur ou d'une fonction interpolée (par exemple les opérations de dérivation) sera entachée d'une erreur pouvant être très importante. De ce point de vue, l'extrapolation, dont nous n'avons pas parlé, doit être manipulée avec une extrême prudence, surtout si le point c est loin de l'intervalle de tabulation. On peut trouver à peu près n'importe quoi.

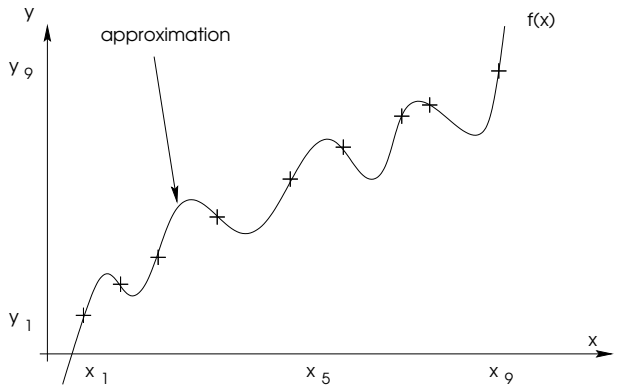


Figure 4.7: Oscillations apparaissant quand on tente de faire passer un polynôme de degré élevé par un ensemble de points.

4.4.4 Méthode générale

Enfin, les polynômes comme base de fonctions interpolantes n'est évidemment pas le seul choix possible; on peut en pratique utiliser d'autres types de fonctions $\tilde{f}_k(x)$ (voir §4.6). Le principe reste le même. On cherchera a représenter la fonction tabulée par une fonction analytique du type

$$f(x) \approx \sum_k a_k \tilde{f}_k(x) = \mathcal{I}(x) \tag{4.21}$$

où l'on imposera $f_i = \mathcal{I}_i$ pour touts les points tabulés. L'écriture des N contraintes fournira les coefficients a_k.

Signalons qu'il existe des méthodes qui utilisent des informations sur la ou les dérivées successives de la fonction à interpoler (comme les formules de Bessel et de Everett). Rares toutefois sont les cas où l'on dispose de ce type d'information. Autrement dit, si des données doivent être ré-utilisées et extrapolées, il peut être judicieux de génerer en même temps que f, sa dérivée première (au moins) et d'appliquer ces méthodes spécifiques.

4.5 Principes de l'ajustement

Dans le processus d'interpolation, on utilise tous les points de mesure pour construire un polynôme P_N ou une fonction interpolante $\mathcal{I}(x)$. Comme nous l'avons mentionné, cette opération peut s'avérer numériquement risquée pour N grand. Elle peut également être dénuée de sens physique, par exemple si les points tabulées décrivent une loi linéaire tout en affichant une certaine dispersion. Il est fréquent de rencontrer ce type de situation lorsque les points sont issus d'une expérience où chaque mesure est entachée d'une erreur. On peut alors souhaiter extraire une corrélation simple entre deux variables x et y d'un échantillon sans pour autant vouloir une fonction interpolante qui passe par tous les points. On cherche alors à réaliser un *ajustement* (ou "curve fitting" en anglais). Cet ajustement peut mettre en jeu une fonction tout-à-fait quelconque. Le plus souvent, on travaille avec un polynôme P_k de degré k tel que $k + 1 < N$, N étant le nombre de points de collocation. La question qui se pose alors est la suivante:

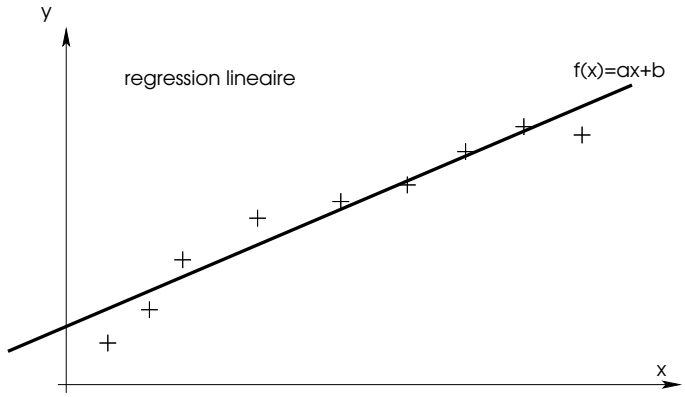


Figure 4.8: Dans les problèmes d'ajustement, on cherche à minimiser la distance entre les points de collocation et la fonction d'ajustement f.

comment, une fois la fonction sélectionnée, trouver ses paramètres de manière à réaliser le meilleur ajustement ? Le procédé est quasiment toujours le même. On commence par calculer, pour chaque point de l'échantillon, la distance d_i à la courbe de la fonction, ce qui se solde par l'écriture de N équations. Puis, l'on impose que la distance totale $\sum_i |d_i|$ soit la plus petite possible.

Par exemple, pour trouver la droite moyenne d'équation $y = ax + b$ (a et b étant les paramètres de la fonction à déterminer) passant au mieux un échantillon contenant N points (x_i, y_i), on peut montrer qu'il faut en fait résoudre les deux équations suivantes

$$\begin{cases} a\sum_i^N x_i^2 + bN\langle x\rangle - \sum_i^N x_i y_i = 0 \\ a\langle x\rangle - b - \langle y\rangle = 0 \end{cases}$$

où $\langle x\rangle$ est la moyenne des $\{x_i\}$, et $\langle y\rangle$ celle des $\{y_i\}$. La solution de ce système (les coefficients a et b) est généralement triviale à calculer. Toutes les calculatrices scientifiques sont équipées d'un petit programme de ce type que l'on invoque en appuyant sur la touche *régression linéaire*.

4.6 Splines

Récapitulons: l'approximation polynomiale devient critique lorsque le nombre de points de mesure est grand et l'ajustement par un polynôme de bas degré peut être insatisfaisant. Une première alternative est donné par l'interpolation linéaire mais elle conduit à une fonction interpolée en ligne brisée. Comment faire alors pour arrondir les angles ? La réponse est proposée par la construction d'une fonction *spline cubique* $S(x)$: c'est une fonction définie par morceaux, comme l'interpolant linéaire, mais elle possède les caractéristiques suivantes

- sur chaque intevalle $[x_i, x_{i+1}]$, la fonction $S(x)$ est un polynôme du troisième degré. On le note $P_{3;i,i+1}(x)$
- la fonction $S(x)$ passe par tous les points de l'échantillon, quelque soit leur nombre. C'est-à-dire que $P_{3;i,i+1}(x_{i+1}) = P_{3;i+1,i+2}(x_{i+1})$.

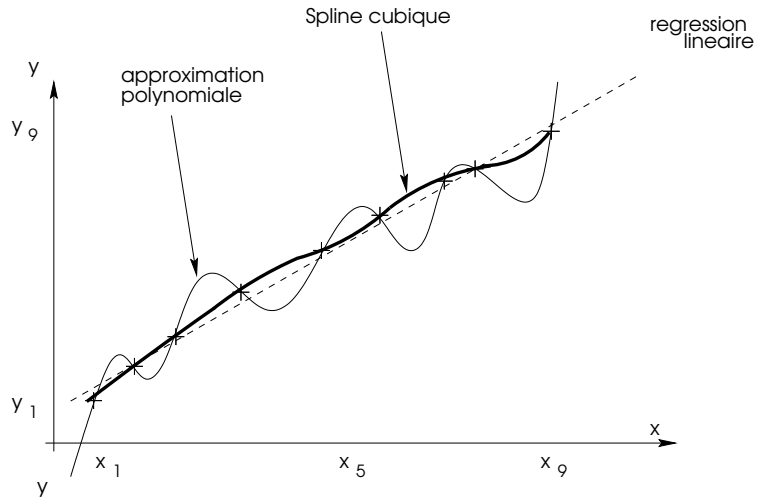


Figure 4.9: Regression, approximation polynomiale et spline cubique.

- $S'(x)$ est continue sur chaque noeuds, soit $P'_{3;i,i+1}(x_{i+1}) = P'_{3;i+1,i+2}(x_{i+1})$
- $S''(x)$ est continue sur chaque noeuds, soit $P''_{3;i,i+1}(x_{i+1}) = P''_{3;i+1,i+2}(x_{i+1})$

L'ensemble de ces contraintes définit un certain nombre de relations entre les $N-1$ polynômes $P_{3;i,i+1}(x)$ qui permettent en principe de les caractériser complètement, c'est-à-dire de trouver les $4(N-1)$ coefficients. Le résultat est généralement assez spectaculaire: le spline passe par tous les points tabulés et présente les caractéristiques (au moins visuelles) d'une fonction de lissage sans produire d'oscillations pour N élevé. C'est probablement l'une des meilleures méthodes qui existe (voir la figure 4.9).

4.7 Exercices et problèmes

■ Ecrivez un algorithme de calcul d'un polynôme de Lagrange mettant en oeuvre un schéma récursif de type Horner.

■ Ecrivez les polynômes de Legendre sous la forme de Horner.

■ Calculer les valeurs du polynôme de Legendre $\mathcal{P}_{10}(x)$ sur $[-1,1]$ de trois manières différentes: (i) sous la forme développée du type (4.1), (ii) en utilisant la relation de récurrence (4.13) et (iii) à partir de la forme de Horner (4.16). Comparer le nombre d'opération et la précision obtenue.

■ Retrouver la relation (4.17).

■ Retrouvez par l'approximation de Lagrange, l'équation de la parabole qui passe par les points $(-1,1)$, $(0,0)$ et $(1,1)$.

■ Comparez, en terme de nombdre d'opérations élémentaires, l'approximation de Lagrange et la méthode des différences divisées de Newton.

■ Etablissez la formule d'interpolation linéaire à 2 dimensions.

■ Retrouvez les 2 équations ci-dessus. Comment se transforment-elles si la regression est quadratique en x ? en y ? si elle est en loi de puissance ?

■ Dressez le bilan des contraintes sur les coefficients de $S(x)$ et montrer que le système est, en fait, indéterminé.

Chapitre 5

Quadratures

5.1 Rappels

Rappelons que l'intégrale définie I d'une fonction f de la variable x sur un intervalle $[a, b]$ représente l'aire, comptée algébriquement, comprise entre la courbe représentative de la fonction, l'axe des abscisses et les droites d'équations $x = a$ et $x = b$ (voir la figure 5.1), soit la quantité

$$I = \int_a^b f(x)dx = F(b) - F(a) = \int_a^b \frac{dF}{dx}dx \tag{5.1}$$

C'est une *intégrale définie* dont le résultat est un *scalaire*. On pourra donc facilement former une nouvelle fonction F de la variable b, *primitive* de f (ou encore *intégrale indéfinie*), telle que

$$F(b) = \int_a^b f(x)dx + F(a) \tag{5.2}$$

Le recouvrement des techniques présentées ici avec celles dédiées à la résolution des équations différentielles ordinaires (voir le chapitre 7) est immédiat puisque F est solution de l'équation

$$\frac{dF}{dx} = f(x) \tag{5.3}$$

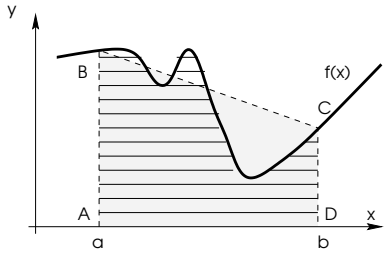


Figure 5.1: Intégrale définie d'une fonction f (*surface hachurée*). Dans la méthode d'integration des trapèzes, on assimile cette fonction à une fonction linéaire et l'intégrale est donné par l'aire du trapèze ABCD (*surface grisée*).

où la primitive prend la valeur $F(a)$ en $x = a$. En pratique, l'approximation de I s'effectue le plus souvent à l'aide de combinaisons linéaires des valeurs de f, et éventuellement de ses dérivées, calculées dans l'intervalle $[a, b]$.

Il est important d'ajouter que l'intégrale où la primitive d'une fonction analytique peut ne pas être analytique, contrairement au calcul des dérivées. Les méthodes présentées ici concernent de telles situations, incluant les cas très fréquents où la fonction est définie par un ensemble de points $\{(x_i, y_i = f(x_i))\}$, par exemple issus de mesures expérimentales.

5.2 Méthode des trapèzes

La méthode la plus directe est la *méthode des trapèzes* qui se résume au schéma suivant

$$I \approx \frac{b-a}{2}\left(f(a) + f(b)\right), \tag{5.4}$$

explicité ici à un seul et unique trapèze dont les arètes sont les quatre points A$(a, 0)$, B$(a, f(a))$, C$(b, f(b))$ et D$(b, 0)$. Comme le montre la figure 5.1, cela revient à assimiler la fonction à une droite. De ce fait, la méthode des trapèzes ne donne un résultat exact que lorsque la fonction est une fonction linéaire. On peut aisément se convaincre (au moins graphiquement) que l'approximation (5.4) sera d'autant meilleure que la fonction f aura un gradient essentiellement constant sur l'intervalle $[a, b]$. Tout dépend de la précision souhaitée.

La relation approchée (5.4) peut-être interprétée légèrement différemment. En supposant que f est constante sur l'intervalle, on a

$$\int_a^b f(x)dx = f(x)\int_a^b dx = f(c)(b-a), \tag{5.5}$$

où c est à déterminer. On retrouve le résultat précédent en choisissant c tel que $f(c) = \frac{1}{2}\left(f(a) + f(b)\right)$ (c'est la valeur moyenne de f pour l'échantillon à deux points).

Notez qu'il n'y a aucune raison des privilégier les bords. En d'autres termes, on peut très bien considérer le schéma suivant

$$\int_a^b f(x)dx = f(m)(b-a), \tag{5.6}$$

où m est le milieu de l'intervalle, soit $m = a + \frac{b-a}{2}$. C'est la *méthodes des rectangles*.

5.2.1 Précision

La méthode des trapèzes nécessite deux évaluations de la fonction, aux points $x_1 = a$ et $x_N = b$. C'est un *schéma à 2 points*. Il est d'*ordre* h^3 (avec $h = b-a$), ce qui signifie que l'erreur commise sur I est proportionnelle à h^3. On peut s'en convaincre en considérant b infiniment proche de a, de tel sorte qu'un développement de Taylor autour de a donne

$$f(x) = f(a) + (x-a)f'(a) + \mathcal{O}((x-a)^2) \tag{5.7}$$

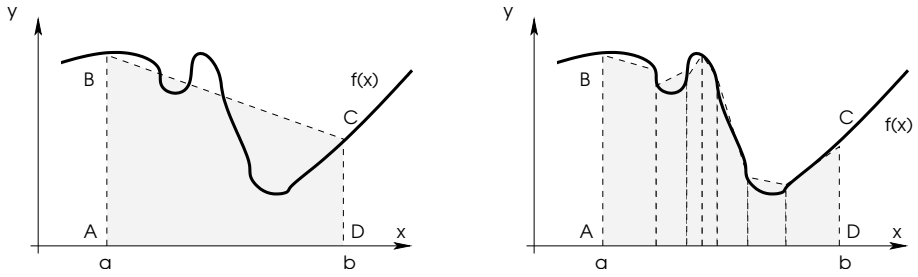


Figure 5.2: Pour les fonctions qui présentent de fortes variations sur l'intervalle d'intégration ou si l'intervalle est trop grand, la méthode des trapèzes peut donner "n'importe quoi", a moins qu'elle ne soit utilisée sur des sous-intervalles réguliers (*méthode composée*) ou irréguliers où l'on concentre le découpage dans les zones de forts gradients (*méthode composée adaptative*).

pour $x \in [a, b]$. En reportant cette expression dans la relation (5.5), on trouve

$$\int_a^b f(x)dx = \frac{b-a}{2}\left(f(a)+f(b)\right) + \int_a^b \mathcal{O}\left((x-a)^2\right)dx \tag{5.8}$$

$$\approx \frac{b-a}{2}\left(f(a)+f(b)\right) + \mathcal{O}(h^3) \tag{5.9}$$

Notez qu'il y a une différence essentielle entre dérivation et d'intégration: pour un schéma à deux points, la première est d'ordre h alors que la seconde est d'ordre h^3.

5.2.2 Version composée

On peut augmenter l'ordre de la méthode, c'est-à-dire réduire l'erreur, grâce à la méthode des trapèzes *composée* qui consiste à découper l'intervalle $[a, b]$ en 2, 3, ... ou $N-1$ sous-intervalles de largeur constante $h = x_{i+1} - x_i$, avec $i \in [1, N-1]$. Dans ces conditions, on montre facilement que

$$I \approx \frac{h}{2}\left(f(a)+f(b)\right) + h\sum_{i=2}^{N-1} f_i \tag{5.10}$$

où $f_i \equiv f(x_i)$. Formellement, l'ordre n'a pas changé: c'est toujours h^3, mais h est plus petit. On voit donc que si $N \gg 1$, alors $h \ll b-a$ et la précision est meilleure. Une autre façon de s'en convaincre est de remarquer que $h = \frac{b-a}{N-1}$, c'est-à-dire que pour un intervalle donné, la précision varie approximativement comme N^{-3}. Notez que ce gain n'est pas sans efforts: il faudra à présent $N+1$ évaluations de la fonction, ce qui se solde par un temps de calcul plus grand.

5.2.3 Découpage adaptatif

On peut procéder à un découpage *adaptatif* de l'intervalle en sous-intervalles de largeurs inégales, par exemple dans les zones où la fonction présente de forts gradients. C'est un peu plus difficile à gérer car il faut savoir repérer numériquement de telles zones. On peut à cet effet, faire appel à la dérivée seconde de f.

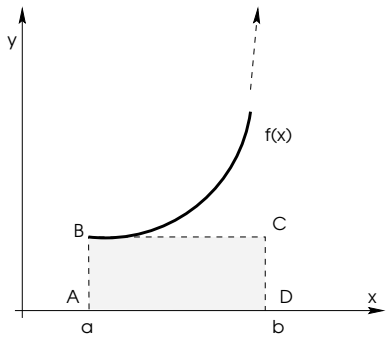


Figure 5.3: Quand une fonction est singulière, on peut, en première approximation, utiliser une formule ouverte ou semi-ouverte.

5.2.4 Schémas ouvert et semi-ouvert

Rien n'oblige à utiliser tous les points de l'échantillon. On peut en effet écarter un ou plusieurs point(s) particulièrement gênant(s). Par exemple, si f diverge en b (voir la figure 5.3), on pourra remplacer le schéma (5.4) par

$$I \approx hf(a) \tag{5.11}$$

C'est ce que l'on appelle un sch'ema *semi-ouvert*. Dans un *schéma ouvert*, deux points sont laissés de coté. Bien-sûr, dans ce genre de traitement, il ne faut pas s'attendre à un résultat précis. Il existe des techniques permettant de traiter correctement les singularités.

5.3 Méthodes de Simpson

La méthode des trapèzes est simple et efficace, mais elle nécessite souvent un échantillonnage très dense. On peut avoir recours utilise à des schémas plus précis, comme par exemple au premier schéma de *Simpson*

$$I \approx \frac{h}{3}\left(f(a) + 4f_2 + f(b)\right), \tag{5.12}$$

où $f_2 \equiv f(x_2)$, $h = x_3 - x_2 = x_2 - x_1$, $a \equiv x_1$ et $b \equiv x_3$. C'est un *schéma à 3 points*. Il est d'ordre h^5 et suppose un échantillonnage régulier. On pourrait montrer qu'il correspond en fait à l'approximation de la fonction par un polynôme de degré 2, passant par les points d'abscisses x_1, x_2 et x_3. Le schéma de Simpson est, de fait, rigoureusement exact pour une fonction du type: $y(x) = \alpha x^2 + \beta x + \gamma$.

Comme pour la méthode des trapèzes, on peut procéder à un découpage de l'intervalle d'intégration en $N-1$ sous-intervalles égaux et appliquer le schéma élémentaire ci-dessus sur chaque sous-intervalle. On réalise alors un *schéma de Simpson composé*. Le découpage peut également être adaptatif.

Il existe une *seconde règle de Simpson*, d'ordre h^5 également. Le schéma est

$$I = \frac{3h}{8}\left(f_1 + 3f_2 + 3f_3 + f_4\right), \tag{5.13}$$

C'est donc un *schéma à 4 points*. Il est obtenu par un polynôme interpolant de degré 3.

5.4 Formules de Newton-Cotes

On peut construire des schémas encore plus précis grâce à l'interpolation polynômiale. La règle générale est la suivante: un schéma à $k+1$ points régulièrement espacés doit donner un résultat exact pour tout polynôme $P_k(x)$ de degré $\leq k$. Les schémas associés font partie de ce que l'on appelle classiquement les *quadratures de Newton-Cotes*. La méthode des trapèzes correspond à $k = 1$, celles de Simpson à $k = 2$ et 3. Le schéma de *Boole*

$$I = \frac{2h}{45}\left(7f_1 + 32f_2 + 12f_3 + 32f_4 + 7f4\right), \tag{5.14}$$

est obtenu avec $k = 4$, etc. Plus le degré du polynôme est élevé, plus l'ordre du schéma est élevé et donc, plus la précision est grande. La seule limitation est le nombre p de points disponibles dans l'échantillon: il doit satisfaire $k \leq p - 1$. On évitera toutefois de construire des schémas mettant en jeu des polynômes de degrés très élevés.

5.5 Schémas récursifs

Pour les formules de Newton-Cotes, on peut mettre en place des méthodes récursives qui permettent d'une part de limiter le nombre d'évaluation de la fonction et d'autre part de contrôler la précision de l'intégration. Prenons par exemple la méthode des trapèzes et notons $I[h]$ l'approximation de l'intégrale de f obtenue en découpant $[a, b]$ en $J = N - 1$ sous-intervalles de largeur $h = \frac{b-a}{J}$. Pour calculer $I[h]$, il aura été nécessaire d'évaluer la fonction en $J + 1$ points $x_1, \ldots, x_N$. On peut alors montrer que si l'on découpe chaque intervalle en 2 faisant apparaître J nouveaux points f_j^{new} (il y en a maintenant $2J$ au total), une meilleure approximation de I est donnée par

$$I\left[\frac{h}{2}\right] = \frac{1}{2}I[h] + \frac{h}{2}\sum_{j=1}^{J} f_j^{\text{new}} \tag{5.15}$$

On gagne ainsi un facteur 8 en précision (soit presque un chiffre significatif). On peut bien-sûr appliquer cette technique aux règles de Simpson et autres, puis les combiner, toujours dans l'objectif d'accroître la précision du calcul. C'est le principe de construction des tables de *Romberg*.

5.6 Méthode générale

Une formulation plus générale du problème des quadratures repose sur l'existence de N poids w_i tels que l'intégrale

$$\int_a^b g(x)f(x)dx = \sum_{i=1}^{N} w_i f_i \tag{5.16}$$

où g est une fonction quelconque, doit être exacte pour un certain ensemble de fonctions. La plupart du temps, cet ensemble de fonction est l'espace vectoriel

Quadrature	conditions
Gauss-Legendre	$g(x) = 1$ ou $a = -b = -1$
Gauss-Laguerre	$g(x) = e^{-x}$ ou $g(x) = x^p e^{-x}$
Gauss logarithmique	$g(x) = \ln x$, $a = 0 = b - 1$ avec $a = 0$ et $b = \infty$
Newton-Cotes	$g(x) = 1$, avec $a = -b = -1$
Tchebysheff	$g(x) = 1$ et $w_i = C^{te}$ avec $a = -b = -1$
Tchebysheff-Radau	$g(x) = x$ ou $\frac{x}{\sqrt{1-x^2}}$, avec $a = -b = -1$
Radau	$g(x) = 1$, $a = -b = -1$)
Filon	$g(x) = \sin kx$
"Monte-carlo"	

Table 5.1: Quelques quadratures standards.

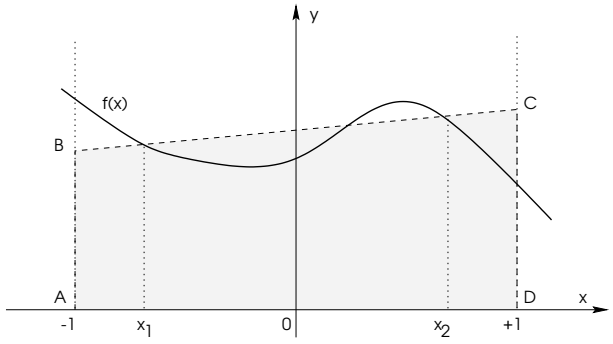


Figure 5.4: Dans l'intégration de Gauss-Legendre à 2 points, on cherche les noeuds x_1 et x_2 tels que la surface du trapèze ABCD coincide au mieux avec l'intégrale de la fonction. La méthode est exacte pour une polynôme de degré au plus 3.

P_k des polynômes de degré k. Il est facile de voir que les formules de Newton-Cotes sont bien du type de (5.16), avec

- $w_i = \{\frac{1}{2}, 1, 1, \ldots, 1, \frac{1}{2}\}$ pour la méthode des trapèzes,
- $w_i = \{\frac{1}{3}, \frac{4}{3}, \frac{2}{3}, \frac{4}{3}, \ldots, \frac{1}{3}\}$ pour la première méthode de Simpson,
- $w_i = \{\frac{3}{8}, \frac{9}{8}, \frac{9}{8}, \frac{3}{4}, \ldots, \frac{9}{8}, \frac{3}{8}\}$ pour la seconde méthode de Simpson,
- $w_i = \{\frac{14}{45}, \frac{64}{45}, \frac{24}{45}, \frac{64}{45}, \frac{28}{45}, \ldots, \frac{64}{45}, \frac{14}{45}\}$ pour la méthode de Boole,

avec $g(x) = 1$ dans tous deux cas. Comme nous l'avons vu, ces méthodes sont exactes pour les polynômes P_k de degré $k = N - 1$ pourvu que les x_i soient disposés régulièrement.

D'autres méthodes travailllent avec un espacement irrégulier: les points de collocation x_i et les poids sont imposés quelque soit la fonction f. Le tableau 5.1

regroupe quelques quadratures de ce type. Parmi les plus classiques, on trouve la quadrature de Gauss-Legendre. Dans cette méthode, $a = -b = -1$, $g(x) = 1$ et les abscisses x_i, encore appelées *noeuds*, sont les racines des polynômes de Legendre. Dans le cas $N = 2$, les noeuds et les poids sont

$$x_1 = -\frac{\sqrt{3}}{3} = -x_2 \qquad \text{et} \qquad w_1 = w_2 = 1 \tag{5.17}$$

et la méthode est exacte pour tout polynôme de degré $k \leq 3$. La figure 5.4 montre la construction graphique associée à la quadrature de Gauss-Legendre à 2 points. On peut étendre la méthode à un intervalle $[a, b]$ quelconque moyennant un changement de variable. On opère alors une *translation de Gauss-Legendre*.

5.7 Noyaux singuliers

Il est fréquent de rencontrer des intégrales dont les intégrandes (ou *noyaux*) présentent une ou plusieurs *singularités* sur un intervalle donné. Nous avons déjà signalé que des schémas ouverts ou semi-ouverts peuvent être utilisés dans ce cas, mais ils sont très imprécis. Le traitement correct d'une singularité nécessite que l'on connaisse sa forme (*explicite* ou *implicite*), sa nature et sa localisation. Par exemple, les intégrales

$$\int_a^{2a} \ln(x-a)dx, \qquad x \geq a \tag{5.18}$$

et

$$\int_a^{2a} \frac{dx}{x-a}, \qquad x \geq a \tag{5.19}$$

possèdent un noyau singulier localisé en $x = a$. Dans les deux cas, la singularité est explicite; la première est de nature logarithmique et la seconde de nature hyperbolique. Ici, la quadrature peut être calculée analytiquement. Numériquement, on pourrait utiliser des méthodes avec pas h adaptatif où l'on sera inévitablement tenté par un schéma semi-ouvert pour calculer

$$\lim_{h\to 0} \int_a^{a+h} \ldots dx \tag{5.20}$$

ou bien les méthodes de Gauss (voir le tableau 5.1).

Un exemple d'intégrale à singularié implicite et localiée (un cas fréquemment rencontré) est

$$I(a) = \int_a^1 x\mathcal{K}(x)dx \tag{5.21}$$

où $\mathcal{K}(x)$ est l'intégrale elliptique complète de première espèce. Rappelons que $\mathcal{K}(x)$ est une *fonction spéciale* définie par

$$\mathcal{K}(x) = \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - x^2 \sin\phi}} \tag{5.22}$$

et qui intervient souvent dans les problèmes à symétrie cylindrique. On sait que

$$\lim_{x\to 1} \mathcal{K}(x) = \ln 4 - \frac{1}{2}\ln\left(1 - x^2\right) \tag{5.23}$$

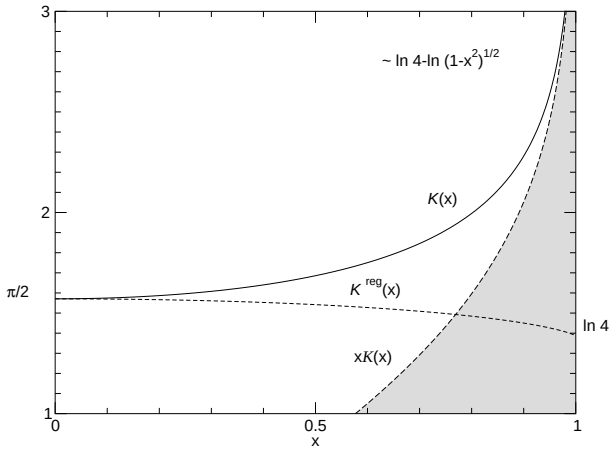


Figure 5.5: Intégrale elliptique complète de première espèce $\mathcal{K}(x)$ possédant une divergence logarthmique pour $x \to 1$, et la fonction régulière $\mathcal{K}^{\rm reg}(x)$.

rendant la singularié de nature logarithmique (voir la figure 5.5). Numériquement, on pourra procéder de la façon suivante. La nature de la singularité étant connue, on commence par former le noyau régulier $\mathcal{K}^{\rm reg}(x)$ par soustraction de la singularité, en posant

$$\mathcal{K}^{\rm reg}(x) = \mathcal{K}(x) + \ln(1-x^2) \tag{5.24}$$

et la relation (5.21) devient

$$I(a) = \int_a^1 x \left[\mathcal{K}^{\rm reg}(x) + \ln(1-x^2)\right] dx \tag{5.25}$$

La suite est relativement naturelle et directe. On sépare l'intégration en deux, pour obtenir successivement

$$\begin{aligned} I(a) &= \int_a^1 x\mathcal{K}^{\rm reg}(x)dx + \int_a^1 x\ln(1-x^2)dx \\ &= \int_a^1 x\mathcal{K}^{\rm reg}(x)dx - \frac{1}{2}\int_{1-a^2}^0 \ln u du, \qquad u = 1-x^2 \\ &= \int_a^1 x\mathcal{K}^{\rm reg}(x)dx - \frac{1}{2}\left[u(\ln u - 1)\right]_{1-a^2}^0 \\ &= \int_a^1 x\mathcal{K}^{\rm reg}(x)dx + \frac{1-a^2}{2}\left[\ln(1-a^2) - 1\right] \end{aligned} \tag{5.26}$$

où la partie $\int_a^1 x\mathcal{K}^{\rm reg}(x)dx$ possède par construction un noyau parfaitement régulier et pourra être calculée avce précision grâce aux méthodes standards. Il est impératif ici que la fonction $\mathcal{K}^{\rm reg}(x)$ soit calculée avec la plus haute précision (par exemple via des bibliothèques numériques).

5.8 Exercices et problèmes

■ Démontrez la relation (5.10).

■ Démontrez la relation (5.12). Pouvez-vous trouver un exemple simple où la méthode de Simpson est moins précise que celle du trapèze ?

■ Ecrivez la version composite de la règle de Simpson pour une fonction définie en N points. Comparez sa précision par rapport à la formule ne mettant en jeu que 3 points de l'intervalle $[a, b]$.

■ Trouvez le schéma relatif à la seconde règle de Simpson.

■ Etablissez les schémas correspondant à $k = 4$ et 5.

■ Etablissez la relation de récurrence ci-dessus pour le schéma de Simpson à 3 points.

■ Retrouvez les noeuds x_1, x_2 et les poids w_1 et w_2 de l'intégration de Gauss-Legendre à deux points.

■ Calculez numériquement l'intégrale $I(a)$ donnée par la relation (5.26). Pour $a = 0$, comparez à la valeur exacte $I(0) = 2G \simeq 1.831931\ldots$

Chapitre 6

Zéro d'une fonction. Systèmes non-linéaires

Lorsque l'on cherche à caractériser l'état d'équilibre d'un système, il est fréquent de rencontrer des équations vectorielles du type

$$F(\vec{X}) = 0 \tag{6.1}$$

où $\vec{X}$ désigne un vecteur à N composantes x_i et F est un ensemble de N fonctions f_i des N variables x_1, x_2, ..., x_N. Plus familiers peut-être sont les problèmes à une dimension (i.e. $N = 1$) du type

$$f(x) = 0 \tag{6.2}$$

où il n'y a qu'une seule variable en jeu, $x \equiv x_1$. En raison du nombre de degré de liberté élevé (i.e. supérieur à 2), la résolution d'une équation vectorielle du type (6.1) est généralement plus délicate que pour son analogue à une dimension. Dans les deux cas en revanche, la determination de la solution $\vec{R} = \{r_i\}_N$ s'effectue à partir d'un intervalle de recherche donné par des méthodes itératives dont l'initialisation se borne au choix d'un (voire de plusieurs) point(s) de départ $\vec{X}^{(0)}$ (voire $\vec{X}^{(1)}$, etc.). En fait, on ne trouve jamais la valeur qui annule exactement le système, mais une approximation. Comme l'illustre la figure 6.1, il peut exister plusieurs racines proches dans un intervalle donné. La convergence sera d'autant meilleure que le choix sera judicieux, c'est-à-dire que $\vec{X}^{(0)}$ sera déjà "suffisamment" proche de $\vec{R}$. Elle dépend également très fortement des propriétés de la fonction non seulement au voisinage de la racine cherchée mais aussi au voisinage du point de départ. Enfin, la convergence du schéma nécessite un critère, qui se fait soit sur la variable, soit sur la fonction, ou bien sur les deux.

6.1 Existence des solutions et convergence

6.1.1 Condition d'existence

A une dimension, la condition d'existence d'une solution (au moins) pour l'équation (6.2) peut s'énoncer de la façon suivante: si une fonction f est continue et qu'il existe un intervalle $[a, b]$ tel que

$$f(a) \times f(b) \leq 0 \tag{6.3}$$

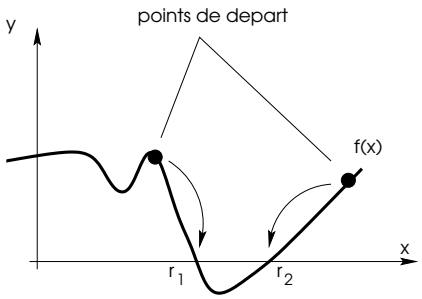


Figure 6.1: Racines r_1 et r_2 d'une fonction f de la variable x. Selon le point de départ, on tombe sur l'une ou sur l'autre.

alors il existe au moins une quantité réelle $r \in [a, b]$ telle que $f(r) = 0$ (r est la *racine* ou le *zéro* de la fonction). Notez que l'inégalité (6.3) ne traduit pas une condition nécessaire, et ne dit rien sur l'unicité de la racine (il y en a forcément un nombre impair). Par exemple, sur $[-1, 1]$, la fonction $y = x^2$ possède une racine alors que $f(-1) \times f(1) > 0$.

6.1.2 Critère de convergence

Comme nous l'avons signalé, on ne peut pas trouver numériquement la solution exacte r mais seulement à une approximation suffisamment bonne de cette racine $x^{(n)}$, grâce à un processus itératif que l'on aura délibéremment stoppé à une certaine étape n. Pour définir cette étape, il faut définir un critère de convergence. Si ce critère est mal choisi, l'approximation de la racine sera mauvaise, ce qui peut avoir des répercussions importantes sur la suite des calculs. Eliminons immédiatement tout critère qui porterait sur n: on ne peut pas prévoir à l'avance le nombre d'itérations qu'il faudra effectuer, du moins dans le cas le plus général. Cela pourrait par ailleurs conduire à une mauvaise identification de cete racine (par exemple, l'équation $x^2 + 1 = 0$ n'a pas de racine réelle et un schéma itératif pourrait donner à peu près n'importe quoi, comme l'absisse du minimum $x = 0$).

Le critère classique porte généralement sur la racine elle-même, c'est-à-dire que le processus itératif est interrompu lorsque

$$\Delta_n(x) \equiv |x^{(n)} - x^{(n-1)}| \leq C_x \tag{6.4}$$

ou mieux, si la racine est non nulle, lorsque

$$\epsilon_n(x) \equiv \left| \frac{x^{(n)} - x^{(n-1)}}{x^{(n-1)}} \right| \leq E_x \tag{6.5}$$

où C_x et E_x sont des constantes. Ces critères sont tout-à-fait recevables. Toutefois, ils n'assurent pas que l'image par f de l'intervalle $[r(1 - E_x), r(1 + E_x)]$ autour de la racine soit également "petit", ce qui peut-être problématique si l'on doit par ailleurs utiliser les valeurs de f. Ceci dépend du gradient de la fonction au voisinage de la racine, comme l'illustre la figure 6.2. Une procédure plus fiable consiste à mettre en place un critère portant aussi sur f, par exemple

$$\Delta_n(f) \equiv \left| f(x^{(n)}) - f(x^{(n-1)}) \right| \leq C_f \tag{6.6}$$

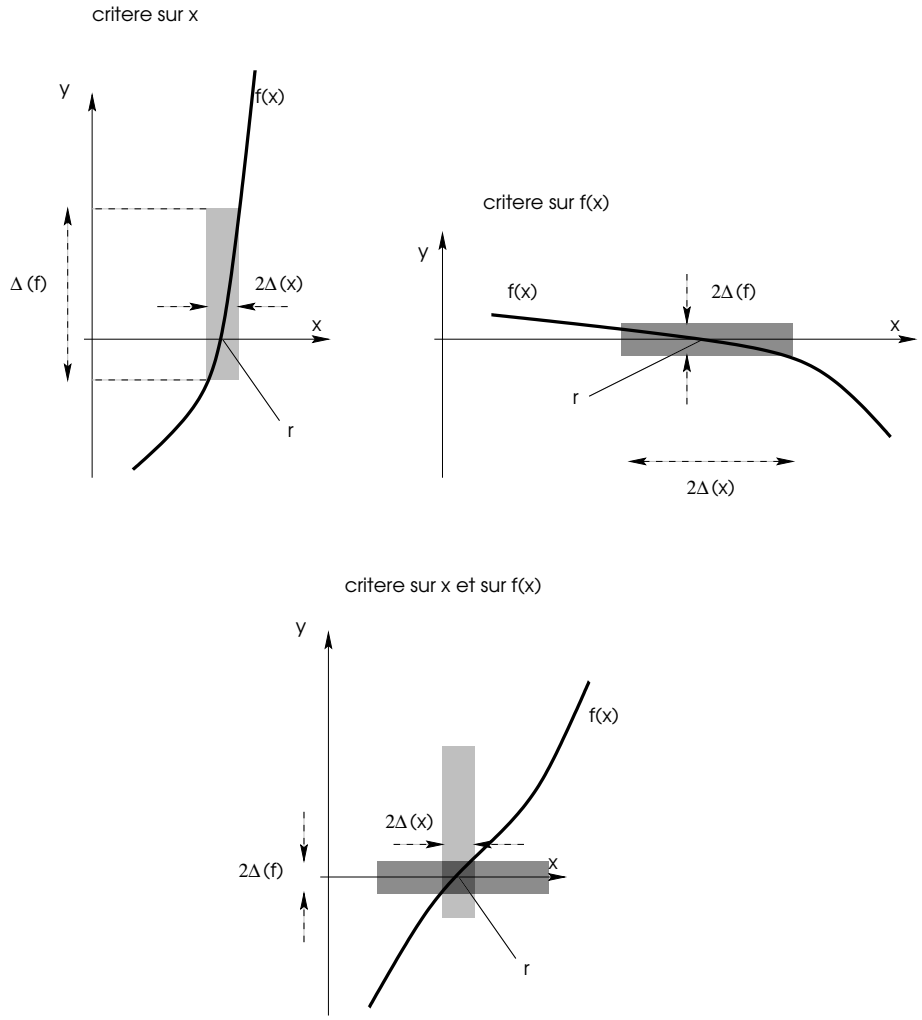


Figure 6.2: Dans la recherche de zéro d'une fonction, il est recommandé d'appliquer un critère de convergence portant à la fois sur la variable et sur la fonction.

ou bien si f est non nulle

$$\epsilon_n(f) \equiv \left| \frac{f(x^{(n)}) - f(x^{(n-1)})}{f(x^{(n-1)})} \right| \leq E_f \tag{6.7}$$

où C_f et E_f sont également des constantes à définir. On pourra par exemple choisir $C_f = \frac{1}{100}\left[\sup(f) - \inf(f)\right]$ sur $[a, b]$ et $E_f = 10^{-4}$.

6.1.3 Sensibilité

On ne peut pas ici calculer le nombre de conditionnement η tel qu'il a été définit au chapitre 2, car la fonction s'annule au niveau de la racine. Par ailleurs, on peut facilement se convaincre que la recherche de la racine sera d'autant plus facile que la fonction sera sensible à tout changement de la variable x, c'est-à-dire plus sont gradient (au niveau de la reacine) sera important. Par conséquent, le conditionnement (ou la sensibilité) dans la recherche de zéro est en fait inverse de celui relatif à l'évaluation d'une fonction. Le nombre pertinent sera plutôt

$$\frac{1}{|f'(r)|} \tag{6.8}$$

Notez que le conditionnement est particulièrement mauvais lorsque la tangente à la courbe au point r est presque horizonale, et en particulier pour une racine double (c'est-à-dire lorsque $f(r) = f'(r) = 0$). C'est par exemple le cas des fonctions du type $f(x) = x^p$, avec $p \geq 2$.

Pour un problème à plusieurs dimensions, le "bon" nombre de conditionnement est

$$\frac{1}{||J(\vec{R})||} \tag{6.9}$$

où J est la *matrice jacobienne*

$$J(\vec{R}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_N} \\ \frac{\partial f_N}{\partial x_1} & \frac{\partial f_N}{\partial x_2} & \cdots & \frac{\partial f_N}{\partial x_N} \end{pmatrix}_{\vec{R}} \tag{6.10}$$

Un problème se produit lorsque le déterminant de cette matrice est nul.

6.1.4 Taux de convergence

On peut contrôler la convergence du processus de recherche de racine en analysant le comportement de l'erreur d'une étape à l'autre. Pour cela, on forme la séquence

$$\Delta_1(x), \Delta_2(x), \Delta_3(x), \ldots, \Delta_{n-1}(x), \Delta_n(x)$$

Si $|\Delta_k(x)| \propto |\Delta_{k-1}(x)|$ approximativement, alors la convergence est dite *linéaire*. Dans ce cas, l'approche de la racine s'effectue par mise en place régulière des chiffres significatifs. Si $|\Delta_k(x)| \propto |\Delta_{k-1}^2(x)|$, la convergence est dite *quadratique* et le nombre de chiffre significatif double en moyenne. Plus généralement, on appelle *taux de convergence*, le nombre ζ tel que

$$\frac{|\Delta_k(x)|}{|\Delta_{k-1}(x)|} \sim C|\Delta_{k-1}(x)|^{\zeta-1} \tag{6.11}$$

où $|C| \leq 1$ (sinon l'erreur augmente d'une étape à l'autre et le processus diverge). Pour $\zeta = 1$, la convergence est linéaire; elle est quadratique pour $\zeta = 2$; cubique (c'est plus rare) pour $\zeta = 3$, etc.

6.2 Problèmes à une dimension

6.2.1 Méthode de bissection

Supposons que l'on connaisse un intevalle $[a, b]$ contenant la racine r de l'équation du type (6.2). La racine est dite *encadrée*. Si l'on coupe l'intervalle en deux, avec c comme abscisse du point milieu, deux situations peuvent se présenter: soit $f(a) \times f(c) < 0$, la racine se trouve alors sur l'intervalle $[a, c]$, soit $f(b) \times f(c) < 0$ et la racine se trouve alors sur l'intervalle $[c, b]$. Quelqu'il en soit, nous disposons d'un intervalle plus petit $[a^{(1)}, b^{(1)}]$ (s'est soit $[a, c]$ soit $[c, b]$) sur lequel nous pouvons a nouveau opérer un découpage en deux. Au bout de i *bissections*, l'intervalle *courant* a une largeur $\frac{b-a}{2^i}$ et la racine est localisée sur l'intervalle $[a^{(i)}, b^{(i)}]$. Lorsque la précision est jugée suffisante, à l'étape n, la racine vaut $r \approx a^{(n)} \approx b^{(n)}$. Un exemple de programmation de cette méthode est

```
epsilon=1.00D-04
WHILE (relerror>=epsilon) DO
  midvalue=a+b(-a)/2
  IF (f(a)*f(midvalue)<=0) THEN
    a=midvalue
  ELSE
    b=midvalue
  ENDIF
  relerror=ABS(b-a)/MAX(ABS(a),ABS(b))
ENDDO
PRINT*,'ROOT,ABS. & REL. ERROR::',(a+b)/2,relerror,(b-a)/2
```

La précision est de l'ordre de la largeur du demi-intervalle, soit $\frac{1}{2}(b^{(n)} - a^{(n)})$. En assimilant la fonction à une droite sur cet intervalle, on a même

$$r \approx \frac{f(a^{(n)})(a^{(n)} - b^{(n)})}{f(b^{(n)}) - f(a^{(n)})} + a^{(n)} \tag{6.12}$$

Ce schéma est très efficace. Il est même imparable. Il garantit toujours de trouver une racine. Si l'intervalle $[a, b]$ contient plusieurs racines, la méthode parviendra toujours a en isoler une. Mais elle est très lente à converger. En effet, le nombre d'itération nécessaire pour encadrer la racine à la précision E près est

$$n \approx 1.4 \times \ln \frac{b - a}{E} \tag{6.13}$$

Par exemple, pour un intervalle initial de largeur unité et une précision relative de 10^{-6}, on obtient, $n \sim 21$ itérations. Cela semble peu, mais d'autres méthodes donne un résultat aussi précis avec quelques itérations seulement.

Enfin, mentionnons qu'il peut être dangereux de chercher une racine avec comme seul critère le signe du produit $f(a) \times f(b)$, pour deux raisons au moins. D'une part, la fonction peut avoir une racine double[1]. D'autre part, bien que

[1] Une racine r est simple lorsque $f(r) = 0$, mais $f^{(n)}(r) \neq 0$ pour $n \geq 1$. Une racine r est double lorsque $f(r) = f'(r) = 0$, mais $f^{(n)}(r) \neq 0$ pour $n \geq 2$. Une racine r est multiple lorsque $f(r) = f'(r) = f^{(2)}(r) = \cdots = f^{(k)}(r) = 0$, mais $f^{(k)}(r) \neq 0$ pour $n \geq k + 1$.

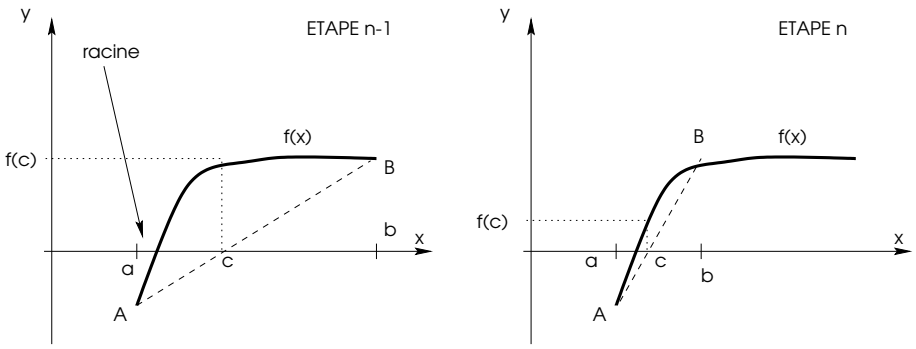


Figure 6.3: Dans la méthode des fausses positions, on cherche l'intersection du segment [AB] avec l'axe $y = 0$. Selon le signe des produits $f(a) \times f(c)$ et $f(c) \times f(b)$, on selectionne le nouvel intervalle de recherche, et ainsi de suite jusqu'à convergence.

l'on puisse avoir $f(a) \times f(b) < 0$, on peut se trouver en présence d'une fonction qui diverge à l'infini de part et d'autre d'une valeur $c \in [a, b]$ (c'est par exemple le cas de la fonction $f(x) = \frac{1}{x}$ en $x = c = 0$) et chercher une racine sur cet intervalle serait vain.

6.2.2 Méthode des "fausses positions"

C'est une méthode de résolution proche de la précédente car également basée sur la comparaison des signes de la fonction f aux deux bornes de l'intervalle de recherche. Toutefois, ici, au lieu de prendre le point milieu, on cherche le point $\mathrm{C}(c, 0)$, intersection du segment [AB] joingnant les points $\mathrm{A}(a, f(a))$ et $\mathrm{B}(b, f(b))$ et l'axe des abscisses, comme indiqué à la figure 6.3. On réitère le processus sur un intervalle plus petit $[a, c]$ ou $[c, b]$, selon le signes de $f(a)$, $f(c)$ et $f(b)$. A l'étape n, lorsque la taille de l'intervalle de recherche est "suffisamment" petit, la racine vaut $r \approx c^{(n)} \pm \frac{1}{2}(b^{(n)} - a^{(n)})$. On peut se convaincre de l'efficacité de cette méthode par rapport à la méthode de bissection (et donc à sa plus grande rapidité) sur une fonction linéaire par exemple.

6.2.3 Méthode du "point fixe"

La méthode du point fixe concerne la résolution d'équations du type

$$g(x) - x = 0 \tag{6.14}$$

qui constituent un sous-ensemble[2] de (6.2). En d'autres termes, il s'agit de trouver l'intersection d'une fonction g avec la première bissectrice des axes d'équation $y = x$. Le schéma associé est simple et intuitif: on part d'une abscisse initiale $x^{(0)}$ proche de la racine r cherchée puis l'on calcule $g(x^{(0)}) = x^{(1)}$; puis connaissant $x^{(1)}$, on calcule $g(x^{(1)}) = x^{(2)}$, et ainsi de suite jusqu'à ce que l'on obtienne une stabilisation du processus, c'est-à-dire $g(x^{(n)}) \approx x^{(n)}$ avec une

[2]Toute équation du type (6.2) peut être explicitement ramenée à une forme du type (6.14) en posant $g(x) = f(x) + x$; mais l'emploi de la méthode du point fixe serait dans ce cas une erreur.

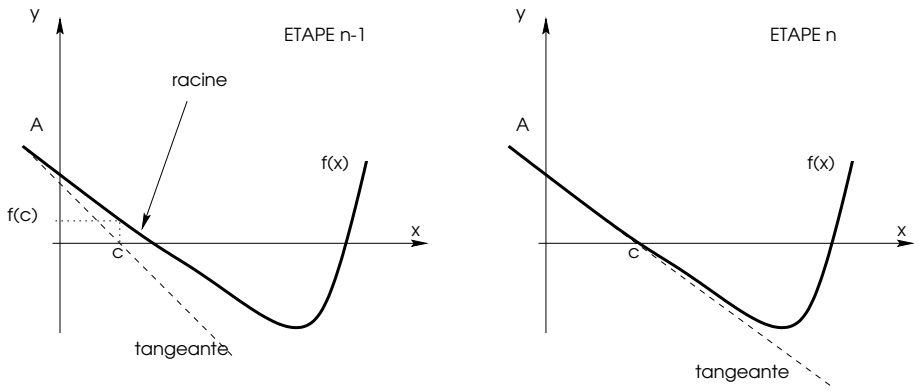


Figure 6.4: Construction graphique associée à la méthode des gradients.

précision donnée. La racine r cherchée n'est alors autre que $r \approx x_n$. On écrit donc ce schéma

$$x^{(n+1)} = g(x^{(n)}) \tag{6.15}$$

La méthode du point fixe n'est pas toujours stable (et donc utilisable). On peut en effet montrer que le schéma (6.15) ne converge que lorsque la condition

$$-1 \leq g'(r) \leq 1 \tag{6.16}$$

est satisfaite. Comme nous l'avons souligné plus haut, le choix du point de départ $x^{(0)}$ est crucial: il peut très bien conduire à une racine autre que celle souhaitée ou bien à une divergence du schéma. Notez que la relation (6.16) est une condition sur la dérivée en la solution, non au point de départ $x^{(0)}$.

6.2.4 Méthode des gradients

Si l'on a à faire à une racine simple pour f (voir la note 1), alors la solution de l'équation (6.2) coïncide avec la solution d'une équation du type (6.14) si g est définie par

$$g(x) = x - \frac{f(x)}{f'(x)}, \tag{6.17}$$

où g est la *fonction de Newton-Raphson*. L'intérêt de cette définition est immédiat: on se ramène ainsi à le méthode du point fixe (voir §6.2.3). La condition de convergence du schéma, déduite de (6.16), devient

$$\left| \frac{f(x)f''(x)}{f'^2(x)} \right|_r < 1 \tag{6.18}$$

L'interprétation graphique de la méthode des gradients est la suivante. En un point de départ $\mathrm{M}^{(0)}(x^{(0)})$, on construit la tangente à la courbe $y = f(x)$ et on cherche l'intersection de cette tangente avec l'axe des abscisses. Ceci définit un nouveau point $\mathrm{M}^{(1)}(x^{(1)})$, plus proche de la racine que $\mathrm{M}^{(0)}$. On trace alors la tangente à la courbe en ce point, puis son intersection avec l'axe $y = 0$ défini un troisième point, et ainsi de suite. On peut montrer que cette démarche produit le schéma numérique suivant

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}, \tag{6.19}$$

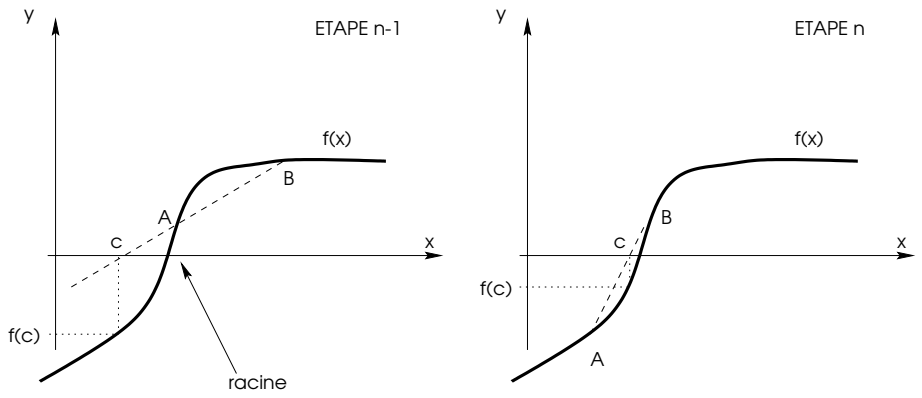


Figure 6.5: Illustration de la méthode des sécantes.

ce qui nous ramène bien à la fonction g définie plus haut. L'ordre de la méthode est quadratique (i.e. $\zeta = 2$), c'est-à-dire que le nombre de chiffres significatifs double à peu près d'une itération à l'autre.

La méthode des gradients, également appelée méthode de Newton-Raphson, ne converge pas dans certaines situations, par exemple si la fonction a une racine multiple (on peut partiellement y remédier en considerant, avec précaution, la fonction $x \times f(x)$) ou bien si $x^{(0)}$ est mal choisi (trop loin de la racine cherchée). Plus rarement, on peut rencontrer des séquences cycliques, des séquences périodiques ou même divergentes.

Notez que la relation (6.17) fait apparaître le nombre de conditionnement η définit auparavant

$$g(x) = x\left(1 - \frac{1}{\eta}\right), \tag{6.20}$$

d'où

$$g'(x) = \left(1 - \frac{1}{\eta}\right) + \frac{x}{\eta^2}\frac{d\eta}{dx} \tag{6.21}$$

Nous en déduisons que la condition de convergence de la méthode est liée à sa sensibilité au niveau de la racine et en particulier à $\frac{d\ln|\eta|}{d\ln x}$.

Exemple

Soit à trouver le zéro de la fonction $f(x) = \sin x - \cos x$ sur $[0, 1]$. Le schéma itératif est donc

$$x^{(k+1)} = x^{(k)} - \frac{\sin x^{(k)} + \cos x^{(k)}}{\sin x^{(k)} - \cos x^{(k)}} \tag{6.22}$$

Pour $x_0 = 0$, la résolution numérique donne

```
k  x                       f(x)
0 +0.0000000000000000D+00 +0.0000000000000000D+00
1 +1.0000000000000000D+00 -1.0000000000000000D+00
2 +7.8204190153913800D-01 +3.0116867893975678D-01
3 +7.8539817599970183D-01 -4.7464621278041701D-03
```

```
4 +7.8539816339744828D-01 +1.7822277877512036D-08
5 +7.8539816339744828D-01 -4.3313876790795902D-17
```

Pour $x_0 = 1$,on obtient

```
k  x                        f(x)
0 +1.0000000000000000D+00 +0.0000000000000000D+00
1 +7.8204190153913800D-01 +3.0116867893975678D-01
2 +7.8539817599970183D-01 -4.7464621278041701D-03
3 +7.8539816339744828D-01 +1.7822277877512036D-08
4 +7.8539816339744828D-01 -4.3313876790795902D-17
5 +7.8539816339744828D-01 -4.3313876790795902D-17
```

Notez que seulement 5 itérations sont nécessaires pour atteindre la précision de la machine et que la convergence est bien quadratique.

6.2.5 Méthode des sécantes

Cette méthode est identique à la méthode des gradients à ceci près que l'on choisit non pas la tangente à la courbe en un point initial mais une droite définie par 2 points $\mathrm{M}^{(0)}(x^{(0)})$ et $\mathrm{M}^{(1)}(x^{(1)})$ de la courbe. On utilise alors le point d'intersection de cette droite avec l'axe $y = 0$, soit le point $\mathrm{M}^{(2)}(x^{(2)})$ comme nouveau point et l'on réitère l'opération avec les point $\mathrm{M}^{(1)}$ et $\mathrm{M}^{(2)}$ (voir la figure 6.5). Le schéma associé est

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})(x^{(n)} - x^{(n-1)})}{f(x^{(n)}) - f(x^{(n-1)})}, \tag{6.23}$$

Cette méthode présente l'avantage de ne faire appel qu'à l'évaluation de la fonction f et non à sa dérivée qui n'est pas toujours accessible analytiquement. Toutefois, elle ne guarantit pas que les points encadrent toujours la solution. Elle peut de se fait diverger dans certains cas.

6.2.6 Méthode de Müller

Cette méthode est semblable à la méthode des sécantes mais avec 3 points. On commence par construire une parabole passant par 3 points de la courbe $\mathrm{M}^{(0)}(x^{(0)}, f(x^{(0)}))$, $\mathrm{M}^{(1)}(x^{(1)}, f(x^{(1)}))$ et $\mathrm{M}^{(2)}(x^{(2)}, f(x^{(2)}))$ préalablement choisis et l'on cherche l'intersection de la parabole avec l'axe horizontal. Ceci donne un nouveau point $\mathrm{M}^{(3)}(x^{(3)}, f(x^{(3)}))$. On recommence le processus avec $\mathrm{M}^{(1)}$, $\mathrm{M}^{(2)}$ et $\mathrm{M}^{(3)}$ jusqu'à convergence. Le schéma faisant passer du triplet ($\mathrm{M}^{(n-2)}$, $\mathrm{M}^{(n-1)}$, $\mathrm{M}^{(n)}$) au triplet suivant ($\mathrm{M}^{(n-1)}$, $\mathrm{M}^{(n)}$, $\mathrm{M}^{(n+1)}$) est

$$x^{(n+1)} = x^{(n)} - \frac{2f(x^{(n)})}{\beta^2 \pm \sqrt{\beta^2 - 4\alpha f(x^{(n)})}} \tag{6.24}$$

où

$$\alpha = \frac{f(x^{(n-2)}) - f(x^{(n)})}{(x^{(n-2)} - x^{(n)})(x^{(n-2)} - x^{(n-1)})} - \frac{f(x^{(n-1)}) - f(x^{(n)})}{(x^{(n-1)} - x^{(n)})(x^{(n-2)} - x^{(n-1)})} \tag{6.25}$$

et

$$\beta = \frac{[f(x^{(n-1)}) - f(x^{(n)})](x^{(n)} - x^{(n-2)})^2}{(x^{(n-1)} - x^{(n)})(x^{(n-2)} - x^{(n)})(x^{(n-2)} - x^{(n-1)})} - \frac{[f(x^{(n-2)} - f(x^{(n)})](x^{(n-1)} - x^{(n-2)})^2}{(x^{(n-1)} - x^{(n)})(x^{(n-2)} - x^{(n)})(x^{(n-2)} - x^{(n-1)})} \tag{6.26}$$

où il conviendra de choisir le signe correspondant à la solution la plus proche de la racine cherchée.

Dans la méthode de Müller, la fonction interpolante est quadratique en x. On peut aussi chercher une fonction interpolante quadratique en y. C'est l'esprit de la méthode de Van Wijngaarden-Bekker. Sa convergence est également quadratique. Notez que les points n'encadrent pas nécessairement la racine mais peuvent être tous au dessous ou au dessus de l'axe $y = 0$. De ce fait, des séquences divergentes peuvent apparaître dans certains cas.

On peut généraliser ce type de méthode en utilisant un interpolant polynômial de degré supérieur à 2.

6.3 Systèmes non-linéaires

Nous nous sommes intéressé jusqu'à présent à des méthodes permettant de trouver le zéro d'une fonction d'une variable. Pour résoudre une équation du type (6.1), on peut en principe utiliser les méthodes décrites plus haut, généralisées à N fonctions couplées des variables x_1, x_2, ..., x_N. La convergence est souvent plus délicate.

6.3.1 Généralisation de la méthode du point fixe

Pour la méthode du point fixe

$$G(\vec{X}) = \vec{X}, \tag{6.27}$$

le schéma est

$$\begin{cases} x_1^{(n+1)} = g_1\left(x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}\right) \\ x_2^{(n+1)} = g_2\left(x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}\right) \\ x_3^{(n+1)} = g_3\left(x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}\right) \\ \quad \ldots \\ x_N^{(n+1)} = g_N\left(x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}\right) \end{cases}$$

Une condition nécessaire à la convergence de ce schéma est que le *rayon spectral* de la matrice jacobienne associée à G soit de norme inférieure à l'unité, soit

$$\sum_{k=1,N} \left|\frac{\partial g_i}{\partial x_k}\right| < 1, \qquad i = 1, N \tag{6.28}$$

qui n'est que la généralisation à N dimensions de la condition (6.16) rencontrée plus haut. Le taux de convergence est 2 si cette quantité est nulle.

6.3.2 Méthode de Seidel

On peut améliorer la convergence du schéma précédent de la façon suivante: on calcule $x_2^{(n+1)}$ non pas à partir de $x_1^{(n)}$ mais à partir de $x_1^{(n+1)}$, soit

$$x_2^{(n+1)} = g_2(x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}) \tag{6.29}$$

puis $x_3^{(n+1)}$ à partir de $x_1^{(n+1)}$ et $x_2^{(n+1)}$, soit

$$x_3^{(n+1)} = g_2(x_1^{(n+1)}, x_2^{(n+1)}, x_3^{(n)}, \ldots, x_N^{(n)}) \tag{6.30}$$

et ainsi de suite jusqu'au dernier. Le schéma devient donc

$$\begin{cases} x_1^{(n+1)} = g_1(x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}) \\ x_2^{(n+1)} = g_2(x_1^{(n+1)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}) \\ x_3^{(n+1)} = g_3(x_1^{(n+1)}, x_2^{(n+1)}, x_3^{(n)}, \ldots, x_N^{(n)}) \\ \quad \ldots \\ x_N^{(n+1)} = g_N(x_1^{(n+1)}, x_2^{(n+1)}, x_3^{(n+1)}, \ldots, x_{N-1}^{(n+1)}, x_N^{(n)}) \end{cases}$$

C'est la *méthode de Seidel* du point fixe. Elle est sensée conduire à une amélioration du traitement car elle "anticipe" dans $n-1$ directions.

6.3.3 Méthode de Newton généralisée

Autre adaptation possible à N dimensions, celle de la méthode des gradients exposée au §6.2.4. Elle fait intervenir la matrice jacobienne du système, les dérivées partielles étant évaluées au point courant $\mathbf{X}^{(n)}$. Le schéma itératif est

$$\vec{X}^{(n+1)} = \vec{X}^{(n)} - J^{-1}(\vec{X}^{(n)})F^{(n)} \tag{6.31}$$

dont l'équivalent à une dimension est la relation (6.19). Il met en jeu une inversion de matrice. Comme nous l'avons souligné, le cas $||J(\vec{X}^{(n)}|| = 0$ est problématique. Notez que dans de nombreux cas pratiques, on n'a pas accès au dérivées analytiques; il faut allors procéder numériquement pour estimer J.

6.4 Exercices et problèmes

■ Calculez (à la main) à 10^{-4} près la racine r positive de l'équation $g(x) = x$ pour $g(x) = \frac{x}{2} + \frac{1}{x}$.

■ Combien d'itérations sont nécessaires pour trouver la racine à ϵ près d'une fonction du type $f(x) = ax + b$ près avec la méthode des fausses positions ? Même question si f est un polynôme de degré $n > 1$.

■ Quels sont le taux de convergence ζ et la constante C pour la méthode de bissection ? et pour la méthode des fausses positions ? et pour la méthode des sécantes ? et pour la méthode de Müller ?

■ Quels sont le taux de convergence ζ et la constante C pour la méthode du point fixe ? Même question lorsque $g'(r) = 0$.

■ Quels sont le taux de convergence ζ et la constante C pour la méthode des gradients en présence d'une racine multiple ?

■ Calculez (à la main) à 10^{-4} près la racine r positive de l'équation $g(x) = x$ pour $g(x) = \frac{x}{2} + \frac{1}{x}$ et comparer avec la méthode de bissection.

■ Démontrez la relation (6.19)

■ Montrez que la fonction $g(x) = \frac{x}{2} + \frac{1}{x}$ considérée dans les exercices précédents est la fonction de Newton-Raphson d'une certaine fonction f que l'on explicitera. En déduire que la méthode du point fixe pour une fonction $g(x) = \frac{x}{2} + \frac{A}{2x}$ est un moyen de déterminer $\sqrt{A}$.

■ Démontrez la relation (6.23)

■ Etablissez les schémas de Müller et de Van Wijngaarden-Bekker.

Chapitre 7

Equations aux dérivées ordinaires

7.1 Définitions

Les *équations différentielles ordinaires*, ou ODEs en anglais (pour "ordinary differential equations"), sont des équations mettant en jeu une fonction y de la variable x et au moins l'une de ses dérivées. Elles sont de la forme

$$\mathcal{F}\left(\frac{d^n y}{dx^n}, \frac{d^{n-1} y}{dx^{n-1}}, \ldots, y' \equiv \frac{dy}{dx}, y, g(x), x\right) = 0 \tag{7.1}$$

où g est une fonction de x (n est l'*ordre* de l'équation). La forme la plus simple sur laquelle nous discuterons largement dans ce chapitre est

$$\frac{dy}{dx} = f(y, x) \tag{7.2}$$

où f englobe toute forme explicite de x et de y. Elle représente en fait, dans le plan (x, y), le *champ des pentes* de la fonction y (voir la figure 7.1).

On trouve également des systèmes d'équations différentielles ordinaires où il y a $\mathcal{F}_i$ fonctions du type (7.1). Un tel système s'écrit

$$\begin{cases} \mathcal{F}_1\left(\frac{d^{n_1} y_1}{dx^{n_1}}, \frac{d^{n_1-1} y_1}{dx^{n_1-1}}, \ldots, y_1' \equiv \frac{dy_1}{dx}, y_1, g_1(x), x\right) = 0 \\ \mathcal{F}_2\left(\frac{d^{n_2} y_2}{dx^{n_2}}, \frac{d^{n_2-1} y_2}{dx^{n_2-1}}, \ldots, y_2' \equiv \frac{dy_2}{dx}, y_2, g_2(x), x\right) = 0 \\ \ldots \\ \mathcal{F}_N\left(\frac{d^{n_N} y_N}{dx^{n_N}}, \frac{d^{n_N-1} y_N}{dx^{n_N-1}}, \ldots, y_N' \equiv \frac{dy_N}{dx}, y_N, g_N(x), x\right) = 0 \end{cases}$$

Son ordre est l'ordre le plus élevé, soit $\sup\{n_i\}$. Autre variété possible de problème, celui des systèmes d'équations différentielles ordinaires à plusieurs

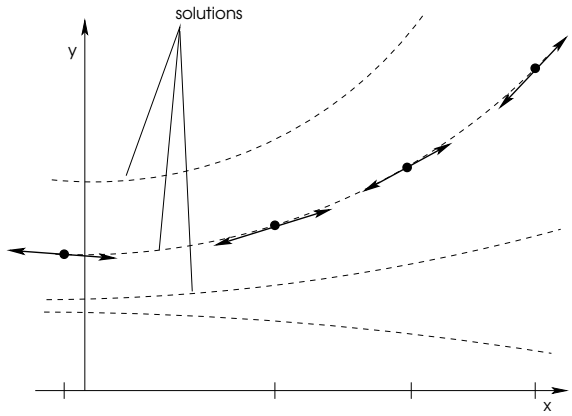


Figure 7.1: Illustration de la notion de champ de pentes.

variables indépendantes du type

$$\begin{cases} \mathcal{F}_1\left(\dfrac{d^{n_1}y_1}{dx_1^{n_1}}, \dfrac{d^{n_1-1}y_1}{dx_1^{n_1-1}}, \ldots, y_1' \equiv \dfrac{dy_1}{dx_1}, y_1, g_1(x_1), x_1\right) = 0 \\ \mathcal{F}_2\left(\dfrac{d^{n_2}y_2}{dx_2^{n_2}}, \dfrac{d^{n_2-1}y_2}{dx_2^{n_2-1}}, \ldots, y_2' \equiv \dfrac{dy_2}{dx_2}, y_2, g_2(x_2), x_2\right) = 0 \\ \ldots \\ \mathcal{F}_N\left(\dfrac{d^{n_N}y_N}{dx_N^{n_N}}, \dfrac{d^{n_N-1}y_N}{dx_N^{n-1}}, \ldots, y_N' \equiv \dfrac{dy_N}{dx_N}, y_N, g_N(x_N), x_N\right) = 0 \end{cases}$$

L'aspect complexe de ce système est trompeur. En raison du caratère indépendant des variables (chaque fonction y_i est toujours dérivée par rapport à la même variable x_i et ne dépend que d'elle), on se retrouve face à N problèmes découplés du type 7.1 et on utilisera de ce fait les mêmes méthodes que pour résoudre une équation à variable unique.

En contraste avec les équations aux dérivées partielles ou PDEs (pour "partial differential equations" en anglais), les équations différentielles ordinaires (ou les systèmes) ne font intervenir qu'une seule variable et sont, de ce fait, beaucoup plus simples à résoudre. Du point de vue purement mathématique, plusieurs fonctions y (souvent une infinité) peuvent être solution de l'équation (7.1). Toutefois, les solutions mathématiques ne correspondront pas forcément à la situation physique recherchée. Pour le Physicien, il s'agira de sélectionner les fonctions qui satisfont un certain nombre de contraintes appelées *conditions aux limites*, par exemple une condition sur la valeur de la fonction et/ou la valeur de sa ou ses dérivée(s) en certains points de intervalle d'étude $[a, b]$. Pour une équation différentielle ordinaire d'ordre n (ou d'un système de n équations d'ordre 1), il faut nécessairement n conditions aux limites pour dfinir une solution $y(x)$. Compte-tenu de ces contraintes imposées par le modèle physique, il pourra finalement ne pas exister de solution au problème posé ou, mieux, en exister une et une seule. Dans certains cas, les solutions resteront *dégénérées* (i.e. multiples).

7.2 Le "splitting"

Quand l'*ordre* n de l'équation différentielle est strictement supérieur à un, on aura fortement intérêt (du moins dans l'objectif d'une recherche de solutions numériques) à former un système de n équations différentielles du *premier ordre*. Ceci est toujours réalisable en faisant appel à $n-1$ fonctions intermédiaires y_j, par exemple en posant

$$\frac{dy_{j-1}}{dx} \equiv y_j, \qquad j = 1, n-1 \tag{7.3}$$

avec $y_0 \equiv y$. On réalise alors un *splitting*. Par exemple, l'équation *non-linéaire* du second ordre à *coefficients non constants*

$$f'' + f' - xf = \sin(x) \tag{7.4}$$

se ré-écrira sous la forme d'un système couplé

$$\begin{cases} y_1' = xy_0 - y_1 + \sin(x) \\ y_0' = y_1 \end{cases}$$

avec $y_0 \equiv f$.

On peut voir le problème de résolution d'une équation différentielle ordinaire comme un problème d'intégration et l'on peut légitimement envisager d'utiliser les méthodes dédiées au calcul des *quadratures*. En effet, la relation (7.2) s'écrit aussi

$$y(x) = \int_a^x f(x', y)dx' + y(a), \tag{7.5}$$

Toutefois, une différence fondamentale est que la fonction y n'est pas connue au préalable. La forme implicite de la relation (7.6) (f dépend de y !) rend en principe impossible le calcul directe de y de cette manière. Pour majorité des techniques, la construction y se fera de proche en proche, à partir d'un *point de départ* $(a, y(a))$ ou d'un *point d'arrivée* $(b, y(b))$. Tout la difficulté sera alors de calculer le plus précisément possible l'*incrément* δ_i, permettant depasser de y_i à y_{i+1}, soit

$$y_{i+1} = y_i + \delta_i \tag{7.6}$$

7.3 Les conditions aux limites

7.3.1 Valeurs initiales et conditions aux contours

On distingue habituellement trois classes de problèmes selon l'endroit de l'intervalle d'étude $[a, b]$ où s'appliquent les conditions aux limites (voir la figure 7.2):

- les problèmes aux *conditions initiales* ou encore IVPs (pour "Initial Value Problems" en anglais). Dans ce cas, les conditions aux limites sont imposées au même endroit, en $x = a$ ou en $x = b$.

- les problèmes aux *conditions aux contours* encore appelés (T)BVPs (pour "(Two) Boundary Value problems" en anglais). Les contraintes sont cette fois données à deux endroits différents (au moins), en a et en b.

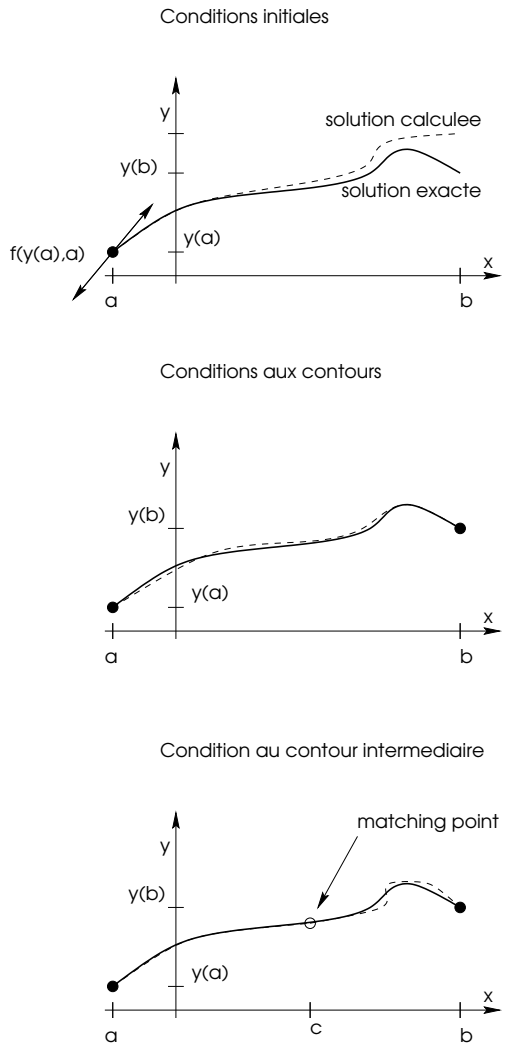


Figure 7.2: Trois types de conditions aux limites dans les problèmes aux dérivées ordinaires: conditions initiales (*en haut*), conditions aux (deux) contours (*au milieu*) et condition au contour intermédiaire (*en bas*).

- les problèmes avec *condition au contour intermédiaire* où une condition est donnée à l'intérieur de l'intervalle d'intégration en un point d'abscisse $x = c \in [a, b]$ appelé *matching point*. On pourrait classer ce type de problème dans l'une des deux catégories précédentes, car il faudra en pratique découper le problèmes en deux (l'un pour $x \in [a, c]$ et l'autre pour $x \in [c, b]$) et assurer leur jonction correcte en c.

Notez que ces distinctions n'ont de sens que si l'équation est au moins du second ordre.

Il est important de réaliser que, comme dans toute approche numérique, la résolution de problème, aussi propre soit-elle, ne donnera jamais la solution exacte, mais une *solution approchée*, entachée en chaque point d'une certaine erreur $\Delta y(x)$. Le choix de la méthode d'intégration sera alors guidé par la recherche d'un compromis entre rapidité et précision. De ce point de vue, dans un problème aux conditions initiales, on aura facilement tendance à s'écarter de la solution exacte (généralement inconnue), d'autant plus que l'intervalle d'intégration sera grand et que la fonction présentera de fortes variations. En raison de l'accumulation inévitable des erreurs, la précision ne pourra que se dégrader au fil de l'intégration. Au contraire, dans un problème aux conditions aux contours, la solution étant contrainte de part et d'autre de l'intervalle $[a, b]$, les écarts à la solution exacte seront généralement moindres, et les erreurs seront soit localisées dans les régions de fortes variations (cas des méthodes de tir; voir §7.6), soit distribuées sur tout l'intervalle (cas des méthodes employant les différences finies; voir §7.7). Avec une (ou plusieurs) condition(s) intermédiaire(s), on augmente les contraintes et l'on limite l'accumulation des erreurs (on permet à la solution de repartir sur des bases saines à chaque matching point). Ces considérations sont illustrées à la figure 7.2.

7.3.2 Conditions de Dirichlet et de Neumann

Les conditions aux limites qui portent sur la fonction sont dites de *conditions de Dirichlet*; lorsqu'elles portent sur la dérivée, on parle de conditions de *conditions de Neumann*. On peut bien-sûr rencontrer des situations où les deux types de conditions co-existent. Une condition aux contours s'écrit de manière générale

$$\mathcal{C}\left(y, \frac{dy}{dx}, x, c(x)\right) = 0 \qquad \text{en } x = a, b \text{ ou } c \tag{7.7}$$

où $c(x)$ est une fonction quelconque.

7.4 Méthodes "mono-pas"

7.4.1 Méthode d'Euler

La méthode d'intégration la plus intuitive et la plus simple aussi est la *méthode d'Euler* qui reprend la notion de taux d'accroissement sans passage à la limite, soit

$$y_{i+1} = y_i + hf(y_i, x_i) \tag{7.8}$$

où l'on a posé $h = x_{i+1} - x_i$. On peut évidemment voir la relation (7.8) comme le développement de Taylor de y à l'ordre le plus bas. Elle implique un *schéma aux différences finies*. L'incrément est très simple à calculer: $\delta_i = hf(y_i, x_i)$. En

partant de $x_0 \equiv a$ et en répétant ce schéma N fois, on parvient alors à $x_N \equiv b$ après avoir calculé successivement y_1, y_2, ..., y_N, sachant y_0 (la condition initiale). Notez qu'il n'est pas spécifié ici que les *points de la grille*, c'est-à-dire les x_i, soient tous équidistants. On peut en effet choisir, selon la variation locale de la fonction y que l'on est en train de calculer[1], de réduire l'intervalle, ou de l'agrandir. On réalise alors une méthode au *pas adaptatif*.

La méthode d'Euler n'a pas bonne réputation car elle propage très bien les erreurs et les amplifie généreusement. On peut s'en convaincre en estimant, du moins grossièrement, l'erreur totale Δy produite en $x = b$ à l'issue de l'intégration, en considérant que cette erreur est donnée par l'erreur de troncature à chaque pas, soit

$$\Delta y \sim \frac{h^2}{2} \sum_{i=1}^{N} f'(x_i, y_i) \approx \frac{Nh^2}{2} \langle f'(x,y) \rangle_{[a,b]} \tag{7.9}$$

où $\langle f'(x,y) \rangle_{[a,b]}$ est la pente moyenne de f' sur l'intervalle d'étude. Il est simple de voir par cette analyse très approchée que si la fonction présente une variation notable, l'erreur croît linéairement avec la distance d'intégration, à pas fixe. Toutefois, la méthode d'Euler est une méthode simple, de précision modulable ($\Delta y \propto h^2$), qui conviendra si la fonction n'est pas trop variable et si l'intervalle de variation n'est pas trop grand. Dans bien des cas, elle suffira, au moins en première approximation, pour obtenir une première idée des solutions d'un problème.

7.4.2 Méthode de Heun

Dans cette méthode (encore appelée *méthode de Newton modifiée*), on évalue l'incrément en deux étapes. La première étape consiste à écrire

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} f(y_i, x_i) dx \tag{7.10}$$

en estimant l'intégrale par la *méthode des trapèzes*, soit

$$y_{i+1} = y_i + \frac{h}{2} \left(f(y_i, x_i) + f(y_{i+1}, x_{i+1}) \right) \tag{7.11}$$

Cette formule est une formule *implicite* en y_{i+1} et elle n'est généralement pas inversible. De ce fait, on ne connait encore pas y_{i+1} à ce stade. On trouve une valeur approchée de $f(y_{i+1}, x_{i+1})$ grâce à la méthode d'Euler. C'est la seconde étape. On a

$$p_{i+1} = y_i + h f(y_i, x_i) \tag{7.12}$$

$$\approx y_{i+1} \tag{7.13}$$

La méthode de Heun met en jeu ce que l'on appelle classiquement un schéma de type *prédicteur/correcteur*. Le schéma *prédicteur* donne une estimation de y_{i+1} qui permet d'intuiter ou de prédire la valeur de la fonction en x_{i+1}. Le schéma *correcteur* associé est donné par la relation (7.11). Globalement, on a donc

$$y_{i+1} = y_i + \frac{h}{2} \left[f(y_i, x_i) + f(y_i + h f(y_i, x_i), x_{i+1}) \right] \tag{7.14}$$

[1] On pourra utiliser la dérivée seconde de y, soit f', pour tracer ces zones de forts gradients.

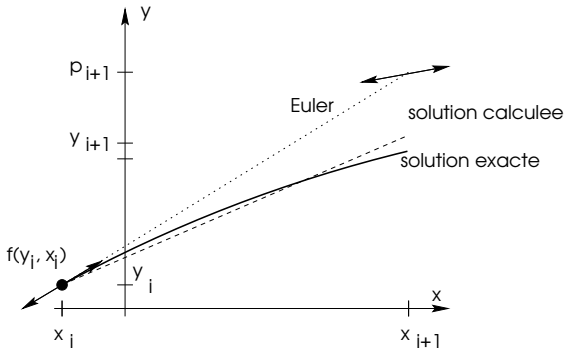


Figure 7.3: Dans la méthode de Heun, la fonction y est d'abord estimée en $x_i + h$ grâce à la méthode d'Euler, soit $y_{i+1} \approx p_{i+1}$. La valeur retenue pour y_{i+1} est alors calculée par la méthode des trapèzes à partir de y_i et de p_{i+1}.

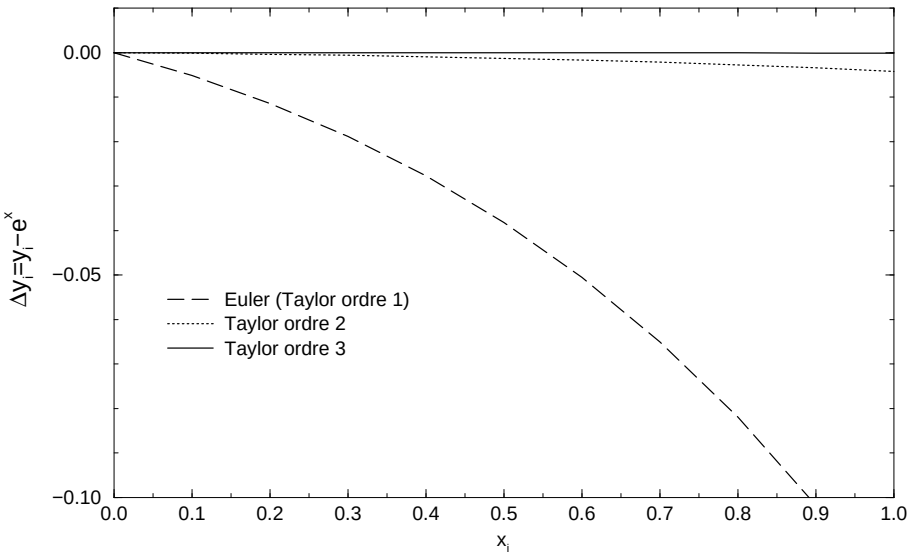


Figure 7.4: Comparaison de la méthode d'Euler et des méthode de Taylor d'ordre 2 et 3 pour l'équation telle que $f(y, x) = y$, avec $y(0) = 1$. Est donné ici l'erreur absolue par rapport à la solution exacte $y = e^x$ sur [0,1] pour une pas $h = 0.1$.

7.4.3 Méthode des séries entières

On voit que l'erreur commise dans l'écriture de la relation (7.8) correspond à l'omission des termes d'ordre $n \geq 2$ de la série de Taylor associée et dépend donc des dérivées seconde, troisième, etc. On a en effet, en tout rigueur

$$y_{i+1} = y_i + f(y_i, x_i)h + \frac{h^2}{2}\left.\frac{df}{dx}\right|_{x_i,y_i} + \sum_{n=3}^{\infty} \frac{h^n}{n!}\left.\frac{d^n y}{dx^n}\right|_{x_i,y_i} \tag{7.15}$$

Par conséquent, la méthode d'Euler n'est rigoureusement exacte que lorsque la fonction f est une constante, c'est-à-dire, en pratique, jamais (sinon, on peut vraiment se passer de méthode numérique !). L'idée de la méthode de Taylor est simple: comme f est une fonction de y, f' est une fonction de y', donc de f que l'on connaît. De même, f'' est une fonction de y' et de f', donc de f. De manière générale, on a

$$\frac{d^n f}{dx^n} = \left(\frac{\partial}{dx} + \frac{dy}{dx}\frac{\partial}{dy}\right)^n f(y,x) \tag{7.16}$$

$$= \left(\frac{\partial}{dx} + f\frac{\partial}{dy}\right)^n f(y,x) \tag{7.17}$$

ce qui permet de produire un schéma du même type que celui d'Euler, mais incluant autant de terme que l'on veut. On augmente ainsi considérablement la précision de l'estimation. Prenons un exemple. Pour résoudre $y' = x \sin y$ par cette méthode, on calcule d'abord $f' = y''$, soit

$$\begin{aligned} y'' &= \sin y + xy' \cos y \\ &= \sin y + x(x \sin y)\cos y \\ &= \sin y(1 + x^2 \cos y) \equiv f'(x,y) \end{aligned} \tag{7.18}$$

où l'on a remplacé y' par son expression, puis $f'' = y^{(3)}$. De la même façon, on trouve

$$y^{(3)} = \sin y \left[x^3 \sin y + (3x + x^3)\cos y\right] \equiv f''(x,y) \tag{7.19}$$

et aisni de suite jusqu'à l'ordre voulu. Reste à écrire le développement de Taylor associé. En l'occurrence, à l'ordre 3 inclus, on obtient le schéma suivant

$$\begin{aligned} y_{i+1} = y_i + hx_i \sin y_i + \frac{h^2}{2}\sin y_i(1 + x_i^2 \cos y_i) \\ + \frac{h^3}{6}\sin y_i \left[x_i^3 \sin y_i + (3x_i + x_i^3)\cos y_i\right] \end{aligned} \tag{7.20}$$

La méthode de Taylor est assez efficace (voir la figure 7.4). A l'ordre 2, elle coincide avec la méthode de Heun. Elle ne peut en revanche être mise en place que lorsque l'on peut effectivement calculer la dérivée d'ordre n de la fonction y, ce qui n'est pas toujours le cas.

7.4.4 Méthodes de Runge-Kutta

L'idée générale qui sous-tend les méthodes de Runge-Kutta d'ordre N repose sur la possibilité d'exprimer l'incrément $y_{i+1} - y_i$ sous la forme suivante

$$y_{i+1} - y_i = h\sum_{j=1}^{N} w_j f_j \tag{7.21}$$

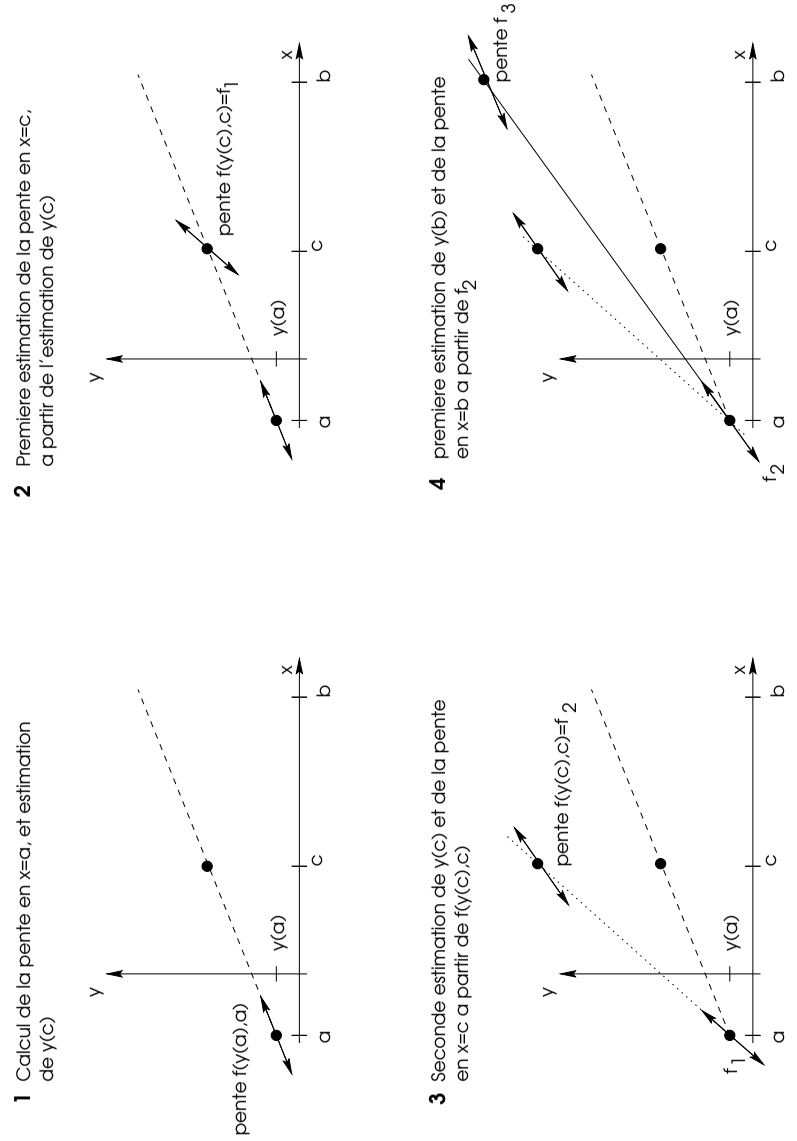


Figure 7.5: Interprétation graphique de la méthode de Runge-Kutta d'ordre 4.

où $f_j = f(Y_j, x_i + h\alpha_j)$, avec $0 \leq \alpha_j \leq 1$ et

$$Y_j = y_i + \sum_{k=1}^{j-1} \beta_{jk} Y_k \tag{7.22}$$

En d'autres termes, on estime la pente de la fonction y en N points de l'intervalle $[x_i, x_i + h]$, puis l'on calcule une pente moyenne (c'est le terme $\sum w_j f_j$) qui sert alors à calculer l'incrément par la méthode d'Euler. Les coefficients α_j, β_{jk} et les poids w_i sont déterminés en imposant que le schéma (7.21) soit équivalent à celui qui est associé à la méthode des séries entières d'ordre N. On bénéficie ainsi, à précision égale, d'une meilleure souplesse qu'avec la méthode de Taylor, en particulier parce qu'il s'agit uniquement d'évaluer la fonction et pas ses dérivées. En fait, le système d'équations qui relient les coefficients α_j, β_{jk} et w_i entre-eux est sous-déterminé (il y a une équation de moins par rapport au nombre d'inconnues). En conséquence, on doit introduire à la main l'un des coefficients, ce qui revient à pouvoir construire, pour un ordre donné, autant de méthode de Runge-Kutta qu'on le souhaite.

Les méthodes de Runge-Kutta les plus utilisées sont celle d'ordre 2 et celle d'ordre 4 (dont l'interprétation graphique est proposée à la figure 7.5). La méthode de Runge-Kutta d'ordre 2 "classique" est équivalente à la méthode de Heun, par construction. Existe aussi la méthode dite de *Newton-Cauchy modifiée* ou *méthode du point milieu* qui a pour schéma

$$y_{i+1} = y_i + hf\left(y_i + \frac{h}{2} f(y_i, x_i), x_i + \frac{h}{2}\right) \tag{7.23}$$

Le schéma de Runge-Kutta d'ordre 4 est

$$y_{i+1} = y_i + \frac{h}{6}\left[f(y_i, x_i) + 2f_1 + 2f_2 + f_3\right] \tag{7.24}$$

où

$$f_1 = f(y_i + \frac{h}{2} f(y_i, x_i), x_i + \frac{h}{2}) \tag{7.25}$$

$$f_2 = f(y_i + \frac{h}{2} f_1, x_i + \frac{h}{2}) \tag{7.26}$$

$$f_3 = f(y_i + hf_2, x_i + h) \tag{7.27}$$

Un exemple est donné à la figure 7.6.

7.4.5 Méthode de Burlish-Stoer

C'est probablement la méthode la plus fiable. Pour un intervalle d'intégration donné $[x_i, x_{i+1}]$, l'incrément δ_i est calculé pour différents pas d'intégration

$$h_1 = h, h_2 = \frac{h}{2}, h_3 = \frac{h}{4}, h_4 = \frac{h}{6}, \ldots, h_n = \frac{h}{n}$$

par exemple en utilisant la méthode de Runge-Kutta, puis on cherche la limite de l'incrément quand le pas tends vers zéro, c'est-à-dire pour $n \to \infty$. Cette procédure est résumée à la figure 7.7. Le nombre effectif n d'intégration n'est pas connu à l'avance, car il dépend du taux de convergence de la série $\{\delta_i\}_n$.

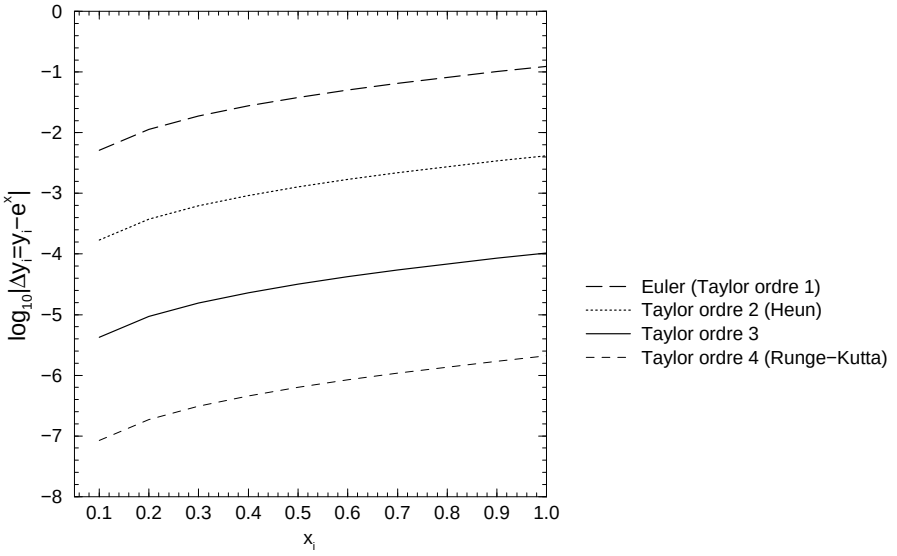


Figure 7.6: Comparaison de la méthode de Runge-Kutta d'ordre 4 avec des méthodes de Taylor d'ordre inférieur. Même conditions de calcul que pour la figure 7.4.

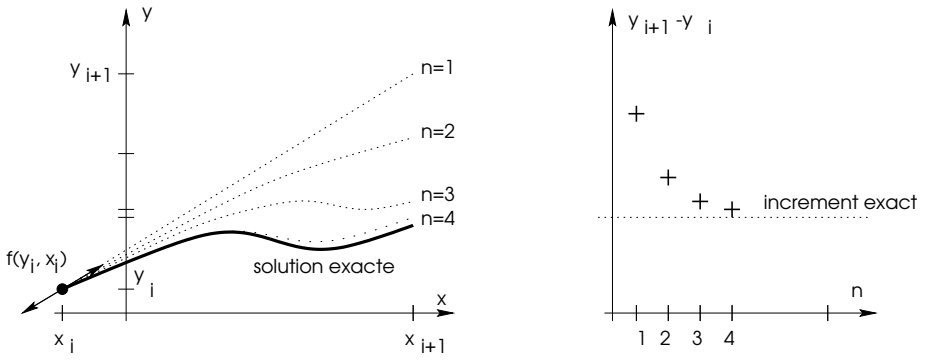


Figure 7.7: Princide de l'intégrateur de Burlish-Stoer. Par approximations successives, on cherche la valeur de la fonction y_{i+1} dans la limite d'un pas d'intégration nul.

7.5 Méthodes "multi-pas"

Les méthodes décrites précédemment permettent d'estimer la fonction en x_{i+1} connaissant uniquement la fonction et sa dérivée au point précédent x_i. Elles constituent ce que l'on appelle des *méthodes mono-pas*. Par opposition, les *méthodes multi-pas* permettent d'accéder à y_{i+1} à partir de toute l'information disponible sur y. On utilise un certain nombre de points calculés précédemment, généralement 3 ou 4, soientt (x_{i-1}, y_{i-1}), (x_{i-2}, y_{i-2}), (x_{i-3}, y_{i-3}), et éventuellement (x_{i-4}, y_{i-4}). L'idée est de représenter la pente de la fonction y (soit la fonction f elle-même) par un polynôme P_N de degré N (par exemple par l'approximation de Lagrange). Pour $N+1$ points successifs (x_k, f_k), on impose donc

$$P_N(x_{i-(N+1-k)}) = f(y_i, x_{i-(N+1-k)}), \qquad k = 1, N+1 \tag{7.28}$$

Ainsi, la relation (7.6) devient intégrable analytiquement

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} P_N(x')dx' \tag{7.29}$$

En pratique, on travaille à bas ordres, pour éviter le phénomène de Runge. Pour $N = 1$, on retrouve la méthode d'Euler. Pour $N = 2$, 3 et 4 on trouve les *formules d'Adams-Bashforth*

- pour $N = 2$

$$y_{i+1} = y_i + \frac{h}{2}\left(3f(y_i, x_i) + f(y_{i-1}, x_{i-1})\right) \tag{7.30}$$

- pour $N = 3$

$$y_{i+1} = y_i + \frac{h}{12}\left(23f(y_i, x_i) - 16f(y_{i-1}, x_{i-1}) + 5f(y_{i-2}, x_{i-2})\right) \tag{7.31}$$

- pour $N = 4$

$$\begin{aligned} y_{i+1} = y_i + \frac{h}{24}\big(&55f(y_i, x_i) - 59f(y_{i-1}, x_{i-1}) \\ &+37f(y_{i-2}, x_{i-2}) - 9f(y_{i-3}, x_{i-3})\big) \end{aligned} \tag{7.32}$$

Un variante consiste à utiliser ces formules comme prédicteurs où $p_{i+1} \approx y_{i+1}$. Par exemple, pour $N = 4$, le prédicteur est donné par la relation (7.32). Pour produire un schéma correcteur associé, on introduit le point (x_{i+1}, p_{i+1}) et on construit un second polynôme qui met en jeu les quatre points (x_{i-2}, f_{i-2}), (x_{i-1}, f_{i-1}), (x_i, f_i) et (x_{i+1}, f_{i+1}) où $f_{i+1} \equiv f(p_{i+1}, x_{i+1})$. On obtient sur le même principe

$$y_{i+1} = y_i + \frac{h}{24}\left(f_{i-2} - 5f_{i-1} + 19f_i + 5f_{i+1}\right) \tag{7.33}$$

Les formules (7.32) et (7.33) constituent le schéma prédicteur-correcteur de Adams-Bashforth-Moulton d'ordre 4. Il en existe d'autres ...

Les méthodes multi-pas sont très précises, généralement plus précises que les méthodes mono-pas (exception faite peut-être de l'intégrateur de Burlish-Stoer). En revanche, elles peuvent être instables et conduire à des oscillations, par exemple si l'intervalle d'intégration est grand. On peut les stabiliser en choisissant un pas plus petit. Notez que l'on ne peut jamais utiliser une méthode multi-pas dès le début de l'intégration. On peut alors démarrer avec un schéma de Runge-Kutta puis passer, quand le nombre de points est suffisant, à un schéma multi-pas.

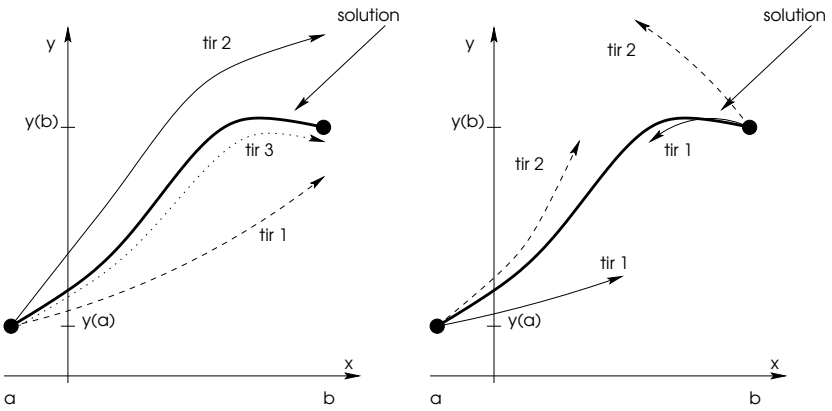


Figure 7.8: Dans la méthode de tir simple, on intègre l'équation à partir d'un contour en cherchant à atteindre la condition à l'autre contour. On peut également intégrer des deux cotés simultanément en cherchant à "joindre les deux bouts" en un point intermédiaire.

7.6 Méthodes de tir

Jusqu'à présent, nous avons essentiellement parlé des méthodes relatives aux problèmes aux conditions initiales. Les choses se compliquent un peu pour les problèmes aux conditions aux contours. Cela se conçoit facilement si l'on souvient que la fonction est contraintes par sa valeur ou par la valeur de sa dérivée en $x = a$ et en $x = b$. L'exemple du tir ballistique permet de fixer les idées: on sait les conditions très particulières qu'il faut réunir simultanément pour qu'un projectile partant d'un point $\mathrm{A}(a, y(a))$ parvienne en un point $\mathrm{B}(b, y(b))$; il faudra "régler" la vitesse ou/et l'angle de tir. Il n'est d'ailleurs pas garanti que l'on puisse toujours atteindre l'objectif, par exemple si la vitesse de tir est limitée ou/et l'angle imposé. Si c'est le cas, on opèrera en pratique par approximations successives: pour une vitesse initiale donnée (ou un angle donné), on règlera progressivement l'angle de tir (respectivement la vitesse) afin d'atteindre la cible. Deux aspects fondamentaux des problèmes de conditions aux contours sont illustrés ici. Il y a d'une part la non-existence possible de solution, une difficulté qui n'est pas présente pour les problèmes aux conditions initiales. D'autre part, on voit que le problème aux conditions aux contours peut-être assez naturellemnt converti en un problème aux conditions initiales devant satisfaire une condition finale. C'est justement ce que l'on appelle une *méthode de tir simple*, illustrée à la figure 7.8.

Pour trouver la solution d'un problème avec conditions aux contours à l'aide d'une méthode de tir simple, on procédera par dichotomie en cherchant d'abord deux jeux de conditions initiales qui permettront d'encadrer la condition "finale" souhaitée. Par exemple, si y_b désigne la condition à la limite en $x = b$ que doit satisfaire y, on cherchera $y^+(a)$ et $y^-(a)$ (ou des conditions sur la dérivée en $x = a$), c'est-à-dire finallement deux fonctions y^- et y^+ telles que, une fois intégrées jusqu'en $x = b$, elles satisfassent

$$(y^+(b) - y_b)(y^-(b) - y_b) \leq 0 \tag{7.34}$$

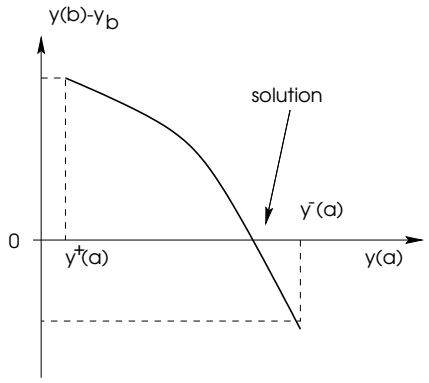


Figure 7.9: Dans la méthode de tir simple, on est amené à résoudre une fonction du type $f(x) = 0$, où x désigne une condition sur un contour et f l'erreur sur la valeur de la condition à l'autre contour. Ici $x \equiv y(a)$ et $f \equiv y(b) - y_b$.

Des tirs de plus en plus ressérés autour de y^+ et y^- permettront alors d'approcher la solution, sous réserve que l'équation différentielle soit suffisamment régulière. En pratique, il s'agira finalement d'annuler (du moins, de minimiser) l'écart entre la condition souhaitée sur l'autre contour et cette calculée, comme indiqué à la figure 7.9. On pourra alors utiliser toutes sortes de stratégies dédiées à la recherche de zéro ou à la minimisation de fonctions ou de système de fonctions.

On peut également procéder à un *double tir*, l'un correspondant à une intégration depuis $x = a$ vers b et l'autre à une intégration dans le sens inverse, en cherchant un point de raccordement quelque part sur l'intervalle.

7.7 Schémas aux différences finies

Une autre façon d'aborder les problèmes aux conditions aux contours, plus rapide aussi, mais généralement moins précise, consiste à écrire le schéma aux différence finies associé à l'équation différentielle en N *points de grille* x_i (régulièrement) répartis sur $[a, b]$. Par exemple, pour une équation d'ordre 2 telle que l'équation de Poisson à une dimension,

$$\phi''(x) - s(x) = 0 \tag{7.35}$$

où ϕ est le potentiel et s une *fonction source* connue, on commence par écrire le schéma aux différences associé. A l'ordre 2 inclus (sur les bords comme au centre), on a

$$\begin{cases} \dfrac{2\phi_1 - 5\phi_2 + 4\phi_3 - \phi_4}{h^2} - s_1 = 0, & i = 1 \\ \dfrac{\phi_{i-1} - 2\phi_i + \phi_{i+1}}{h^2} - s_i = 0, & i \in [2, N-1] \\ \dfrac{2\phi_N - 5\phi_{N-1} + 4\phi_{N-2} - \phi_{N-3}}{h^2} - s_N = 0, & i = N \end{cases}$$

Si le potentiel ϕ est spécifié aux deux bords x_1 et x_N, le système précédent se ré-écrit

$$\begin{cases} -5\phi_2 + 4\phi_3 - \phi_4 = s_1 h^2 - 2\phi_1, \\ -2\phi_2 + \phi_3 = s_2 h^2 - \phi_1, \qquad i = 2 \\ \dots \\ \phi_{i-1} - 2\phi_i + \phi_{i+1} = s_i h^2, \qquad i \in [3, N-3] \\ \dots \\ \phi_{N-2} - 2\phi_{N-1} = s_{N-1} h^2 - \phi_N, \qquad i = N-1 \\ -5\phi_{N-1} + 4\phi_{N-2} - \phi_{N-3} = s_N h^2 - 2\phi_N, \end{cases}$$

On aboutit alors à un système de N équations linéaires dont les inconnues sont les $N-2$ valeurs ϕ_i (avec $i \in [2, N-1]$). On élimine alors 2 équations, celles correspondant à $i = 2$ et $i = N-1$ que l'on ré-injecte sur les bords, pour obtenir

$$\begin{cases} 3\phi_3 - 2\phi_4 = (2s_1 - 5s_2)h^2 + \phi_1, \\ \phi_{i-1} - 2\phi_i + \phi_{i+1} = s_i h^2, \qquad i \in [3, N-3] \\ -2\phi_{N-1} + 3\phi_N = (2s_N - 5s_{N-1})h^2 + \phi_N, \end{cases}$$

On écrira ce système sous forme matricielle

$$A\bar{\phi} = B \tag{7.36}$$

et on déterminera les solutions du système par des méthodes appropriées. Pour les schémas de bas ordre, la résolution est souvent facilitée par le fait que la matrice est nulle sauf sur, et autour de la diagonale. Dans ce type de méthode, l'erreur est donnée par la précision du schéma aux différences.

7.8 Exercices et problèmes

■ Estimez la trajectoire d'une planète en orbite circulaire autour d'une étoile, par la méthode d'Euler, en utilisant différents pas d'intégration. Comparez à la solution exacte.

■ Ecrivez un schéma de type prédicteur/correcteur en évaluant l'intégrale dans la formule (7.10) par la méthode de Simpson.

■ Programmez l'exemple ci-dessus. Comparez la méthode pour différents pas.

■ Ecrivez le schéma d'ordre 4 correspondant à un développement de Taylor pour l'équation différentielle suivante: $y'(x) = 1 + y^2(x)$. Programmez la méthode et comparez les résultats à la solution exacte.

■ Etablissez les relations (7.24) et (7.23).

■ Etablissez un schéma de Runge-Kutta d'ordre 5.

■ Construisez votre propre intégrateur de type Burlish-Stoer.

■ Montrez que méthode d'Adams-Bashforth obtenue pour $N = 1$ coincide effectivement avec la méthode d'Euler.

■ En vous inspirant de la démarche suivie D, établissez les formules d'Adams-Bashforth-Moulton pour $N = 3$. Même question concernant les relations (7.32) et (7.33).

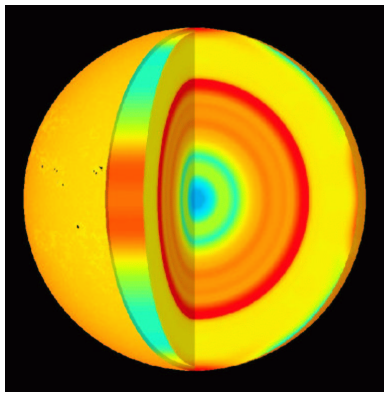
■ Ecrire le schéma aux différences finies pour l'équation de Poisson à une dimension $\Delta\Psi(x) = 4\pi G\rho(x)$ sur l'intervalle $x \in [-1, 1]$ avec $\rho(x) = 1 - x^2$ et résoudre le problème pour $\frac{d\Psi}{dx}(0) = 0$ et $\Psi(1) = 0$. Comparez avec la solution exacte en fonction du nombre de points de grille.

Chapitre 8

Applications astrophysiques

8.1 Structure interne des étoiles

Il s'agit de résoudre les équations "les plus simples" régissant la structure interne des étoiles à symétrie sphérique et chimiquement homogènes. En régime stationnaire, quatre équations forment un système différentiel ordinaire du premier ordre hautement non-linéaire que l'on propose de résoudre par la méthode du tir. Cette méthode consiste à intégrer simultanément les équations du système à partir du centre et de la surface de l'étoile puis à chercher, par itération successives sur les quatre conditions aux limites (deux sont données au centre, et deux en surface), à réaliser la continuité des quatre fonctions intégrées en un point intermédiaire.



Différents niveaux de réalisme dans la description physique peuvent être envisagés (équation d'état, opacité, production d'énergie, convection, etc.)

Références

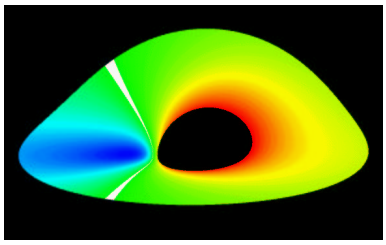
- Péquignot D., Cours de DEA, "Structure et évolution des étoiles"
- Cox J.P., Giuli R.T., 1968, in "Principles of stellar structure. Vol. I", Gordon & Breach

8.2 Image et spectre d'un disque autour d'un trou noir de Schwarzschild

Il s'agit de calculer l'apparence d'un disque d'accrétion autour d'un trou noir de Schwarz-schild tel que le verrait un observateur situé à l'infini. Techniquement, il "suffit" de réaliser un programme de tracé de rayons dans un espace courbe décrit par la métrique de Schwarzschild. On déterminera ensuite le spectre émis par le disque dans deux cas limites: i) le disque émet en chaque point un spectre de corps noir, et ii) le disque émet en chaque point une raie infiniment mince.

Références

- Luminet J.P., 1979, *Astron. & Astrophys.*, **75**, 228
- Fu & Taam, R.E., 1990, *Astrophys. J.*, **349**, 553
- Hameury, J.-M., Marck, J.-A., Pelat, D., 1994, *Astron. & Astrophys.*, **287**, 795



Cas d'un trou noir de Kerr

8.3 Instabilité thermique d'un disque d'accrétion de Sakura-Sunyaev

Un disque d'accrétion est une structure applatie dans laquelle la matière est transportée vers le centre et le moment cinétique dans la direction opposée. Dans le modèle "standard" (Shakura & Sunyaev, 1973), on suppose que le disque est képlérien, c'est-à-dire que les trajectoires de la matière autour de l'objet central sont des orbites képlériennes. Dans beaucoup de cas, cela est une très bonne approximation. Pour certains taux d'accrétion, le disque devient localement thermiquement instable, c'est-à-dire que lorsque sa température augmente, il se réchauffe plus vite qu'il ne se refroidit. Un tel comportement devrait conduire à la destruction du disque. Toutefois, l'analyse globale montre que le disque peut survivre à l'instabilité thermique en adoptant un comportement non-stationnaire.

Il faudra dans un premier temps construire un modèle numérique de disque stable et stationnaire (Taam & Lin, 1984). On augmentera ensuite le taux d'accrétion jusqu'au seuil d'instabilité et on adaptera le modèle numérique de façon à pouvoir suivre le comportement dynamique du disque. D'un point de vue numérique, il faudra s'initier à la résolution de systèmes d'équations aux dérivées partielles par des méthodes aux différences finies suivant les schémas implicites et/ou explicites.

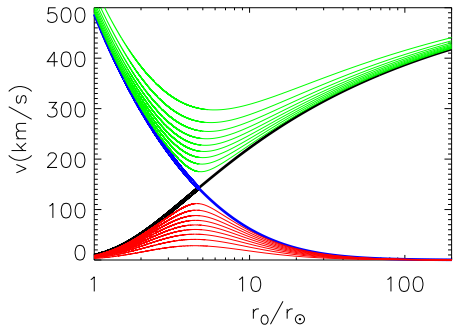
Références

- Taam, R.E. & Lin D.N.C., 1984, *Astrophys. J.*, **287**, 761
- Shakura N.I., Sunyaev R.A., 1973, *A&A*, 24, 337
- Pelat D., polycopié

8.4 Vent solaire

C'est à Parker que l'on attribue les idées fondatrices de la théorie du vent solaire dont l'existence fut définitivement établie dans les années 1950 par l'observation des comètes puis confirmée par les mesures de la sonde spatiale Mariner-2 la décennie suivante. Le vent emporte avec lui le champ magnétique de Soleil. Il est accéléré à la base de la couronne et s'étend dans tout le système solaire fouettant notamment les planètes sur son passage. Initialement, la vitesse du vent est faible, mais celle-ci croît au fur et à mesure qu'il s'écoule dans l'espace interplanétaire pour atteindre plusieurs centaines de kilomètres par seconde. A cause de ce phénomène, le Soleil perd inexorablement de sa masse, à un taux toutefois relativement faible, de l'ordre de 10^{-14} $M_\odot$/an (aujourd'hui).

On étudiera l'écoulement du vent solaire d'après le modèle hydrodynamique simplifié de Parker (1960): le fluide est supposé à symétrie sphérique, stationnaire, non-magnétisé, et isotherme (on pourra toutefois envisager une solution polytropique). On traitera d'abord le cas stationnaire et on étudiera en particulier la topologie des solutions en fonction des paramètres de contrôle que sont la vitesse du vent à la base de la couronne et la température du gaz. Puis, on abordera le cas dépendant du temps et en particulier l'installation de l'écoulement ainsi que la sensibilité de la dynamique du vent aux conditions aux limites. On pourra ainsi voir comment un vent peut se transformer en une accrétion (Bondi, 1952).



Vents de Parker
(d'après I. Zouganelis, DEA prom. 2001/2002)

Du point de vue méthodologique, le problème stationnaire met en jeu la résolution d'une équation différentielle ordinaire mono-dimensionnelle (ODE) avec franchissment d'une singularité (correspondant en fait au passage du point sonique). Le problème dépendant du temps est un problème aux dérivées partielles (PDEs) avec conditions aux contours (TBVP).

Références

- Philips K.J.H., 1995, in 'Guide to the sun", CUP
- Parker E.N., 1960, *ApJ*, **132**, 821
- Bondi H., 1952, *MNRAS*, **112**, 195

8.5 Equation de Saha

La connaissance de l'équation d'état des gaz est fondamentale pour comprendre la nature des fluides astrophysiques comme ceux qui composent le milieu interstellaire, les étoiles, les étoiles à neutrons, les planètes géantes, les disques d'accrétion, etc. Dans nombre de situations, les gaz sont non-dégénérés et quasiment à l'équilibre thermodynamique local (ETL). C'est notamment le cas du

gaz de l'interieur du Soleil. Selon la température du milieu et sa pression (ou sa densité), le gaz peut revêtir différents aspects physico-chimiques: totalement ou partiellement ionisé, neutre, moléculaire ou condensé sous forme de grains. Les propriétés globales qui en découlent, notamment les propriétés radiatives et thermodynamiques, jouent alors un rôle clé dans l'équilibre ou l'évolution du milieu lui-même. Par exemple, un milieu froid constitué majoritairement de molécules H_2 n'aura pas le même comportement (dynamique, émissif, etc.) qu'un milieu complètement ionisé. Dans le cadre de l'ETL, s'il peut être justifié par ailleurs, la détermination des composants chimiques d'un gaz de type parfait pour une température T et une pression P données est grandement facilitée par le fait que l'on peut s'affranchir de la connaissance des mécanismes microscopiques qui forment et détruisent les espèces individuellement, ceci par le biais de la mécanique statistique (Tsuji, 1964a et 1964b; Cox & Giuli, 1965).

Dans ce projet, on mettra en place un outil permettant de résoudre l'équilibre de Saha (équilibre ions/neutres) et l'équilibre de dissociation (équilibre neutres/molécules) pour un couple (T, P) donné, à partir d'un mélange initial d'éléments (H:He:C:N:O:...). On pourra éventuellement traiter la condensation (équilibre gas/grain) et en déduire des quantités thermodynamiques fondamentales comme le poids moléculaire moyen, l'énergie interne, l'entropie, les gradients adiabatiques, etc. Il sera intéressant de pouvoir modifier à loisir le mélange d'éléments (H:He:C:N:O:...).

Du point de vue méthodologique, le problème consiste essentiellement à résoudre une équation vectorielle du type $\vec{F}(\vec{X}) = \vec{0}$. Par ailleurs, le système est "raide".

Références

- Tsuji T., 1964a, Pub. Soc. Japan, **18**, 127
- Tsuji T., 1964b, Ann. Tokyo Astron. Obs. 2nd Ser., **9**, 1
- Cox J.P., Giuli R.T., 1968, in "Principles of stellar structure. Vol. I", Gordon & Breach

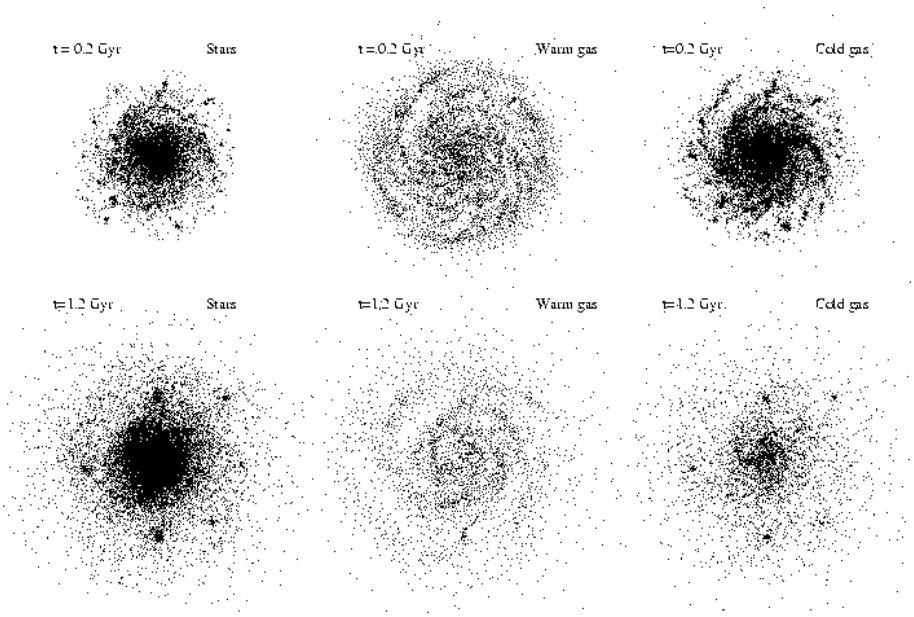
8.6 Dynamique des disques stellaires

On peut simuler les instabilités spirales ou de barres d'un disque d'étoiles par une modélisation numérique de type N-corps et comprendre ainsi certains aspects de la dynamique galactique. Dans ce projet, il s'agira d'étudier l'évolution temporelle d'un ensemble de N particules matérielles (les étoiles) évoluant dans un champ comprenant éventuellement un champ central (trou noir) et le champ propre du disque stellaire. On envisagera tout d'abord une modélisation de type "PP" ou les interactions sont calculées "particule à particle". Il faudra construire un intégrateur de mouvement particulièrement fiable et traiter soigneusement le problème des rencontres à deux corps. On pourra ensuite implémenter une méthode de type PM, voire une structure hiérarchique de type "tree-code", permettant d'augmenter le nombre de particules (et ainsi d'accroître le degré de réalisme de l'expérience) sans augmenter le temps de calcul de manière prohibitive. De nombreuses extensions sont possibles.

Du point de vue méthodologique, le problème consiste essentiellement à résoudre des équations différentielles ordinaires (ODEs).

Références

- Hockney R.W., Eastwood J.W., 1981, in "Computer simulations using particles, New-York, Mc Graw Hill
- Hernquist L., 1987, *ApJ*, **64**, 715
- Selwood J.A., Carlberg R.G., 1984, *ApJ*, **282**, 61

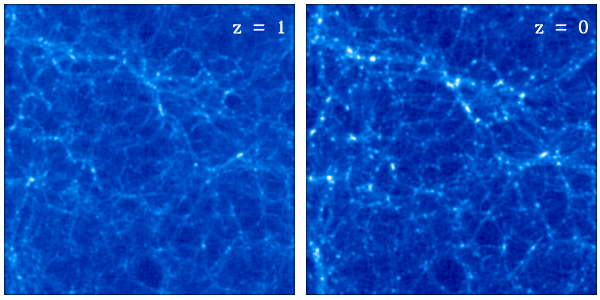


Instabilités dans les disques galactiques
(d'après Semelin & Combes, 2002)

8.7 Fluide cosmologique et formation des grandes structures

La formation des galaxies dans l'Univers primordial a été déterminée par la structuration de la matière à grande échelle. On peut en étudier le mécanisme grâce à un suivi de l'évolution spatio-temporelle de perturbations initiales de densité. On mettra en place un outil permettant la simulation d'un fluide de matière noire (non-collisionnelle) à une puis à deux dimensions. L'originalité de l'étude vient du fait que les équations du mouvement doivent inclure l'expansion de l'Univers, régulée par les paramètres cosmologiques "habituels". On pourra ainsi, selon les conditions initiales, observer l'effondrement ou la diffusion de structures à petite échelle.

Du point de vue méthodologique, le problème consiste essentiellement à résoudre des équations aux dérivées partielles (PDEs).



Distribution spatiale de la matire noire à différents redshit
dans un scnario ΛCDM (Courty, 2002).

Références

- Hockney R.W., Eastwood J.W., 1981, in "Computer simulations using particles, New-York, Mc Graw Hill
- Efstafhiou G. et al., 1985, *ApJ*, **57**, 241
- Courty S., polycopié

Bibliographie

[1] Numerical methods for Mathematics Science and Engineering, 1992, J.H. Mathews, Press-Hall International Inc., 2nd edition

[2] Numerical methods that work, 1970, J.H.Mathews, Press-Hall International Inc., 2nd edition

[3] Numerical receipes, the art of scientific computing, 1992, W.H. Press et al., CUP, 2nd edition

[4] Analyse numérique pour ingénieurs, 1995, A. Fortin, Ed. de l'Ecole Polytechnique de Montréal

[5] Méthodes numériques. Interpolation — Dérivées, 1959, J. Kuntzmann, Dunod

[6] Introduction to numerical analysis, 1956, F.B. Hildebrand, McGraw-Hill

[7] Numerical analysis, 1961, Z. Kopal, Chapman & Hall LTD

[8] Computing methods. Vol. I, 1965, O.M. Blunn, Booth A.D., Pergamon Press LTD

[9] Computing methods. Vol. II, 1965, O.M. Blunn, Booth A.D., Pergamon Press LTD

Revues périodiques

[10] Computer Physics Communications

[11] Computer Physics Communications

[12] Computing in Science and Engineering

[13] Journal of Computational Physics

[14] Journal of Mathematical Physics

[15] Mathematics of Computation

Appendice A

GAMS: Project Summary

Extrait de `http://gams.nist.gov/`.

[...]. The Guide to Available Mathematical Software project of the National Institute of Standards and Technology (NIST) studies techniques to provide scientists and engineers with improved access to reusable computer software which is available to them for use in mathematical modeling and statistical analysis. One of the products of this work is an on-line cross-index of available mathematical software. This system also operates as a virtual software repository. That is, it provides centralized access to such items as abstracts, documentation, and source code of software modules that it catalogs; however, rather than operate a physical repository of its own, this system provides transparent access to multiple repositories operated by others.

Currently four software repositories are indexed: three maintained for use by NIST staff (but accessible to public), and netlib, a publically accessible software collection maintained by Oak Ridge National Laboratory and the University of Tennessee at Knoxville (netlib in Tennessee) and Bell Labs (netlib at Bell Labs). This represents some 10,000 problem-solving modules from more than 100 software packages. The vast majority of this software represents Fortran subprograms for mathematical problems which commonly occur in computational science and engineering, such as solution of systems of linear algebraic equations, computing matrix eigenvalues, solving nonlinear systems of differential equations, finding minima of nonlinear functions of several variables, evaluating the special functions of applied mathematics, and performing nonlinear regression. Among the packages cataloged are : the IMSL, NAG, PORT, and SLATEC libraries; the BLAS, EISPACK, FISHPACK, FNLIB, FFTPACK, LAPACK, LINPACK, and STARPAC packages; the DATAPLOT and SAS statistical analysis systems; as well as other collections such as the Collected Algorithms of the ACM. Note that although both public-domain and proprietary software is cataloged here, source code of proprietary software products are not available, although related items such as documentation and example programs often are.

All cataloged problem-solving software modules are assigned one or more problem classifications from a tree-structured taxonomy of mathematical and statistical problems. Users can browse through modules in any given problem class. To find an appropriate class, one can utilize the taxonomy as a decision tree, or enter keywords which are then mapped to problem classes. Search

filters can be declared which allow users to specify preferences such computing precision or programming language. In addition, users can browse through all modules in a given package, all modules with a given name, or all modules with user-supplied keywords in their abstracts. [...]

Appendice B

Quadrature de Gauss-Legendre

L'idée générale de la méthode de Gauss est d'imposer des contraintes sur les x_i où la fonction f est évaluée, de façon à ce que la méthode soit exacte pour des polynômes de degrés élévés. Ceci permet d'augmenter le degré de précision des méthodes associées. Dans la méthode de Gauss-Legendre, on calcule l'intégrale définie $\int_{-1}^{1} f(x)dx$ à l'aide de N évaluations de f aux points $x_1, x_2, \ldots, x_N$, soit

$$\int_{-1}^{1} f(x)dx \approx \sum_{i=1}^{N} w_i f_i$$

où $f_i \equiv f(x_i)$. Il y a $2N$ inconnues dans cette formule: les N abscisses x_i (ou noeuds) et les N coefficients w_i (les poids). Il est donc, en principe, possible de s'imposer $2N$ contraintes. Comme un polynôme de degré $2N-1$ est défini par $2N$ coefficients, il semble alors possible d'exiger que la formule d'intégration ci-dessus soit exacte pour les polynômes de degré $2N-1$ (et donc pour tous les polynômes de degré $k \leq 2N-1$).

Soit $Q_0, Q_1, \ldots, Q_{2N-1}$ une base des polynômes de degré $2N-1$. Une condition nécessaire et suffisante pour que la formule d'intégration soit exacte jusqu'aux polynômes de degré $2N-1$ est

$$\int_{-1}^{1} Q_k(x)dx = \sum_{i=1}^{N} w_i Q_k(x_i), \quad \text{pour } k = 0, 1, \ldots, 2N-1.$$

La formule précédente constitue un système de $2N$ équations à $2N$ inconnues. La difficulté majeure est que ce système n'est pas linéaire, donc difficile à inverser. Choisissons par exemple la base $Q_k(x) = x^k$. Il faut alors résoudre

$$\int_{-1}^{1} x^k dx = \sum_{i=1}^{N} w_i Q_k(x_i), \quad \text{pour } k = 0, 1, \ldots, 2N-1.$$

L'intégrale est facilement évaluée. On a en effet:

$$\int_{-1}^{1} x^k \, dt = \frac{1-(-1)^{k+1}}{k+1} = \begin{cases} \dfrac{2}{k+1} & \text{si } k \text{ est pair}\,; \\ \\ 0 & \text{si } k \text{ est impair}\,. \end{cases}$$

Ce qui conduit au système suivant

$$\left.\begin{aligned} \sum_{i=1}^{N} w_i &= 2\,, \\ \sum_{i=1}^{N} w_i x_i &= 0\,, \\ &\dots \\ \sum_{i=1}^{N} w_i x_i^{2N-2} &= \frac{2}{2N-1}\,, \\ \sum_{i=1}^{N} w_i x_i^{2N-1} &= 0\,. \end{aligned}\right\} \tag{B.1}$$

dont la résolution n'est pas triviale. Il est toutefois possible d'aboutir à un système plus simple par un choix judicieux des polynômes de base Q_k. Prenons en effet $Q_k(x) = x^k$ pour $k = 0, 1, \dots, N-1$ et $Q_k(x) = x^{k-N} P_N(x)$ pour $k = N, \dots, 2N-1$, où P_N est le polynôme de Legendre de degré N. Il est clair que les Q_k sont des polynômes de degré k et que, par conséquent, ils forment une base des polynômes de degré $2N-1$. Les polynômes de Legendre sont orthogonaux sur le domaine d'intégration qui nous intéresse, c'est-àdire

$$\int_{-1}^{1} P_n(x) P_m(x) dx = \begin{cases} 0\,, & \text{si } n \neq m; \\ \frac{2}{2n+1}\,, & \text{si } n = m \end{cases}$$

Cette propriété fait que P_N est orthogonal à tous les polynômes de degré $k < N$. En particulier, $P_N(x)$ est orthogonal à x^{k-N} pour $k = N, \dots, 2N-1$. Dans ces conditions, le système à résoudre s'écrit maintenant

$$\left.\begin{aligned} \sum_{i=1}^{N} w_i &= 2\,, \\ \sum_{i=1}^{N} w_i x_i &= 0\,, \\ &\dots \\ \sum_{i=1}^{N} w_i x_i^{N-1} &= \frac{1-(-1)^N}{N}\,, \\ \sum_{i=1}^{N} w_i P_N(x_i) &= 0\,, \\ &\dots \\ \sum_{i=1}^{N} w_i x_i^{N-1} P_N(x_i) &= 0\,. \end{aligned}\right\} \tag{B.2}$$

Ce système de $2N$ équations à $2N$ inconnues se scinde en deux sous-systèmes

1. Si l'on suppose les x_i connus, les N premières équations forment un système linéaire qui, sous réserve de régularité, permet de déterminer les w_i.

2. Les N dernières équations sont toutes de la forme $\sum_{i=1}^{N} w_i x_i^k P_N(x_i) = 0$ pour $k = 0, \ldots, N-1$. Si les x_i sont les N racines de $P_N(t)$, on aura pour tout i, $P_N(x_i) = 0$ et le sous-système sera toujours satisfait. Il se trouve que ces racines sont réelles, distinctes et comprises entre -1 et 1.

Nous allons démontrer d'une part que le premier sous-système est régulier et d'autre part que la condition suffisante $P_n(x_i) = 0$ est aussi nécessaire, en d'autres termes que la solution du système (B.2) est unique.

Premier sous-système. Supposons les x_i connus (par exemple les solutions de $P_N(x) = 0$), la matrice de ce système linéaire s'écrit alors:

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_N \\ x_1^2 & x_2^2 & \ldots & x_N^2 \\ \ldots & \ldots & \ldots & \ldots \\ x_1^{N-1} & x_2^{N-1} & \ldots & x_N^{N-1} \end{bmatrix} .$$

Il s'agit d'une matrice de Vandermonde dont le déterminant D vaut :

$$D = \prod_{i<j} (x_i - x_j) .$$

Ce déterminant n'est pas nul car les x_i sont tous distincts, le système linéaire est donc régulier. Le premier sous-système admet alors une solution unique qui peut être calculée en inversant la matrice de Vandermonde. Il est cependant possible de trouver une expression plus utile grâce aux polynômes de Lagrange.

Soient L_i les polynômes de Lagrange correspondants aux points x_i. Le polynôme L_i est de degré $N-1$ son intégration par la formule de Gauss-Legendre est donc exacte. Il vient

$$\int_{-1}^{1} L_i(x) dx = \sum_{j=1}^{N} w_i L_i(x_j) .$$

Mais on a $L_i(x_j) = \delta_{ij}$ d'où

$$w_i = \int_{-1}^{1} L_i(x)\, dx .$$

Mais si $L_i(x_j) = \delta_{ij}$, il en est de même de $L_i^2(x_j)$ et comme le polynôme L_i^2 est de degré $2N-2$, il est lui-aussi intégrable par Gauss-Legendre. Le même raisonnement s'applique et on trouve alors

$$w_i = \int_{-1}^{1} L_i^2(x)\, dx .$$

Cette dernière formule montre que les w_i sont des réels strictement positifs.

Second sous-système. Nous avons vu que ce sous-système est satisfait si les x_i sont les N solutions de $P_N(x) = 0$, reste à montrer que cette solution est unique. Raisonnons par l'absurde et supposons qu'il existe au moins un x_i tel

N	i	noeuds x_i	poids w_i
1	1	0	2
2	1; 2	$\mp$0.57735 02691 89626	1
3	1; 3	$\mp$0.77459 66692 41483	0.55555 55555 55556 = 5/9
	2	0	0.88888 88888 88889 = 8/9
4	1; 4	$\mp$0.86113 63115 94053	0.34785 48451 37454
	2; 3	$\mp$0.33998 10435 84856	0.65214 51548 62546
5	1; 5	$\mp$0.90617 98459 38664	0.23692 68850 56189
	2; 4	$\mp$0.53846 93101 05683	0.47862 86704 99366
	3	0	0.56888 88888 88889
6	1; 6	$\mp$0.93246 95142 03152	0.17132 44923 79170
	2; 5	$\mp$0.66120 93864 66265	0.36076 15730 48139
	3; 4	$\mp$0.23861 91860 83197	0.46791 39345 72691
7	1; 7	$\mp$0.94910 79123 42759	0.12948 49661 68870
	2; 6	$\mp$0.74153 11855 99394	0.27970 53914 89277
	3; 5	$\mp$0.40584 51513 77397	0.38183 00505 05119
	4	0	0.41795 91836 73469
8	1; 8	$\mp$0.96028 98564 97536	0.10122 85462 90376
	2; 7	$\mp$0.79666 64774 13627	0.22238 10344 53374
	3; 6	$\mp$0.52553 24099 16329	0.31370 66458 77887
	4; 5	$\mp$0.18343 46424 95650	0.36268 37833 78362

Table B.1: Éléments pour le calcul de la formule de Gauss-Legendre pour $N \leq 12$.

que $P_N(x_i) \neq 0$. Le polynôme $L_i(x)P_N(x)$ est de degré $2N-1$; il est intégrable par Gauss-Legendre, il vient alors

$$\begin{aligned}
\int_{-1}^{1} L_i(x)P_n(x)dx &= 0\,, \\
\sum_{j=1}^{N} w_j L_i(x_j)P_N(x_j) &= 0\,, \\
\sum_{j=1}^{N} w_j \delta_{ij} P_N(x_j) &= 0\,, \\
w_i P_N(x_i) &= 0\,.
\end{aligned}$$

Mais cette dernière condition est contradictoire puisque $w_i > 0$ et que l'on a supposé que $P_N(x_i) \neq 0$. Il n'existe donc pas de x_i tels que $P_N(x_i) \neq 0$. Il resterait à démontrer que les x_i sont bien des réels distincts compris entre -1 et 1. Il s'agit en fait d'une propriété classique des polynômes de Legendre!

N	i	noeuds x_i	poids w_i
9	$1;9$	$\mp$0.96816 02395 07626	0.08127 43883 61574
	$2;8$	$\mp$0.83603 11073 26636	0.18064 81606 94857
	$3;7$	$\mp$0.61337 14327 00590	0.26061 06964 02935
	$4;6$	$\mp$0.32425 34234 03809	0.31234 70770 40003
	5	0	0.33023 93550 01260
10	$1;10$	$\mp$0.97390 65235 17172	0.06667 13443 08688
	$2;9$	$\mp$0.86506 33666 88985	0.14945 13491 50581
	$3;8$	$\mp$0.67940 95682 99024	0.21908 63625 15982
	$4;7$	$\mp$0.43339 53941 29247	0.26926 67193 09996
	$5;6$	$\mp$0.14887 43389 81631	0.29552 42247 14753
11	$1;11$	$\mp$0.97822 86581 46057	0.05566 85671 16172
	$2;10$	$\mp$0.88706 25997 68095	0.12558 03694 64905
	$3;9$	$\mp$0.73015 20055 74049	0.18629 02109 27734
	$4;8$	$\mp$0.51909 61292 06812	0.23319 37645 91990
	$5;7$	$\mp$0.26954 31559 52345	0.26280 45445 10247
	6	0	0.27292 50867 77901
12	$1;12$	$\mp$0.98156 06342 46719	0.04717 53363 86512
	$2;11$	$\mp$0.90411 72563 70475	0.10693 93259 95318
	$3;10$	$\mp$0.76990 26741 94305	0.16007 83285 43346
	$4;9$	$\mp$0.58731 79542 86617	0.20316 74267 23066
	$5;8$	$\mp$0.36783 14989 98180	0.23349 25365 38355
	$6;7$	$\mp$0.12523 34085 11469	0.24914 70458 13403

Table B.2: Éléments pour le calcul de la formule de Gauss-Legendre pour $N \leq 12$ (suite).

Appendice C

Formule de Peano

Peano a donné une expression exacte de l'erreur de quadrature $R(f)$. Cette expression dépend d'une fonction $K(t)$ appelée *noyau de Peano*.

Noyau de Peano. Soit une méthode d'intégration numérique conduisant à l'erreur de quadrature R. Le noyau de Peano associé à cette méthode est une fonction $K(t)$ ainsi définie :

$$K(t) = \frac{1}{n!} R\big(x \to (x-t)_+^n\big)\,, \quad (x-t)_+^n = \begin{cases} (x-t)^n & \text{si } x \geq t\,; \\ 0 & \text{si } x < t\,, \end{cases}$$

où n est le plus grand degré des polynômes pour lesquels l'intégration numérique est exacte.

Conformément à la remarque précédente, la notation $R\big(x \to (x-t)_+^n\big)$ veut dire que dans l'expression de R, la fonction à intégrer est $(x-t)_+^n$ et que la variable d'intégration est x.

Calculons le noyau de Peano associé à la méthode de Simpson. Par définition de R on a :

$$R(f) = \tfrac{1}{3} f(-1) + \tfrac{4}{3} f(0) + \tfrac{1}{3} f(1) - \int_{-1}^{1} f(x)\,dx\,.$$

Il s'agit d'une méthode d'intégration à trois points qui est exacte pour les plynômes de degré inférieur ou égal à trois. Le noyau de Peano s'écrit alors :

$$\begin{aligned} K(t) &= \tfrac{1}{6} R\big(x \to (x-t)_+^3\big)\,, \\ &= \frac{1}{6}\big(\tfrac{1}{3}(-1-t)_+^3 + \tfrac{4}{3}(0-t)_+^3 + \tfrac{1}{3}(1-t)_+^3 - \int_{-1}^{1} (x-t)_+^3\,dx\big)\,. \end{aligned}$$

Le noyau $K(t)$ est défini pour t quelconque mais seul le domaine d'intégration $t \in [-1, 1]$ est utile. Dans cet intervalle, il vient :

$$(-1-t)_+^3 = 0\,, \quad (1-t)_+^3 = (1-t)^3,$$

$$(-t)_+^3 = \begin{cases} 0 & \text{si } t \geq 0\,; \\ -t^3 & \text{si } t < 0\,. \end{cases}$$

$$\int_{-1}^{1} (x-t)_+^3\,dx = \int_{t}^{1} (x-t)^3\,dx = \tfrac{1}{4}(1-t)^4.$$

Finalement le noyau de Peano associé à la méthode de Simpson est, dans l'intervalle d'intégration $[-1,1]$, donné par l'expression :

$$K(t) = \begin{cases} \frac{1}{72}(1-t)^3(1+3t) & \text{si } 0 \le t \le 1\,; \\ \frac{1}{72}(1+t)^3(1-3t) & \text{si } -1 \le t \le 0\,. \end{cases}$$

D'où il ressort que, dans ce cas, le noyau de Peano est symétrique et positif.

A l'aide de K et du théorème suivant, on détermine l'erreur de quadrature.

Si, pour une méthode d'intégration numérique donnée, l'erreur de quadrature correspondante R est nulle pour tous les polynômes de degré n, alors pour toutes les fonctions $f \in C^{n+1}[a,b]$, on a :

$$R(f) = \int_a^b f^{(n+1)}(t)K(t)\,dt\,,$$

où $f^{(n+1)}$ est la dérivée d'ordre $n+1$ de f.

Si $K(t)$ ne change pas de signe dans l'intervalle d'intégration, on peut utiliser le théorème de la moyenne et obtenir :

$$R(f) = f^{(n+1)}(\theta)\int_a^b K(t)\,dt\,, \quad \theta \in [a,b]\,.$$

L'intégrale portant sur K ne dépend pas de f et peut tre évaluée en substituant à f n'importe quelle fonction telle que sa dérivée d'ordre $n+1$ n'est pas nulle. Le plus simple est de prendre la fonction x^{n+1}, il vient :

$$R(f) = \frac{R(x^{n+1})}{(n+1)!}f^{(n+1)}(\theta)\,.$$

Pour la méthode de Simpson et pour l'intervalle d'intégration $[-1,1]$, on trouve :

$$\frac{R(x^4)}{4!} = \frac{1}{24}\Big(\tfrac{1}{3}(-1)^4 + \tfrac{4}{3}(0)^4 + \tfrac{1}{3}(1)^4 - \int_{-1}^{1} x^4\,dx\Big) = \frac{1}{90}\,.$$

Finalement pour toute fonction f, $f \in C^{n+1}[-1,1]$:

$$\Big(\tfrac{1}{3}f(-1) + \tfrac{4}{3}f(0) + \tfrac{1}{3}f(1) - \int_{-1}^{1} f(x)\,dx\Big) = \frac{1}{90}f^{(n+1)}(\theta)\,, \quad \theta \in [-1,1]\,.$$

Appendice D

Formules d'Adams-Bashforth-Moulton

Pour obtenir le schéma prédicteur/correcteur d'Adams-Bashforth-Moulton d'ordre 2, on suppose connue la fonction y en 2 points (x_1, y_1) et (x_2, y_2), ce qui permet de calculer un interpolant polynômial du premier degré $P_1(x)$ pour $y' = f$. Pour former cet interpolant, on passe naturellement par les formes de Lagrange (voir la relation (4.8)). Ainsi, l'interpolant s'écrit-il

$$\mathcal{L}_2(x) = \sum_{k=1}^{2} y_k L_{2,k}(x) = P_1(x) \equiv f(x, y) \tag{D.1}$$

où $L_{2,1}(x)$ et $L_{2,2}(x)$ sont les coefficients de Lagrange définis par

$$\begin{cases} L_{2,1}(x) = \dfrac{x - x_2}{x_1 - x_2} = -\dfrac{x - x_2}{h} \\ L_{2,2}(x) = \dfrac{x - x_1}{x_2 - x_1} = \dfrac{x - x_1}{h} \end{cases}$$

où $h = x_2 - x_1$. Le prédicteur $\tilde{y}_3$, c'est-à-dire une première estimation de y_3, est obtenu en calculant l'incrément sur l'intervalle $[x_2, x_3]$, soit

$$\begin{aligned} \delta &= \int_{x_2}^{x_3} f(x, y)dx && \text{(D.2)} \\ &= y_1 \int_{x_2}^{x_3} L_{2,1}(x)dx + y_2 \int_{x_2}^{x_3} L_{2,2}(x)dx && \text{(D.3)} \\ &= -\frac{y_1}{h} \int_{x_2}^{x_3} (x - x_2)dx + \frac{y_2}{h} \int_{x_2}^{x_3} (x - x_1)dx && \text{(D.4)} \\ &= -\frac{y_1}{h} \left[\frac{(x - x_2)^2}{2} \right]_{x_2}^{x_3} + \frac{y_2}{h} \left[\frac{(x - x_1)^2}{2} \right]_{x_2}^{x_3} && \text{(D.5)} \\ &= -\frac{y_1}{h} \frac{h^2}{2} + \frac{y_2}{h} \left[\frac{(2h)^2 - h^2}{2} \right] && \text{(D.6)} \end{aligned}$$

soit

$$\delta = -\frac{h}{2} y_1 + \frac{3h}{2} y_2 \tag{D.7}$$

Nous obtenons donc pour le prédicteur

$$\tilde{y}_3 = y_2 - \frac{h}{2}y_1 + \frac{3h}{2}y_2 \tag{D.8}$$

$$\tag{D.9}$$

Pour obtenir le schéma correcteur associé, on utilise les 2 points suivants (x_2, y_2) et $(x_3, \tilde{y}_3)$ pour former un second interpolant polynômial du premier degré $Q_1(x)$ pour f. Grâce aux formes de Lagrange, on a

$$\mathcal{L}'_2(x) = y_2 L'_{2,2}(x) + \tilde{y}_3 L'_{2,3}(x) = Q_1(x) \equiv f(x, y) \tag{D.10}$$

avec

$$\begin{cases} L'_{2,2}(x) = \dfrac{x - x_3}{x_2 - x_3} = -\dfrac{x - x_3}{h} \\ L'_{2,3}(x) = \dfrac{x - x_2}{x_3 - x_2} = \dfrac{x - x_2}{h} \end{cases}$$

où $h = x_3 - x_2$. Sur l'intervalle $[x_2, x_3]$, l'incrément vaut

$$\delta = \int_{x_2}^{x_3} f(x, y)dx \tag{D.11}$$

$$= y_2 \int_{x_2}^{x_3} L'_{2,2}(x)dx + \tilde{y}_3 \int_{x_2}^{x_3} L'_{2,3}(x)dx \tag{D.12}$$

$$= -\frac{y_2}{h} \int_{x_2}^{x_3} (x - x_3)dx + \frac{\tilde{y}_3}{h} \int_{x_2}^{x_3} (x - x_2)dx \tag{D.13}$$

$$= -\frac{y_2}{h} \left[\frac{(x - x_3)^2}{2} \right]_{x_2}^{x_3} + \frac{\tilde{y}_3}{h} \left[\frac{(x - x_2)^2}{2} \right]_{x_2}^{x_3} \tag{D.14}$$

$$= -\frac{y_2}{h}\frac{h^2}{2} + \frac{\tilde{y}_3}{h}\frac{h^2}{2} \tag{D.15}$$

soit

$$\delta = -\frac{h}{2}y_2 + \frac{h}{2}\tilde{y}_3 \tag{D.16}$$

Nous obtenons donc pour le correcteur

$$\tilde{y}_3 = \frac{h}{2}y_2 + \frac{h}{2}\tilde{y}_3 \tag{D.17}$$

$$\tag{D.18}$$

Liste des tables

Liste des figures

Index