Méthodes mathématiques et numériques pour la physique – S5 Licence de physique Parcours Physique Appliquée – Univ Lille

Quentin Thommen

Méthodes mathématiques et numériques pour la physique

S5 Licence de physique
Parcours Physique Appliquée

Version 1.2

23 octobre 2016

Quentin Thommen
Bât P5 Bureau 70 *quentin.thommen@univ-lille1.fr*

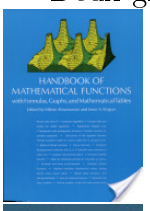
Avertissement

Ce polycopié de cours n'est pas un livre. Il représente une synthèse non exhaustive des méthodes mathématiques et numériques pour la physique. Le texte n'a aucune prétention à l'originalité et il a, pour une grande partie, été pillé dans diverses sources. L'auteur a notamment beaucoup utilisé les ressources numériques telles Wikipédia pour éviter la fastidieuse transcription, des équations en particulier. Certains chapitres très techniques, comme celui sur les polynômes orthogonaux, sont une copie presque intégrale. Le principal travail à consister d'abord à sélectionner, vérifier et uniformiser les contenus puis trouver des exercices adaptés. Dans un sens, ce polycopié est constitué comme un programme qui, tout en incluant quantité de librairies numériques du domaine public, reflète la réflexion propre du concepteur pour atteindre l'objectif fixé.

Bibliographie

Mathématiques

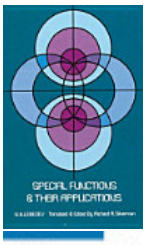
Deux grands classiques



Handbook of mathematical functions with formulas, graphs, and mathematical tables, M. Abramowitz, I. A. Stegun, Dover Publications, 1964 - 1046 pages.

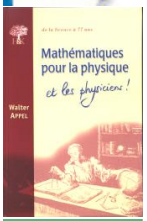
Table of integrals, series and products, I. S. Gradshteyn, I. M. Ryzhik, Academic Press 2007 (seventh edition) 1171 pages.

et des ouvrages utiles

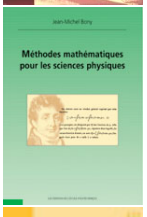


Special functions and their applications, N. N. Lebedev, R. A. Silverman, Dover Publications, 1972 - 308 pages.

Introduction to Perturbation Techniques, A. H. Nayfeh, Wiley 1981, 501 pages



Mathématiques pour la physique et les physiciens, W. Appel, Éditions H&K, 576 pages



Méthodes mathématiques pour les sciences physiques, J. M. Bony, Les éditions de l'école Polytechnique, 2000, 212 pages

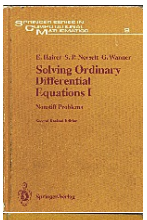
L'analyse au fil de l'histoire, E. Hairer, G. Wanner, Springer 2000, 351 pages.

Analyse numérique

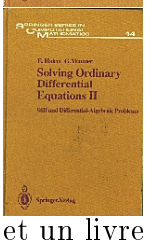
Le grand classique

Numerical recipes : the art of scientific computing, W. H. Press, Cambridge University Press, 2007 (Third Edition) 1256 pages.

Les références mondiales pour l'intégration d'équations différentielles

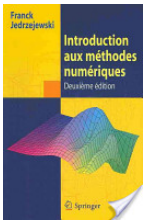


Solving Ordinary Differential Equations I. Nonstiff Problems., E. Hairer, S. P. Norsett, G. Wanner, Springer 1993 (Second edition), 544 pages.



Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems., E. Hairer, G. Wanner, Springer 1996 (Second edition), 630 pages.

et un livre moins complet mais en français



Introduction aux méthodes numériques, F. Jedrzejewski,Springer, 2005 291 pages.

Table des matières

Première partie

Analyse Mathématique

Chapitre 1

Application des nombres complexes à la physique

Sommaire

1.1 Ce qu'il faut retenir

1.1.1 Définitons

Un nombre complexe z est un couple de deux réels a et b de la forme :

$$z = a + ib$$

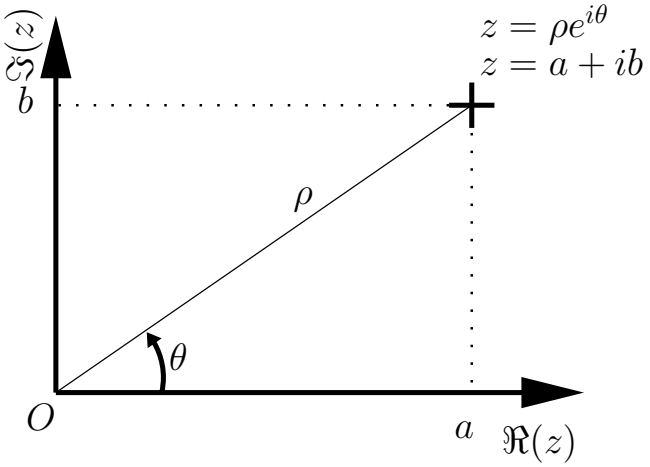
avec i tel que $i^2 = -1$.

On appelle a la *partie réelles* de z ($a = \Re(z)$) et b la *partie imaginaire* de z ($b = \Im(z)$). On parle alors de représentation cartésienne des nombres complexes.

On appelle $z^* = a - ib$ le *complexe conjugué* de z (aussi parfois noté $\bar{z}$).

1.1.2 Représentation graphique

Un nombre complexe z est définit par le couple de réels (a, b), il se représente donc comme un point dans le "*plan complexe*" $(\Re(z), \Im(z))$.



La distance géométrique ρ du point z au point $O = 0 + i0$ est appelée le module de z et noté $\rho = |z|$. L'angle orienté θ est appelé l'argument de z et noté $\theta = \arg(z)$.

1.1.3 Représentation polaire

En utilisant la représentation polaire dans le plan complexe, on déduit simplement : $a = \rho\cos\theta$ et $b = \rho\sin\theta$ on peut donc écrire

$$z = a + ib = \rho\left[\cos\theta + i\sin\theta\right] = \rho e^{i\theta}. \tag{1.1}$$

La dernière relation sera démontrée dans la suite du cours, on l'admet pour l'instant.

Un nombre complexe z peut donc aussi être défini par le couple (ρ, θ) on parle alors de **représentation polaire** de z (mais aussi parfois de représentation géométrique ou exponentielle).

La fonction exponentielle de la définition (1) s'emploie comme la fonction exponentielle usuelle : $e^{i(a+b)} = e^{ia}e^{ib}$; $[e^{ia}]^n = e^{ina}$ ...

1.2 Ce que l'on peut en déduire

Les expressions suivantes se déduisent directement de la définition (1) ou de la représentation dans le plan complexe, essayez de les retrouver !

Sur le module : $\rho = |z| \geq 0$; $|z|^2 = a^2 + b^2 = z.z^*$; $\left|\frac{1}{z}\right| = \frac{1}{|z|}$

Sur l'argument : $\tan\theta = \frac{b}{a}$; $z^* = \rho e^{-i\theta}$; $\cos\theta = \frac{e^{i\theta}+e^{-i\theta}}{2}$; $\sin\theta = \frac{e^{i\theta}-e^{-i\theta}}{2i}$.

Sur les conjugués : $(z^*)^* = z$; $\left(\frac{1}{z}\right)^* = \frac{1}{z^*}$; $(z_1 + z_2)^* = z_1^* + z_2^*$; $(z_1 z_2)^* = z_1^* z_2^*$

1.3 Addition et multiplication de deux complexes

Soit deux nombres complexes z_1 et z_2 définis dans les deux représentations :

$$z_1 = a_1 + ib_1 = \rho_1 e^{i\theta_1}; \quad z_2 = a_2 + ib_2 = \rho_2 e^{i\theta_2}$$

alors :

- $z_1 + z_2 = (a_1 + a_2) + i(b_1 + b_2)$;
- $z_1 z_2 = (a_1 a_2 - b_1 b_2) + i(a_1 b_2 + b_1 a_2) = \rho_1\rho_2 e^{i(\theta_1+\theta_2)}$;
- $\frac{z_1}{z_2} = \frac{\rho_1}{\rho_2} e^{i(\theta_1-\theta_2)}$.

En conclusion, on utilisera de préférence

- **la forme cartésienne pour les additions et les soustractions**
- **la forme polaire pour les multiplications et les divisions**

1.4 Exercices d'applications

1.4.1 Représentation polaire des nombres complexes

1. Donner le module et l'argument des expressions suivantes

   ⋆ $\frac{\omega_0}{\omega_0+i\omega}$ ⋆ $\exp(ia\cos(\omega t))$ ⋆ $\frac{1}{iC\omega}$

2. Pour quels entiers n parmi 2006, 2007, 2008, 2009, le nombre $(1+i)^n$ est–il imaginaire pur ?
3. Donner la forme exponentielle des nombres $z_1 = 1 + e^{2ia}$, $z_2 = e^{ia} + e^{ib}$ et $z_3 = \frac{e^{ia}-1}{e^{ia}+1}$.

1.4.2 Nombres complexes et trigonométrie

1. Calculer les sommes : $C_n = \sum_{k=0}^{n} r^k \cos kx$ et $S_n = \sum_{k=0}^{n} r^k \sin kx$ si k et x sont des paramètres réels.
2. Démontrer que $\cos\left(\frac{\pi}{11}\right) + \cos\left(\frac{3\pi}{11}\right) + \cos\left(\frac{5\pi}{11}\right) + \cos\left(\frac{7\pi}{11}\right) + \cos\left(\frac{9\pi}{11}\right) = \frac{1}{2}$
3. Linéariser $\cos^2 x \sin^2 x$ et $\cos^3 x \sin^4 x$.

1.4.3 Polynômes et racines

Résoudre dans $\mathbb{C}$ les équations d'inconnue z suivantes

⋆ $z^3 = -1$ ⋆ $z - |z|^2 + 1 - i = 0$ ⋆ $z^2 - 2z^\dagger = 0$

et placer les solutions dans le plan complexe.

1.5 Problèmes

1.5.1 Interférences

1.
2. Rappeler la définition mathématique d'une onde électromagnétique progressive sinusoïdale $u(z,t)$ d'amplitude A de pulsation ω_0 se déplaçant dans le vide le long d'une direction notée z.
3. On suppose que cette onde est une onde optique du domaine visible. Rappeler l'ordre de grandeur de la longueur d'onde et de la fréquence de l'onde.
4. On rappelle que l'intensité I de l'onde est proportionnelle à $< u^2(z,t) >$ ou l'opération $<>$ représente la moyenne temporelle sur une durée bien plus longue que la période de l'onde. Montrer que l'intensité est proportionnelle au carré de l'amplitude de l'onde progressive.
5. Rappeler la représentation complexe $\tilde{u}(z,t)$ de l'onde progressive sinusoïdale. A quoi l'intensité est–elle simplement proportionnelle ?
6. On considère deux ondes progressives sinusoïdale de pulsation identique ω_0 et d'amplitudes complexes $\bar{u}_1 = Ae^{i\phi_1}$ et $\bar{u}_2 = Ae^{i\phi_2}$. Quelle est l'intensité lumineuse résultant de la superposition des deux ondes ? A quel type d'expérience le calcul effectué se rattache–t–il ?
7. On suppose maintenant un nombre infini d'ondes présentant des différences de phase identiques et des amplitudes décroissant suivant une progression géométrique ; Leurs amplitudes complexes sont données par $\bar{u}_1 = A$, $\bar{u}_2 = h\bar{u}_1$, $\bar{u}_3 = h\bar{u}_2$, ...avec $h^= re^{i\phi}$ et $r > 1$. Déterminer l'expression de l'intensité du rayonnement résultant de cette superposition. A quel type d'expérience d'optique le calcul effectué se rapporte–t–il ?

1.5.2 Relation de dispersion

1. La représentation complexe du champ électrique associée à la propagation d'une onde électromagnétique plane monochromatique s'écrit $\tilde{E}(z,t) = E_0 e^{i(wt-kz)}$ avec E_0 réel. Donner la direction de propagation.
2. La propagation du champ électrique dans un milieu linéaire est entièrement caractérisé par une relation de dispersion $f(\omega, k) = 0$. Donner la relation de dispersion d'un milieu d'indice n
3. Parmi les relations suivantes, déterminer celles correspondant à des milieux absorbants.
   - ⋆ $\omega^2 + k^2 = 0$;
   - ⋆ $\omega^4 + k^4 = 0$;
   - ⋆ $\omega^4 - k^4 = 0$;
   - ⋆ $\omega^2 + ik - k^2 = 0$.

Chapitre 2

Décomposition d'une fonction en série ; Théorème des résidus

Sommaire

La série de Laurent d'une fonction f est une manière de représenter f au voisinage d'une singularité, ou plus généralement, autour d'un "trou" de son domaine de définition. On représente f comme somme d'une série de puissances (d'exposants positifs ou négatifs) de la variable complexe.

2.1 Séries numériques

Définition : *Soit (u_n) une suite de nombres réels ou complexes. On note $S_n = \sum_{k=0}^{n} u_k$ la série de terme général u_n. La série est dite* ***convergente*** *lorsque la suite S_n est convergente vers S.*

Exercice : Calculer les sommes : $C_n = \sum_{k=0}^{n} r^k \cos kx$ et $S_n = \sum_{k=0}^{n} r^k \sin kx$ si k et x sont des paramètres réels.

2.1.1 Condition nécessaire de convergence :

Si la série $\sum u_n$ converge, alors le terme général u_n tend vers 0 quand n tend vers l'infini.

2.1.2 Critères de convergence d'une série à termes positifs

Soit $\sum u_n$ une **série à termes positifs**, on rappelle les deux règles standards d'étude de la convergence

Règle de Riemann : *Si $n^\alpha u_n$ est majoré avec $\alpha > 1$, alors la série $\sum u_n$ converge.*
Si $n^\alpha u_n$ est minoré par $A > 0$ avec $\alpha \leq 1$, alors la série $\sum u_n$ diverge.

Règle de d'Alembert : *Si $\frac{u_{n+1}}{u_n}$ admet une limite l quand $n \to +\infty$ alors : si $l < 1$, la série converge ; si $l > 1$, la série diverge.*

Pour les séries à termes positifs et négatifs on combine les règles précédentes avec la notion suivante de convergence absolue.

Définition : *Si $\sum |u_n|$ converge, on dit que $\sum u_n$ est **absolument convergente**.*

La convergence absolue est, bien sûr, une condition plus forte que la convergence simple : il est plus "facile" de diverger en n'additionnant que des nombres positifs. En conséquence, toute série absolument convergente sera aussi convergente, mais une série peut converger sans être absolument convergente.

2.1.3 Critères de convergence d'une série alternée

Définition : *Une série $\sum u_n$ à termes réels est alternée si son terme général change de signe alternativement i.e. $u_n = (-1)^n |u_n|$.*

Critère de convergence : *Si la suite de termes positifs (a_n) est décroissante et converge vers 0, alors la série alternée $\sum (-1)^n a_n$ est convergente.*

Exercice : Montrer que la série harmonique alternée $\sum_{n=1}^{n=\infty} \frac{(-1)^{n-1}}{n}$ est convergente, mais n'est pas absolument convergente.

2.2 Séries Entières

Les notions de suite et de série sont fondamentales en analyse numérique, la première servant par exemple pour la résolution d'équations du type $f(x) = 0$ et la seconde pour le calcul numérique d'intégrales du type $\int_a^b f(t)\, dt$.

Plus fondamentalement, elles servent à étayer la théorie mathématique des fonctions. La transition s'effectue *via* la série entière.

2.2.1 Rayon de convergence

Définition : *Une série entière est une série de la forme $\sum_{n=0}^{n=\infty} a_n z^n$ où z est une variable réelle ou complexe et les a_n des constantes réelles ou complexes.*

L'existence et la valeur de la limite S_∞ de la série numérique dépendent de la quantité z qui joue le rôle d'un paramètre. L'ensemble des valeurs de z pour lesquelles la série converge forme la région de convergence. En pratique on utilise la région de convergence absolue qui est un disque centrer sur $z = 0$. En effet, il est évident que la série converge vers 0 pour $z = 0$ et l'utilisation de la convergence absolue de convergence.

Rayon de convergence : *Si $\sum_{n=0}^{n=\infty} a_n z^n$ est une série entière, elle vérifie une et une seule des trois propriétés :*

- ⋆ *la série converge uniquement pour $z = 0$;*
- ⋆ *il existe un nombre réel $R > 0$ tel que la série converge absolument pour tout z tel que $|z| < R$, et diverge pour tout z tel que $|z| > R$;*
- ⋆ *la série converge absolument pour tout z.*

Le nombre R du deuxième cas est appelé rayon de convergence de la série entière.

Exercice : Déterminer de manière générale le rayon de convergence à partir de la règle de d'Alembert.

2.2.2 Rappels sur les développements limités

Définition : *Soit f une fonction définie au voisinage de x_0. On dit que f admet un développement limité d'ordre n au voisinage de x_0 s'il existe une fonction polynôme P_n de degrés inférieur ou égale à n, et une fonction ϵ, définies au voisinage de x_0 telles que :*

$$f(x) = P_n(x) + (x - x_0)^n \epsilon(x) \quad avec \quad \lim_{x \to x_0} \epsilon(x) = 0$$

$P_n(x)$ est la partie régulière et $(x - x_0)^n \epsilon(x)$ le reste.

Exercice : Montrer qu'en posant $x = x_0 - y$, on peut toujours se ramener à l'étude au voisinage de $y = 0$.

Formule de Taylor–Young : *Soit f une fonction dérivable sur I jusqu'à l'ordre n. Alors la fonction ϵ définie au voisinage de 0 par*

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \cdots + \frac{h^n}{n!} f^{(n)}(x_0) + h^n \epsilon(h)$$

est telle que $\lim_{h \to 0} \epsilon(h) = 0$.

La formule de Taylor–Young permet d'obtenir de nombreux développements limités.

Exercice : Retrouver les développements limités des fonctions fonctions suivantes : $\exp(x)$, $\sin(x)$, $\cos(x)$, $(1 + x)^\alpha$, $\ln(1 + x)$, $\tan(x)$.

2.2.3 Développement d'une fonction en série entière

Définition : *Soit f une fonction d'une variable réelle, définie sur un intervalle ouvert U contenant l'origine.*
On dit que f est développable en série entière s'il existe une série entière de rayon de convergence $R \neq 0$ telle que :

$$\forall x \in]-R, R[\bigcap U \quad f(x) = \sum_{n=0}^{n=\infty} a_n x^n.$$

On dit aussi que f est analytique en 0.

Exercice : Montrer que si le développement en série entière de f existe, il correspond au développement de Taylor de f en 0. En déduire l'unicité du développement en série entière.

Exercice : Donner le rayon de convergence des développements en série des fonctions suivantes : $\exp(x)$, $\sin(x)$, $\cos(x)$, $(1 + x)^\alpha$, $\ln(1 + x)$, $\tan(x)$.

Exercice : Démontrer la formule d'Euler $e^{ix} = \cos x + i \sin x$.

Exercice : Montrer que la fonction $f(x) = \frac{x}{1+x}$ ne peut pas être approchée par un polynôme pour tous x. Donner deux polynômes approchant f pour les petites valeurs et pour les grandes valeurs, donner les domaines de validité de ces approximations.

2.3 Fonction holomorphe

Définition : *Soient U un sous-ensemble ouvert (non vide) de l'ensemble $\mathbb{C}$ des nombres complexes et une fonction $f : U \to \mathbb{C}$.*

- *On dit que f est holomorphe (ou dérivable au sens complexe) en un point z_0 de U si la limite suivante, appelée dérivée de f en z_0 existe :*

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

- *On dit que f est holomorphe sur l'ouvert U si elle est holomorphe en tout point z_0 de U. En particulier, on appelle fonction entière une fonction holomorphe dans tout le plan complexe.*

On considère une fonction $f : U \to \mathbb{C}$ d'une variable complexe, où U est un ouvert du plan complexe $\mathbb{C}$. On utilise ici les notations suivantes :

- la variable complexe z est notée $x + i\,y$, où x, y sont réels
- les parties réelle et imaginaire de $f(z) = f(x + i\,y)$ sont notées respectivement $P(x, y)$ et $Q(x, y)$, c'est-à-dire : $f(z) = P(x, y) + i\,Q(x, y)$, où P, Q sont deux fonctions réelles de deux variables réelles.

Propriété : *Les trois propositions suivantes sont équivalentes*

- *f est holomorphe en un point z_0 de U.*
- $\frac{\partial f}{\partial y}(z_0) = i\,\frac{\partial f}{\partial x}(z_0)$
- $\frac{\partial P}{\partial x}(x_0, y_0) = \frac{\partial Q}{\partial y}(x_0, y_0)$ *et* $\frac{\partial P}{\partial y}(x_0, y_0) = -\frac{\partial Q}{\partial x}(x_0, y_0)$

Exercice : Montrer que les fonctions e^z, $\sin z$, $\cos z$ sont des fonctions entières.

Une fonction holomorphe f est analytique, c'est-à-dire développable en série entière au voisinage de chaque point de son domaine de définition. Autrement dit, au voisinage d'un point a où f est définie, on peut écrire $f(z)$ sous la forme :

$$f(z) = \sum_{n=0}^{\infty} a_n (z - a)^n.$$

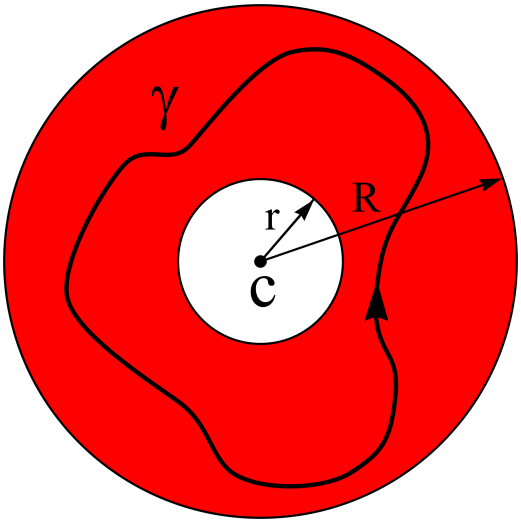
Les séries de Laurent peuvent être vues comme une extension pour décrire f autour d'un point où elle n'est pas (a priori) définie. On inclut les puissances d'exposants négatifs ; une série de Laurent se présentera donc sous la forme :

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z - a)^n.$$

Pour les définir, on a besoin de la notion de couronne.

Définition : *Une couronne centrée en a est un ouvert du plan complexe $\mathbb{C}$ délimité par au plus deux cercles de centre a. En général, une couronne est délimitée par deux cercles de rayons respectifs r, R tels que $r < R$.*

Plusieurs cas dégénérés peuvent toutefois être envisagés :



- Si R vaut l'infini, la couronne considérée est le complémentaire du disque fermé de centre a et de rayon r ;
- Si r vaut 0, la couronne correspond au disque ouvert de centre a et de rayon R, privé de a. On parle aussi dans ce cas de disque épointé ;
- Si r vaut 0 et R l'infini, alors la couronne est le plan complexe privé du point a.

Série de Laurent : *Pour toute fonction holomorphe f sur une couronne C centrée en a, il existe une unique suite $(a_n)_{n\in\mathbf{Z}}$ qui dépend seulement de f telle que :*

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z-a)^n,$$

où la série de fonctions converge normalement sur tout compact de la couronne C. Un tel développement est nommé série de Laurent de la fonction f. Les coefficients a_n de la série de Laurent sont donnés par :

$$a_n = \frac{1}{2\pi i} \oint_\gamma \frac{f(z)\,dz}{(z-a)^{n+1}}$$

où γ est le paramétrage d'un lacet de centre a tracé dans la couronne.
On appelle $f_-(z) = \sum_{n=-\infty}^{-1} a_n (z-a)^n$ la partie principale et $f_+(z) = \sum_{n=0}^{\infty} a_n (z-a)^n$ la partie régulière du développement de f en a.

2.4 Singularité

Une singularité de f est un point z_0 où cette fonction n'est pas analytique, mais l'est en des points dans un voisinage de z_0. La singularité z_0 est isolée si f est analytique en tout point d'un voisinage de z_0 (sauf en z_0). On définit trois types de singularités : singularités superficielles, pôles (ou singularités isolées) et singularités essentielles.

Singularité superficielle : *On dit que f a une singularité superficielle en a si pour tout $n < 0$, les coefficients $a_n = 0$ i.e. la partie principale de f est nulle.*

Pôle d'ordre m : *On dit que f a un pôle d'ordre m en a si pour tout $n < -m$, les coefficients $a_n = 0$ et $a_{-m} \neq 0$. Dans ce cas, $f(z) = \sum_{n=-m}^{\infty} a_n (z-a)^n$.*

Exemple : $f(z) = \frac{1}{z^2+1}$ a deux singularités isolées i et $-i$.

Singularité essentielle : *On dit que f a une singularité essentielle en a si pour tout $n < 0$, les coefficients $a_n \neq 0$.*

Exemple : $f(z) = \exp\left(\frac{1}{z}\right)$ a une singularité essentielle en $z = 0$.

Exercice : Les fonctions $\frac{1}{z^4(1+z^2)}$, $\frac{\sin z}{z}$, $\cos\left(\frac{1}{z}\right)$ possèdent une singularité au point $z_0 = 0$. Pour chacune des fonctions, calculer le développement en série de Laurent en $z = 0$ et déterminer le type de la singularité en $z = 0$.

Résidu

Définition : *Soit f une fonction holomorphe ayant un point singulier en a. On appelle résidu de f au point a, noté $Res\,(f,a)$, le coefficient a_{-1} du développement en série de Laurent de f dans une couronne centrée en a.*

Si f admet un pôle d'ordre m en a alors on montre que

$$Res\,(f, z=a) = \frac{1}{(m-1)!}\lim_{z\to a}\left[\frac{d^{m-1}}{dz^{m-1}}(z-a)^m f(z)\right].$$

Pour deux fonctions f et g à valeurs dans $\mathbb{C}$, on a les relations suivantes :
— Si f a en z_0 un pôle d'ordre 1 : $Res\,(f,z_0) = \lim_{z\to z_0}(z-z_0)f(z)$
— Si f a en z_0 un pôle d'ordre 1 et si g est holomorphe en z_0 : $Res(gf,z_0) = g(a)Res(f,z_0)$
— Si f a en z_0 un zéro d'ordre 1 : $Res\left(\frac{1}{f},z_0\right) = \frac{1}{f'(z_0)}$
— Si f a en z_0 un zéro d'ordre 1 et si g est holomorphe en z_0 : $Res\left(\frac{g}{f},z_0\right) = \frac{g(z_0)}{f'(z_0)}$
— Si f a en z_0 un zéro d'ordre n : $Res\left(\frac{f'}{f},z_0\right) = n$.
— Si f a en z_0 un zéro d'ordre n et si g est holomorphe en z_0 :$Res\left(g\frac{f'}{f},z_0\right) = n\,g(z_0)$.

La motivation de la définition précédente vient du théorème fondamental suivant qui est extrêmement utile pour calculer des intégrales même réelles.

2.5 Théorème des Résidus

Théorème des résidus : *Si f est holomorphe sur et à l'intérieur d'une courbe fermée simple γ, sauf en des singularités isolées $z_1, z_2, \ldots, z_N$ à l'intérieur de γ, alors*

$$\oint_\gamma f(z)dz = 2\pi i\sum_{k=1}^{N} Res\,(f, z=z_k)$$

si γ est parcourue dans le sens direct (sens inverse des aiguilles d'une montre) sinon

$$\oint_\gamma f(z)dz = -2\pi i\sum_{k=1}^{N} Res\,(f, z=z_k).$$

Lemme de Jordan : *Si f est une fonction holomorphe au voisinage d'un point z_0 et possédant un pôle simple en z_0, alors*

$$\int_{\gamma_r} f(z)dz = i\,(\theta_2-\theta_2)\,Res\,(f, z=z_0),$$

où le lacet γ_r est un arc de cercle de rayon r contournant z_0 : $\gamma_r : t\in[\theta_1,\theta_2] \to z_0 + re^{it}$

2.6 Applications du théorème des résidus

Pour évaluer des intégrales réelles, le théorème des résidus s'utilise souvent de la façon suivante : l'intégrante est prolongée en une fonction holomorphe sur un ouvert du plan complexe ; ses résidus sont calculés, et une partie de l'axe réel est étendue à une courbe fermée en lui attachant un demi-cercle dans le demi-plan supérieur ou inférieur. L'intégrale suivant cette courbe peut alors être calculée en utilisant le théorème des résidus. Souvent, la partie de l'intégrale sur le demi-cercle tend vers zéro, quand le rayon de ce dernier tend vers l'infini, laissant seulement la partie de l'intégrale sur l'axe réel, celle que l'on désirait initialement calculer.

2.6.1 Intégrales du premier type

Soit à calculer

$$I = \int_0^{2\pi} R(\cos(t), \sin(t))\, dt,$$

où R est une fonction rationnelle ayant un nombre fini de points singuliers z_n n'appartenant pas au cercle $C(0,1)$ centré à l'origine et de rayon 1. En prenant pour γ le cercle $C(0,1)$ paramétré comme suit

$$\gamma : [0, 2\pi] \to \mathbb{C}, \quad \gamma(t) = e^{it}$$

on obtient par le théorème des résidus

$$I = \oint_\gamma f(z)\, dz = 2i\pi \sum_{|z_j|<1} Res(f, z_j)$$

où f est définie comme

$$f(z) = \frac{1}{iz} R\left(\frac{z + z^{-1}}{2}, \frac{z - z^{-1}}{2i}\right)$$

Exercice : Montrer que si $a > 1$ alors $I = \int_0^{2\pi} \frac{dx}{a+\sin(x)} = \frac{2\pi}{\sqrt{a^2-1}}$

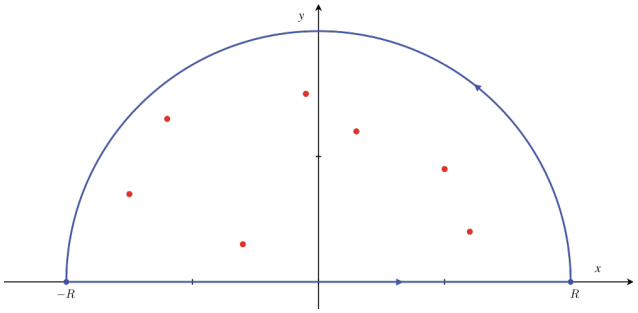
2.6.2 Intégrales du second type

Soit le calcul de l'intégrale réelle suivante :

$$I = \int_{-\infty}^{+\infty} f(x)dx$$

avec $f(z)$ ayant un ensemble de points singuliers isolés z_n purement complexes (en dehors de l'axe des réels) et $\lim_{|z|\to 0} |z||f(z)| = 0$ ce qui garanti la décroissance rapide de f pour les grands modules.

Pour le lacet γ l'axe des réels est complété d'un demi-cercle de rayon infini (voir ci–contre). La fonction décroissant suffisamment rapidement quand $|z| \to \infty$, cet ajout ne modifie pas la valeur de l'intégrale.



L'application du théorème des résidus donne

$$I = 2i\pi \sum_{\Im(z_j)>0} \mathrm{Res}(f, z_j) = -2i\pi \sum_{\Im(z_j)<0} \mathrm{Res}(f, z_j)$$

Exercice : Montrer que si $a > 0$ alors $I = \int_{-\infty}^{+\infty} \frac{dx}{x^2+a^2} = \frac{\pi}{a}$.

2.6.3 Intégrales du troisième type

Soit le calcul de l'intégrale réelle suivante :

$$I = \int_{-\infty}^{+\infty} f(x)\mathrm{e}^{iax}\, dx$$

avec $f(z)$ comportant un ensemble de points singuliers isolés purement complexes. Si

$$\exists M, R > 0 \quad \text{tels que} \quad |f(z)| \leq \frac{M}{|z|} \quad \forall |z| \geq R,$$

alors :

$$(\text{si } a > 0), \quad I = 2i\pi \sum_{\Im(z_j)>0} \text{Res}\left(f(z)\mathrm{e}^{iaz}, z_j\right)$$

et

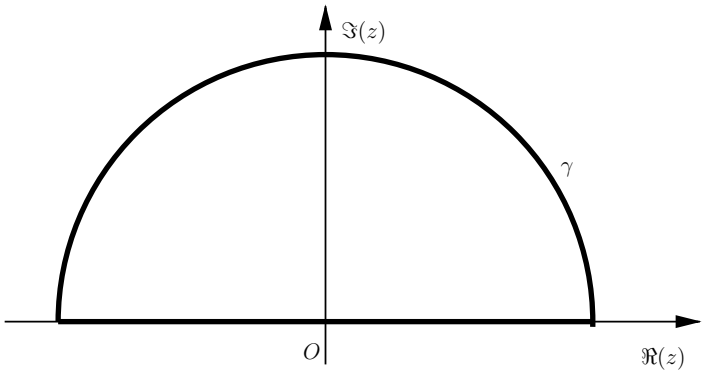
$$(\text{si } a < 0), \quad I = -2i\pi \sum_{\Im(z_j)<0} \text{Res}\left(f(z)\mathrm{e}^{iaz}, z_j\right)$$

Exercice : Montrez que $I = \int_{-\infty}^{+\infty} \frac{\exp(bix)dx}{a^2+x^2} = \frac{\pi}{a}\exp(-ab)$ avec a et b positifs.

Exercices

1 - Intégrale impropre

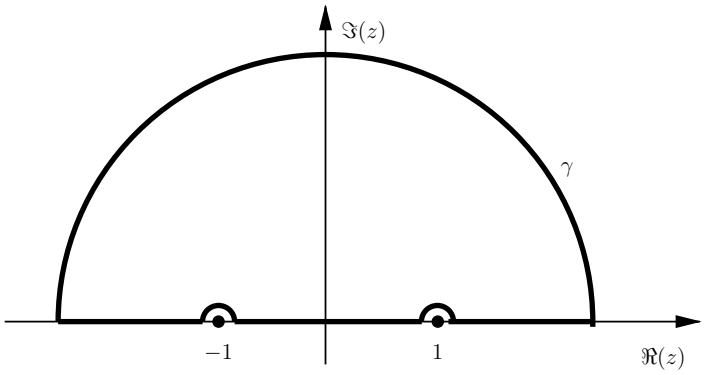
1. Montrer que la fonction $\frac{1}{z^6+1}$ ($z \in \mathbb{C}$) possède six pôles simples.
2. En utilisant le théorème des résidus, calculer l'intégrale
$$\int_{-\infty}^{\infty} \frac{1}{x^6+1}dx$$
3. Justifier l'utilisation du lacet γ ci-contre.



2 - Intégrale de Fourier

Soit k un paramètre réel positif

1. Montrer que la fonction $\frac{e^{ikz}}{z^2-1}$ est holomorphe dans le plan complexe sauf en ses pôles d'ordre 1 : $z = \pm 1$
2. En utilisant le théorème des résidus, calculer l'intégrale
$$\int_{-\infty}^{\infty} \frac{e^{ikx}}{x^2-1}dx$$
3. Justifier l'utilisation du lacet γ ci-contre.

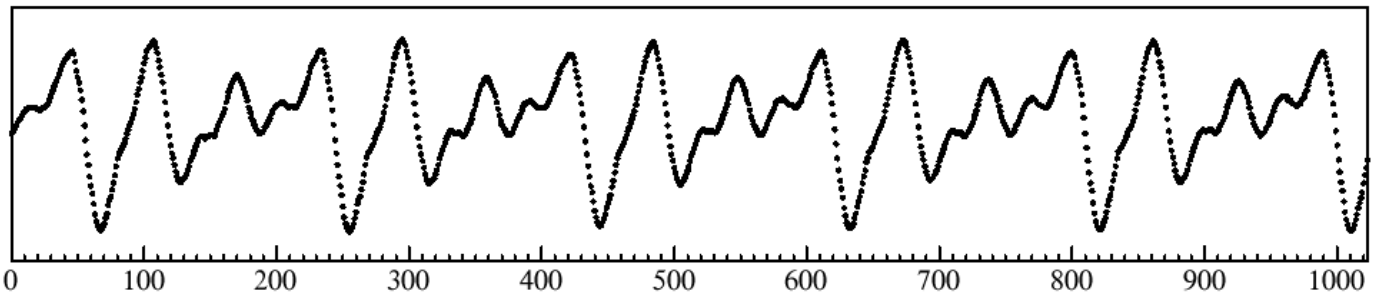


Chapitre 3

Séries de Fourier et systèmes orthogonaux

Sommaire

La théorie de ce chapitre permet de mieux comprendre toute sorte de phénomènes *périodiques*. Elle a son origine au 18e siècle dans l'interpolation de fonctions périodiques en astronomie, dans l'étude de la corde vibrante et du son avant d'entrer en force en science grâce à la *Théorie de la Chaleur* de Fourrier (1822).



Comme exemple, considérons la numérisation d'un son (le son "o"). On a enregistré 22000 impulsions par secondes, dont 1024 sont dessinées. Il n'y a pas de doute que ces données représentent un phénomène périodique. On est souvent intéressé par l'étude du spectre d'un tel signal, par les fréquences dominantes, par la suppression d'un bruit de fond éventuel, etc.

3.1 Séries de Fourier et polynômes trigonométriques

Commençons par quelques généralité su les fonctions périodiques.

Fonctions périodiques. : *Les fonctions* $\sin x$, $\cos x$, *mais aussi* $\sin 2x$, $\cos 5x$ *sont des fonctions* 2π *–périodiques, c–à–d. elles vérifient la relation*

$$f(x+2\pi) = f(x) \quad \forall x \in \mathbb{R}$$

De manière générale, on dit qu'une fonction f *est* L*–périodique si*

$$f(x+L) = f(x) \quad \forall x \in \mathbb{R}$$

De la périodicité de la fonction f découle directement l'invariance par translation de l'intégrale, propriété bien utile pour le calcul des coéfficients de Fourier. Plus précisément, si f une fonction T–périodique alors

$$\int_{a+c}^{b+c} f(x)\,dx = \int_a^b f(x)\,dx \quad \forall c \in \mathbb{R}.$$

3.1.1 Polynômes trigonométriques

De même que le polynôme est le propotype des fonctions, le polynôme trigonométique est le prototype des fonctions périodiques.

Polynôme trigonométrique. : *Les combinaisons linéaires de* $\sin kx$ *et de* $\cos kx$ *(*$k \in \mathbb{Z}$*) sont des fonctions* 2π *–périodiques. Elles sont de la forme*

$$\frac{a_0}{2} + \sum_{k=1}^{N} a_k \cos kx + \sum_{k=1}^{N} b_k \sin kx,$$

où le a_k *et* b_k *sont des coefficients réels. On appelle cette forme un polynôme trigonométrique.*

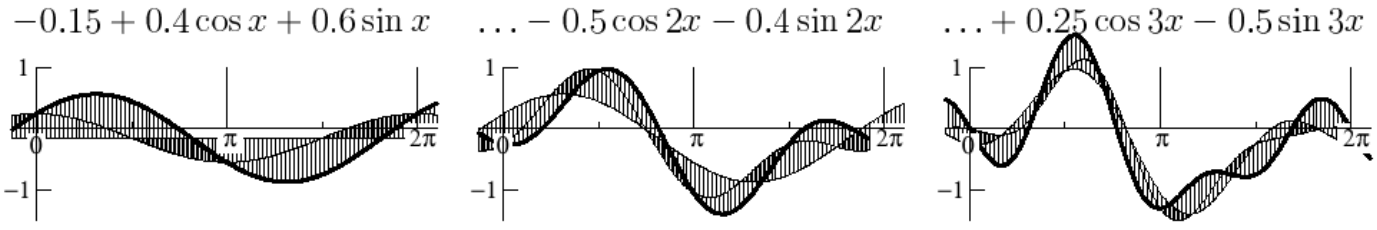


FIGURE 3.1 – Exemples de polynômes trigonométriques

Remarque : Attention si tous les polynômes trigonométriques sont bien des fonctions 2π –périodiques cela ne veut pas dire a priori que toutes les fonctions 2π –périodiques sont des polynômes trigonométriques.

La formule d'Euler

$$e^{ix} = \cos x + i \sin x$$

permet de simplifier l'expression du polynôme trigonométrique. En effet, en additionnant et soustrayant e^{ix} et $e^{-ix} = \cos(-x) + i\sin(-x) = \cos x - i \sin x$, on en déduit

$$\cos x = \frac{e^{ix} + e^{-ix}}{2}, \quad \sin x = \frac{e^{ix} - e^{-ix}}{2i}.$$

Le polynôme trigonométrique peut donc être écrit sous la forme plus élégante

$$\sum_{k=-N}^{N} c_k e^{ikx}, \quad \text{avec} \quad c_k \in \mathbb{C}.$$

Les relations entre les coefficients a_k et b_k d'un côté et les coefficients c_k de l'autre s'obtiennent par identification :

$$c_k = \frac{1}{2}(a_k - ib_k), \quad c_{-k} = \frac{1}{2}(a_k + ib_k)$$

où de manière équivalente

$$a_k = c_k + c_{-k}, \quad b_k = i(c_k + c_{-k}).$$

Avec ces formules, on peut passer de la représentation réelle à la représentation complexe et vice–versa.

3.1.2 Série de Fourier

Considérons en prémice un polynôme trigonométrique $f(t) = \sum_{k=-N}^{N} c_k e^{i\,k\,2\pi\frac{t}{T}}$ et montrons que l'on peut en déterminer les coefficients par un calcul intégrale. En multipliant le polynôme trigonométrique par $e^{-i\,k\,2\pi\frac{t}{T}}$ et en intégrant de 0 à T on trouve les égalités suivantes

$$c_k = \frac{1}{T}\int_0^T f(t)\exp\left(-i\,k\,2\pi\frac{t}{T}\right)dt$$

$$a_k = \frac{2}{T}\int_0^T f(t)\cos\left(2\pi k\frac{t}{T}\right)\,dt$$

$$b_k = \frac{2}{T}\int_0^T f(t)\sin\left(2\pi k\frac{t}{T}\right)\,dt$$

Propriété : Le coefficient a_0 correspond à la moyenne de la fonction sur une période.

Propriété : Les coefficients de Fourier sont calculables par une intégration -$T/2$ à $T/2$, ou de a à $T + a$.

Si $f(x)$ est 2π –périodique, mais pas nécessairement un polynôme trigonométrique, on appelle a_k, b_k et c_k les *coefficients de Fourier* de la fonction $f(x)$.

Série de Fourier : *Soit $f(x)$ une fonction L–périodique telle que les intégrales définissant les coefficients de Fourier existent. On appelle "série de Fourier de $f(x)$" la série*

$$\sum_{k=-\infty}^{\infty} c_k e^{i\,k\,2\pi\frac{x}{L}} \quad \text{où} \quad c_k = \frac{1}{L}\int_0^L f(x)e^{-i\,k\,2\pi\frac{x}{L}}dx$$

et on écrit $f(x) \approx \sum_{k=-\infty}^{\infty} c_k e^{i\,k\,2\pi\frac{x}{L}}$. La série de Fourier peut également être écrite sous la forme

$$f(x) \approx \frac{a_0}{2} + \sum_{k>0} a_k \cos k\,2\pi\frac{x}{L} + \sum_{k>0} b_k \sin k\,2\pi\frac{x}{L},$$

avec

$$a_k = \frac{2}{L}\int_0^L f(x)\cos\left(2\pi k\frac{x}{L}\right)\,dx \quad b_k = \frac{2}{L}\int_0^L f(x)\sin\left(2\pi k\frac{x}{L}\right)\,dx$$

Propriété : Si $f(x)$ est une fonction paire T–périodique on a $b_k = 0$, et si f est impaire $a_k = 0$.

Propriété : Si $f(x)$ est une fonction réelle T–périodique, alors $c_{-k} = c_k^\dagger$.

On appelle spectre de la fonction f la représentation graphique, par un diagrame en baton des modules de cefficients de Fourier en fonction de l'indice. Si la fonction est réelle on ne représente que les indices positifs.

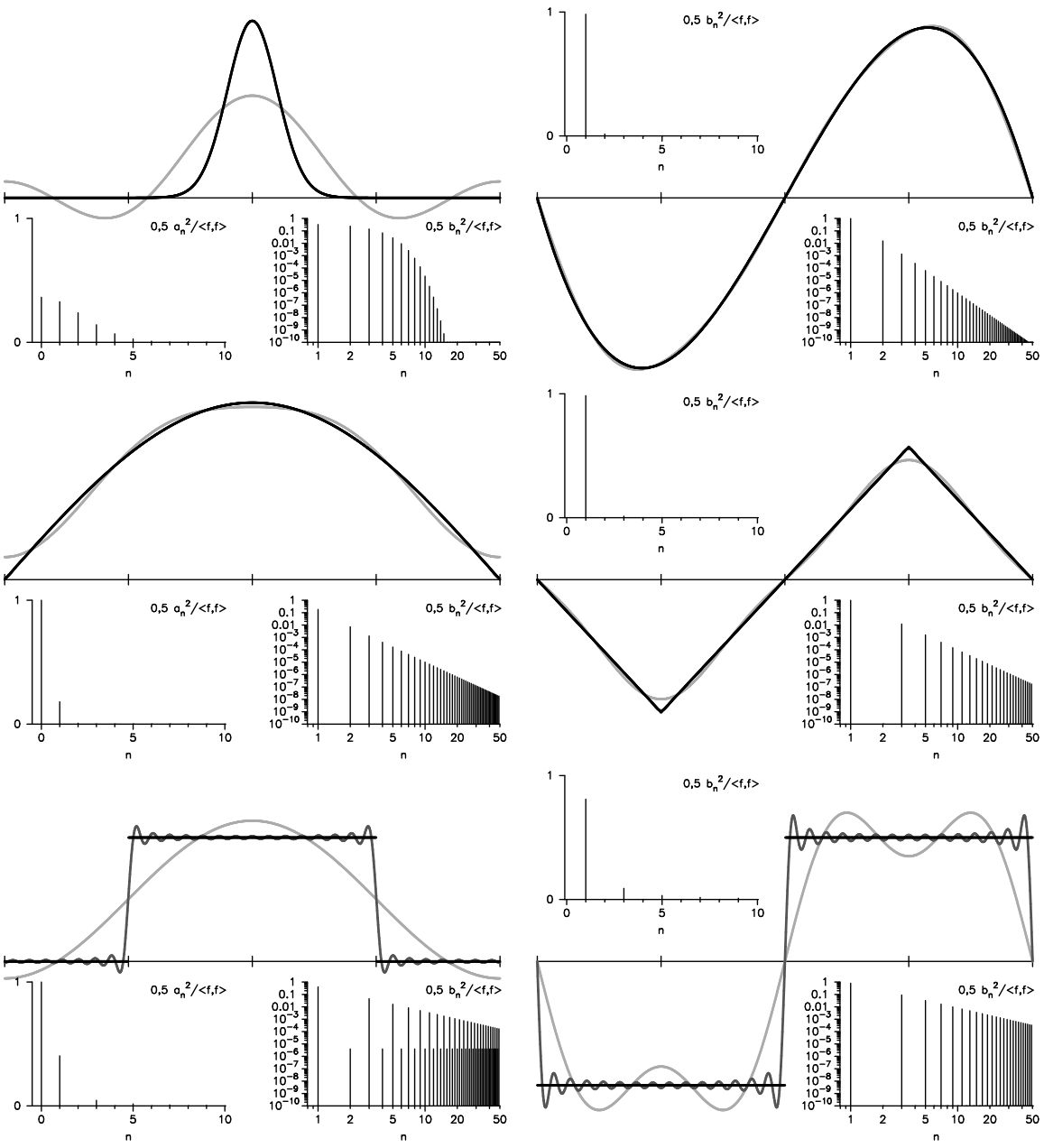
Pour le moment, on sait seulement qu'on a égalité dans $f(x) = \sum_{k=-\infty}^{\infty} c_k e^{i2\pi k \frac{x}{L}}$ pour des polynômes trigonométriques. Le sujet de ce chapitre est d'étudier la convergence de cette série et la question quand cette identité reste vraie pour des fonctions L–périodiques arbitraires.

Étudions comment une fonction $f(x)$ est approchée par sa série de Fourier. La figure suivante montre six fonctions L–périodiques ainsi que plusieurs troncatures de leurs séries de Fourier associées. Exercez–vous à calculer leurs coefficients de Fourier. Les six fonctions sont :

$$f(x) = \exp(10\cos(2\pi x/L)) \qquad f(x) = x\,\left(\tfrac{L}{2} - x\right)\,\left(\tfrac{L}{2} + x\right)$$

$$f(x) = |\cos(\pi \tfrac{x}{L})| \qquad f(x) = \begin{cases} x & \text{si } 0 \le |x| < L/4; \\ L/2 - x & \text{si } L/4 \le x < L; \\ -l/2 + x & \text{si } -L/2 \le x < L/4. \end{cases}$$

$$f(x) = \begin{cases} 1 & \text{si } |x| < L/4; \\ 0 & \text{sinon.} \end{cases} \qquad f(x) = \begin{cases} 1 & \text{si } t \ge 0; \\ -1 & \text{si } t < 0; . \end{cases}$$

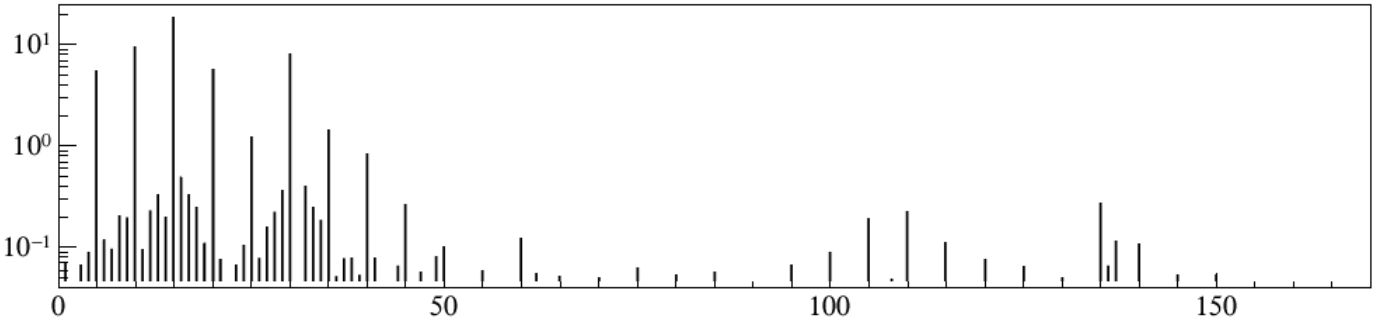


3.1.3 Application de la série de Fourier à la compression des données

Étudions encore la fonction du début de ce chapitre qui est la digitalisation d'un son. Sur l'intervalle comprenant tous les 1024 points elle n'est visiblement pas périodique. Par contre, les premiers 944 points représentent une fonction qui peut être prolongée périodiquement. Sa période est $T = \frac{944}{22000} s$. Si on dénote cette fonction par $F(t)$ (t en secondes), la fonction $f(x) = F(t)$ avec $x = 2\pi t/T$ devient 2π –périodique et on peut appliquer les formules de ce paragraphe. On cherche donc une représentation

$$F(t) \approx \sum_{k=-\infty}^{\infty} c_k e^{ikx} \quad \text{avec} \quad x = 2\pi t/T$$

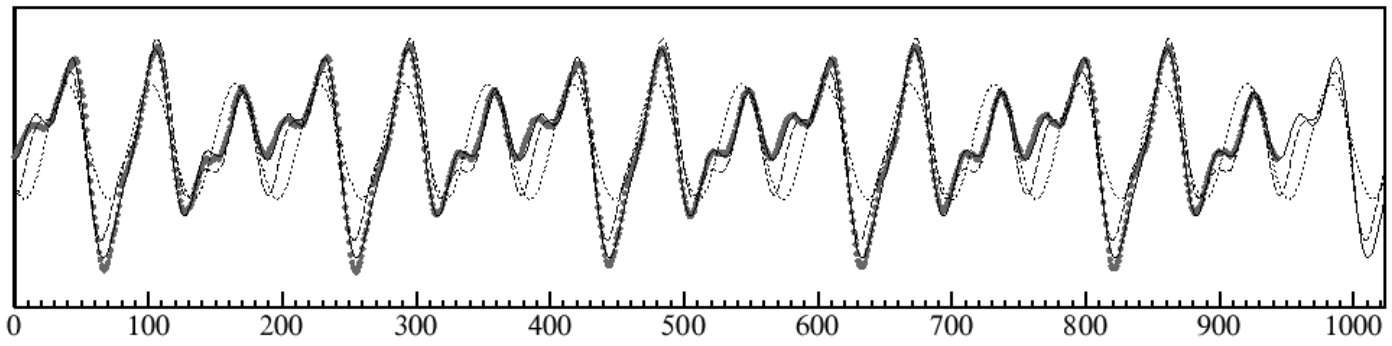
La figure suivante montre les modules de c_k en fonction de k que l'on a calculés numériquement. Comme $f(x)$ est une fonction réelle, on a $c_{-k} = c_k^\dagger$ de sorte que l'on a représenté que les k positifs. On observe que les coefficients de Fourier dominants correspondent tous à des multiples de 5. La fréquence dominante de ce son est donc de $5/T \simeq 116,5\text{Hz}$.



On peut maintenant essayer de retrouver le son à partir de son spectre (c–à–d. à partir des coefficients de Fourier c_k). Les valeurs de c_k, utilisées pour dessiner quelques séries de Fourier tronquées, sont données dans le tableau suivant. D'abord, on prend uniquement les termes avec $k = \pm 15$, ceci donne la fonction pointillée (un sinus pur). Si on tient en plus compte des coefficients avec $k = \pm 10$ et $k = \pm 30$ on obtient la fonction en tirée, et en ajoutant les termes correspondants à $k = \pm 5$ et $k = \pm 20$ on obtient la courbe solide. Elle est déjà une très bonne approximation du son actuel. On voit qu'avec très peu d'information on peut reconstituer le signal original.

k	c_k	$\lvert c_k\rvert$
5	$5.61 - 0.35\,i$	5.62
10	$4.27 - 8.71\,i$	9.70
15	$-13.53 + 12.82\,i$	18.64

k	c_k	$\lvert c_k\rvert$
20	$5.45 - 1.93\,i$	5.78
25	$0.31 - 1.19\,i$	1.23
30	$-7.84 - 2.18\,i$	8.14



Ce principe général de compression du signal est utilisé pour l'encodage ogg et mp3 des fichiers de son, pour l'encodage jpeg des images et mpeg des vidéos.

L'échantillonage du signal et sa durée d'enregistrement limité, impose une borne supérieur au nombre de coefficients de Fourier que l'on peut déterminer. Le signal d'entrée n'est connu qu'au points de mesure,

bien qe l'œil ait tendance à lisser la courbe en reliant les points de mesure par des segments de droite, rien ne justifie un tel lissage ; entre deux points de mesure, le signal nous est inconnu. Le contennu informationnel du signal vaut donc 944 valeurs réelles, on ne peut donc déterminer les coefficients de Fourier jusqu'à l'harmonique $n = 471$.

3.2 Convergence de la série de Fourier

Jusqu'ici tout va bien, on semble en mesure de pouvoir décrire correctement les fonctions périodiques par une liste finie de nombres. Il reste néanmoins à étayer mathématiquement les questions fondamentales que l'on peut résumer comme suit :

Si l'on se donne une fonction "arbitraire" $f(x)$ sur $[0, 2\pi]$, et si l'on calcule les coefficients a_k, b_k ou les c_k, on peut se demander si :

— la série de Fourier va converger ?, converger uniformément ?

— la série de Fourier va converger vers $f(x)$.

Ces questions, affirmées dans un élan de jeunesse par Fourier à partie de 1807, se sont avérées par la suite plus difficiles que prévues.

3.2.1 Comportement asynptotique des coefficients de Fourier

Une première approche pour étudier la convergence de la série de Fourier consiste à dériver des majorations pour $|c_k|$ et d'utiliser le fait que $|c_k e^{ikx}| \leq |c_k|$.

Lemme : *Soit $e : [0, 2\pi] \to \mathbb{R}$ une fonction en escalier sur un nombre fini d'intervalles, alors les coefficients de Fourier sont majorés par*

$$|c_k| \leq \frac{Const}{|k|}, \quad |a_k| \leq \frac{Const}{|k|}, \quad |b_k| \leq \frac{Const}{|k|} \quad \forall k.$$

Ce resultat simple à obtenir n'est pas suffisant pour établir la convergence absolue de la série de Fourier. Le résultat suivant est plus général dans les hypothèses sur la fonction mais n'est toujours pas suffisant pour la convergence.

Lemme de Riemann : *Soit $f : [a, b] \to \mathbb{R}$ intégrable au sens de Riemann alors*

$$\lim_{k\to\infty} \int_a^b f(x) \sin kx dx = 0, \quad \lim_{k\to\infty} \int_a^b f(x) \cos kx dx = 0, \quad \lim_{k\to\infty} \int_a^b f(x) e^{\pm ikx} dx = 0.$$

Ceci implique que les coefficients de Fourier satisfont $a_k \to 0$, $b_k \to 0$ et $c_k \to 0$.

Le résultat principal pour définir la décroissance asymptotique des coefficients implique la notion de fonction à variation bornée que l'on ébauche par les quelques remarques définissantes suivantes

— Une fonction est à variation bornée si et seulement si elle est différence de deux fonctions monotones croissantes,

— si $f : [a, b] \to \mathbb{R}$ est à variation bornée, alors f est intégrable au sens de Riemann,

— Si f est continuent différentiable, alors f est à variation bornée,

— Une fonction f peut être à variation bornée sans être continue (fonction en escalier),

— Une fonction f peut être continue sans être à variation bornée ($x \sin(1/x)$ sur [0,1]).

On a alors le résultat suivant

Théorème : *a) si $f : [0, 2\pi] \to \mathbb{R}$ est à variation bornée, alors les coefficients de Fourier satisfont pour $k \neq 0$*

$$|c_k| \leq \frac{Const}{|k|}, \quad |a_k| \leq \frac{Const}{|k|}, \quad |b_k| \leq \frac{Const}{|k|}.$$

b) si $f : \mathbb{R} \to \mathbb{R}$ est 2π —périodique, p-1 fois continûment différentiable et p fois différentiable par morceaux avec $f^{(p)}|_{[0,2\pi]}$ à variation bornée alors

$$|c_k| \leq \frac{Const}{|k|^{p+1}}, \quad |a_k| \leq \frac{Const}{|k|^{p+1}}, \quad |b_k| \leq \frac{Const}{|k|^{p+1}}.$$

Parmi les six exemples présentés en tête de ce chapitre , les deuxième et cinquième fonctions sont à variation bornée, mais elles ne sont pas continues sur tout l'intervalle. On comprend alors pourquoi les coefficients de Fourier diminuent comme $Const/k$. La quatrième fonction n'est pas bornée et donc pas à variation bornée. Cela n'empêche pas les coefficients de diminuer aussi comme $Const/k$. Les troisième et sixième fonctions possèdent une première dérivée qui est à variation bornée, d'où un comportement en $Const/k^2$. La première fonction est à variation bornée, mais la dérivée ne l'est pas.

3.2.2 Étude élémentaire de la convergence

Après avoir établi quelques règles générales sur la décroissance des coefficients de Fourier, on plonge maintenant dans l'étude de la convergence de la série de Fourier.

Comme $\left|e^{ikx}\right| = 1$, le critère du quotient (Weierstrass) garanti que la série est uniformément convergente si la série des c_k converge absolument. Par conséquent, sous la condition (b) du théorème précédent, il est certain que la série de Fourier converge uniformément sur $[0, 2\pi]$ vers une fonction continue. Mais, on ne sait pas encore si elle converge vraiment vers la fonction $f(x)$. Dans le cas où $|c_k| \simeq C/k$ (cas (a) du théorème précédent), ce raisonnement ne peut pas être appliqué.

Pour étudier la convergence vers $f(x)$ de la série de Fourier, on considère les sommes partielles

$$S_n(x) = \sum_{k=-n}^{n} c_k e^{ikx} \quad \text{où} \quad c_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx.$$

Un résultat intéressant avec une preuve extrêmement simple a été publié par P.R. Chernoff en 1980.

Théorème : *Soit $f(x)$ intégrable sur $[0, 2\pi]$ et différentiable au point $x_0 \in [0, 2\pi]$. Alors les sommes partielles $S_n(x_0)$ convergent vers $f(x_0)$ si $n \to \infty$.*

3.2.3 Théorèmes de Dirichlet et égalité de Parseval

Le résultat suivant est admis

Théorème (Dirichlet 1829) : *Soit $f(x)$ une fonction 2π –périodique et $f|_{[0,2\pi]}$ à la variation bornée. Alors la série de Fourier converge pour tout $x_0 \in \mathbb{R}$ et*

- *si f est continue en x_0, on a $\lim_{n\to\infty} S_n(x_0) = f(x_0)$,*
- *si f est discontinue en x_0, $\lim_{n\to\infty} S_n(x_0) = \frac{1}{2}\left(f(x_0-) + f(x_0+)\right)$*

Au dela de l'interet du théorème de Dirichlet en ce qui concerne la convergence de la série de Fourrier, il met en lumière les diffcultés de convergence pour les fonction discontinue conduisant au phénomène de Gibbs.

Égalité de Parseval : *Soit $f(x)$ une fonction L-périodique de carré sommable, alors*

$$\sum_{n=-\infty}^{+\infty} |c_n|^2 = \frac{1}{L}\int_0^L |f(x)|^2 \,\mathrm{d}x.$$

Dans de nombreuses applications physiques (courant électrique par exemple), L'égalité de Parseval peut s'interpréter comme suit : l'énergie totale s'obtient en sommant les contributions des différents harmoniques.

3.3 Généralisation : systèmes orthogonaux

3.3.1 Définitions et théorèmes fondamentaux

Pour se rapprocher d'une théorie plus générale, et pour simplifier les notations, on considère l'espace des fonctions intégrables sur l'intervalle $[a, b]$:

$$\mathbb{R}([a,b],\mathbb{C}) := \{f : [a,b] \to \mathbb{C};\ f \text{intégrable sur } [a,b]\}$$

muni du produit scalaire (forme sesquilinéaire semi-définie positive)

$$\langle f, g\rangle = \int_a^b f(x)\, g^\dagger(x)\, \omega(x)\, dx.$$

On désigne par

$$||f||_2 = \sqrt{\langle f, f\rangle}$$

la norme correspondante.

On dispose alors des relations de base dans un espace métrique

— l'inégalité de Cauchy–Schwarz $|\langle f, g\rangle| \leq ||f||_2 ||g||_2$

— le théorème de Pythagore : si f et g sont orthogonales c–à–d. $\langle f, g\rangle = 0$ alors $(||f+g||_2)^2 = (||f||_2)^2 + (||g||_2)^2$.

— $||f||_2 = 0$ implique $f = 0$.

Système orthogonal

Définition : *Une suite de fonction $\{\varphi_k\}_{k\leq 0}$ dans $\mathbb{R}([a,b],\mathbb{C})$ forme un système orthogonal si*

$$\langle \varphi_k, \varphi_l\rangle = 0 \quad \textit{pour} \quad k, l \leq 0,\ k \neq l.$$

On a un système orthonormal si de plus $||\varphi_k||_2 = 1$ pour tout $k \leq 0$.

Exercice : Montrer que le système $\{1, e^{ix}, e^{-ix}, e^{i2x}, e^{-i2x}, \ldots\}$ est orthogonal sur $\mathbb{R}([0, 2\pi],\mathbb{C})$.

Exercice : Montrer que le système $\{1, \cos x, \cos 2x, \cos 3x, \ldots\}$ est orthogonal sur $\mathbb{R}([0,\pi],\mathbb{C})$.

Soit $\{\varphi_k\}_{k\leq 0}$ un système orthonormal (un système orthogonal peut toujours être normalisé) et soit f une combinaison linéaire finie de la forme $f = \sum_{k=0}^n c_k\varphi_k$. En prenant le produit scalaire avec φ_j et en utilisant orthonormalité du système on obtient la relation $c_j = \langle f, \varphi_j\rangle$ pour les coefficients. Ceci est la motivation pour généraliser les séries de Fourier à des systèmes orthogonaux arbitraires.

Série de Fourier généralisée

Définition : *Soit $\{\varphi_k\}_{k\leq 0}$ un système orthonormal dans $\mathbb{R}([a,b],\mathbb{C})$ et soit $f \in \mathbb{R}([a,b],\mathbb{C})$. On appelle*

$$f \simeq \sum_{k\geq 0} c_k\varphi_k \quad \textit{avec} \quad c_k = \langle f, \varphi_k\rangle$$

série de Fourier, et les c_k sont les coefficients de Fourier.

Théorèmes d'optimalité des séries de Fourier

Le résultat suivant montre que la série de Fourier tronquée possède une propriété d'optimalité. Elle est la meilleure approximation de f pour la norme $||.||_2$ dans la classe des fonctions qui s'expriment comme des combinaisons linéaires de φ_k jusqu'au degrés n.

Théorème : *Soit $\{\varphi_k\}_{k\leq 0}$ un système orthonormal dans $\mathbb{R}([a,b],\mathbb{C})$, soit $f \in \mathbb{R}([a,b],\mathbb{C})$ et soit $\sum_{k\geq 0} c_k\varphi_k$ sa série de Fourier. Pour tour $n \geq 0$ et pour tout $d_0 \dots d_n$ on a*

$$\left\| f - \sum_{k=0}^{n} c_k\varphi_k \right\|_2 \leq \left\| f - \sum_{k=0}^{n} d_k\varphi_k \right\|_2$$

En particulier, parmi tous les polynômes trigonométriques de degrés n T_n, celui pour lequel

$$\int_0^{2\pi} |f(x) - T_n(x)|^2 \, dx \longrightarrow min$$

est la série de Fourier tronquée.

Théorème : *Soit $\{\varphi_k\}_{k\leq 0}$ un système orthonormal dans $\mathbb{R}([a,b],\mathbb{C})$, soit la fonction f intégrable sur $[a,b]$ et soit $\sum_{k\geq 0} c_k\varphi_k$ sa série de Fourier. Alors*

$$\sum_{k\geq 0} |c_k|^2 \leq ||f||_2^2 \quad \textit{(inégalité de Bessel)}$$

et en particulier la série $\sum_{k\geq 0} |c_k|^2$ converge. On a

$$\underbrace{\sum_{k\geq 0} |c_k|^2 = ||f||_2^2}_{\textit{égalité de Parseval}} \Longleftrightarrow \left\| f - \sum_{k=0}^{n} \langle f, \varphi_k\rangle \, \varphi_k \right\|_2 \to 0 \textit{ quand } n \to \infty.$$

Système complet

Définition : *On dira qu'un système orthonormal $\{\varphi_k\}_{k\leq 0}$ dans $\mathbb{R}([a,b],\mathbb{C})$ est complet (ou total) si pour toute fonction $f \in \mathbb{R}([a,b],\mathbb{C})$ on a*

$$\left\| f - \sum_{k=0}^{n} c_k\varphi_k \right\|_2 \to 0 \textit{ quand } n \to \infty$$

3.3.2 Exemple des Polynomes de Legendre

On considère dans ce paragraphe l'espace des fonctions réelles intégrables sur $[-1,1]$. On veut former une base de polynomes orthogonaux $P_n(x)$ en utilisant le produit scalaire

$$\langle f, g\rangle = \int_{-1}^{1} f(x)\, g(x)\, dx.$$

Il est naturel de partir de la base cannonique $\{1, x, x^2, x^3, \dots\}$ pour former les polynomes recherchés. En notant $\varphi_n(x) = x^n$, on remarque immédiatement que la base canonique n'est pas orthogonale :

$$\langle \varphi_n, \varphi_m\rangle = \int_{-1}^{1} x^n \, x^m \, dx = \left[\frac{x^{n+m+1}}{n+m+1}\right]_{-1}^{1} = \frac{1^{n+m+1} - (-1)^{n+m+1}}{n+m+1} \neq 0 \quad \forall (n,m) \in \mathbb{N}^2.$$

Par contre, on peut utiliser le procédé de Gramm–Schmidt pour l'orgothonaliser, les premier développement donnent

$$\begin{array}{rcll} P'_0(x) & = & \varphi_0(x) = 1 & \\ P'_1(x) & = & \varphi_1(x) = x & P'_1 \perp P'_0 \\ P'_2(x) & = & \varphi_2(x) - \langle P'_0, \varphi_2\rangle \varphi_0(x) = x^2 - \frac{2}{3} & P'_2 \perp P'_1, P'_0 \\ P'_3(x) & = & \varphi_3(x) - \langle P'_1, \varphi_3\rangle \varphi_1(x) = x^3 - \frac{2}{5}x & P'_3 \perp P'_2, P'_1, P'_0. \end{array}$$

puis on normalise chacun des polynomes en posant $P_n = P'_n/\sqrt{\langle P'_n, P'_n\rangle}$ soit

$$\begin{array}{rcl} P_0(x) & = & 1 \\ P_1(x) & = & x \\ P_2(x) & = & \dfrac{3x^2-1}{2} \\ P_3(x) & = & \dfrac{5x^3-3x}{2} \\ P_4(x) & = & \dfrac{35x^4-30x^2+3}{8} \\ \cdots & & \end{array}$$

On obtient l'ensemble des polynômes par la relation de récurence

$$P_{n+1}(x) = \frac{2n+1}{n+1}\, x\, P_n\,(x) - \frac{n}{n+1} P_{n-1}\,(x)\,,$$

et Lles polynomes P_n sont tous orthogonaux sur $]-1,1[$. On remarque que le polynôme P_n est de degrès n et possède la même parité que n. L'orthogonalité s'exprime par

$$\int_{-1}^{1} P_l(x)\; P_m(x) dx = h_l \delta_{l,m}$$

avec $h_n = \frac{2}{2n+1}$ le carré de la norme du polynôme P_n. L'orthogonalité des premiers polynomes est illustrée sur la figure 3.2. Les différents panels représentent deux polynômes de Legendre ainsi que la surface de recouvrement des deux polynômes en grisée. La mesure de cette surface donne la valeur numérique du produit scalaire. L'entrelacement des zéros des polynômes est illustré sur la figure 3.3.

Une fonction $f\,(x)$ définie sur l'intervalle $]a,b[$ est décomposable en série de polynômes de Legendre en deux étape :

— Ramener la fonction sur l'intervalle $]-1,1[$ en posant $g\,(t) = f\left(a + \frac{1}{2}\,(b-a)\,(t+1)\right)$,

— Calculer les projections de g sur les P_n

$$C_n = \frac{\langle g, P_n\rangle}{\langle P_n, P_n\rangle} = \frac{1}{h_n}\int_{-1}^{1} g(x)\; P_n\,(x)\; dx.$$

On obtient alors les décompositions

$$\begin{array}{rcl} g\,(t) & = & \displaystyle\sum_{n=0}^{\infty} C_n\; P_n\,(t)\,; \\ f\,(x) & = & \displaystyle\sum_{n=0}^{\infty} C_n\; P_n\left(\frac{2x-a-b}{b-a}\right). \end{array}$$

La figure 3.4 représente deux exemples de décomposition en utilisant les 11 premiers polynômes.

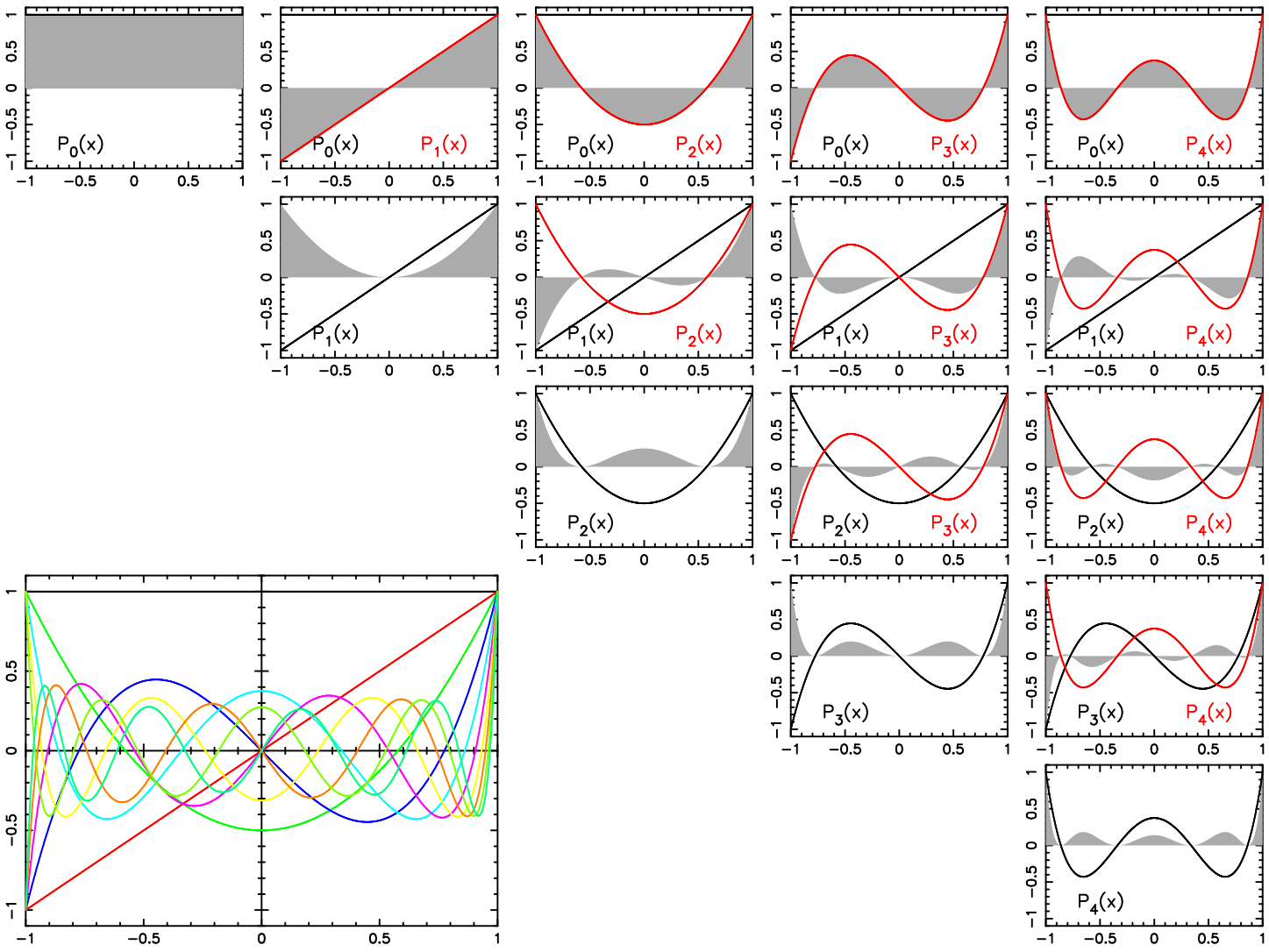


FIGURE 3.2 – Orthogonalité des Polynomes de Legendre.

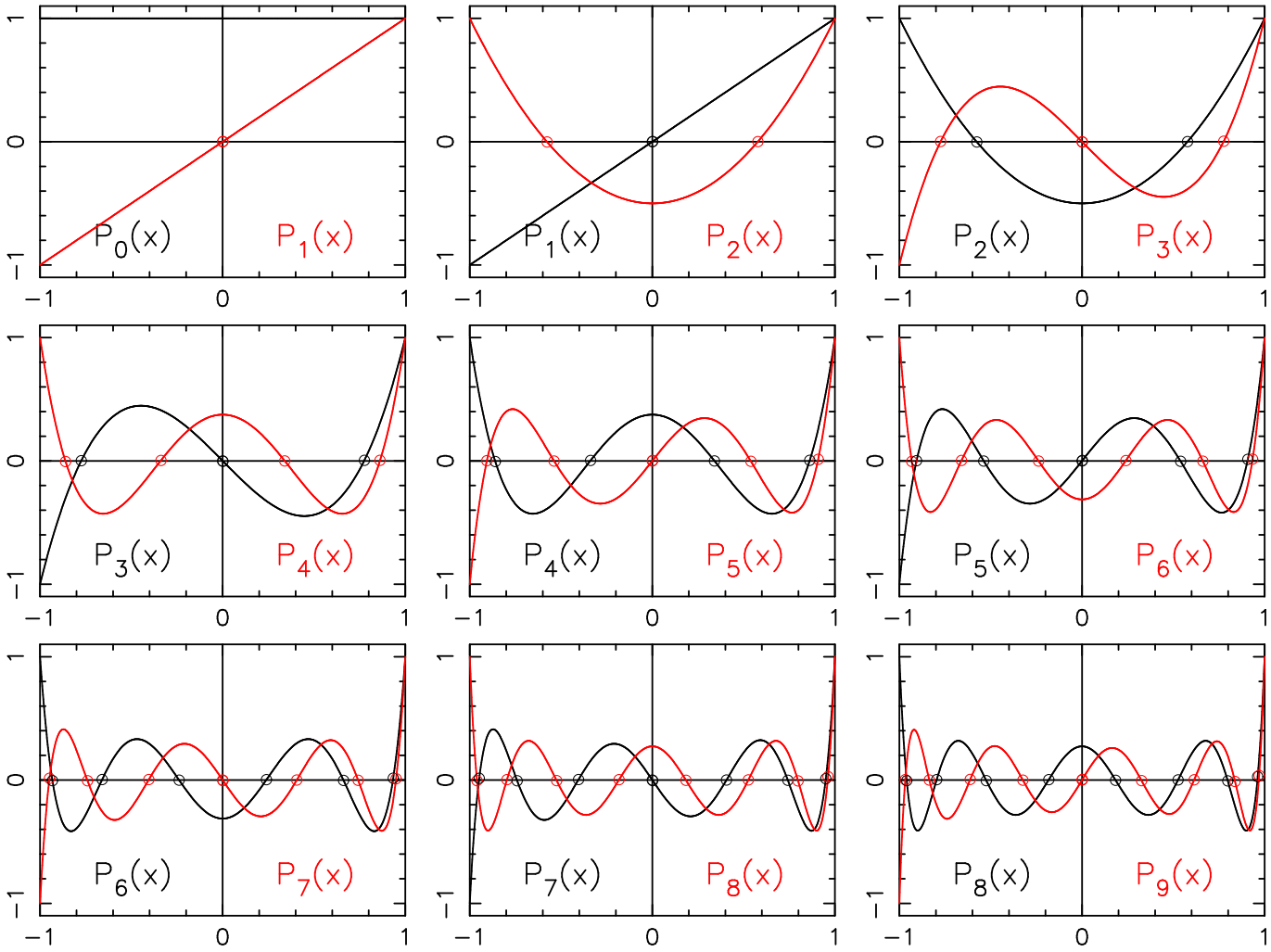


FIGURE 3.3 – Entrelacement des zéros des polynome de Legendre.

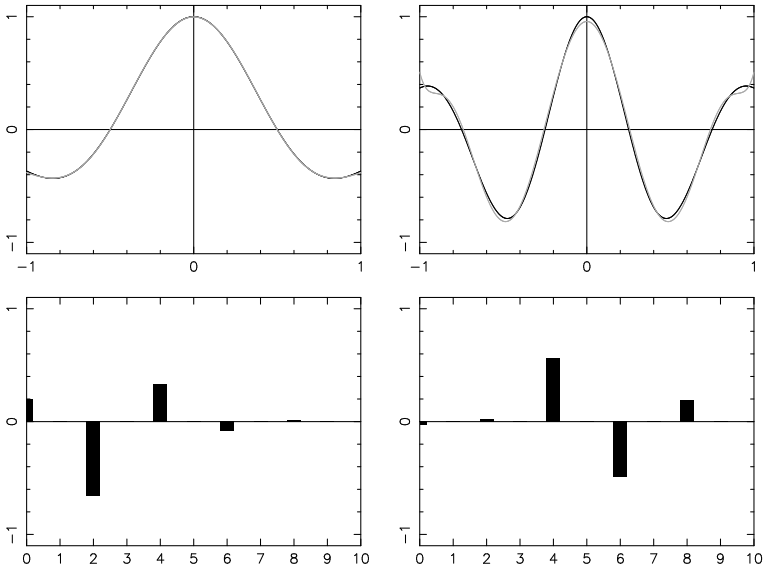


FIGURE 3.4 – Décomposition de fonctions

3.3.3 Exemple de la DFT

La série de Fourier tronquée, ou le polynôme trigonométrique avec les coefficients de Fourier, donne la meilleure approximation, au sens des moindres carrée, d'une fonction f sur un intervalle $[a, b]$, que l'on peut ramener sur $[0, 2\pi]$ par une transformation affine.

$$f(x =\simeq \sum_{n=-M}^{M} c_n\, e^{inx} \quad c_n = \frac{1}{2\pi}\int_0^{2\pi} f(x)\, e^{-inx}\, dx.$$

Cependant, les coefficients de Fourier c_n sont des intégrales qui doivent être approchés dans un calcul numérique. Cela suggère de dérivater une frome discrète de la série de Fourier tronquée pour une fonction f définie sur une grille régulière : $f_j = f(x_j)$, avec $j = 0, 1, 2, \ldots, N$ et $x_j = \frac{2\pi}{N}$. La périodicité de la fonction discrète est spécifiée par $f_0 = f_N$. La fonction étant définie par N valeurs, le nombre de coefficients de Fourier doit aussi être N. Par convenance, le nombre de point de grille est habituellement choisi comme un entier pair. La fonction discrétisée est alors représenté par une série trigonométrique tronquée :

$$f(x_j) = f_j = \sum_{n=-N/2}^{N/2-1} c_n\, e^{inx_j}.$$

Pour calculer les coeficients de Fourier on utilise un produit scalaire discret

$$\langle f, g\rangle_D = \sum_{j=0}^{N-1} f(x_j)\cdot g(x_j)^\dagger, \quad x_j = \frac{2\pi}{N},$$

tel que les fonction de base de la série de Fourier $\varphi_n(x) = e^{inx}$ soient orthogonales :

$$\langle \varphi_n,\, \varphi_m\rangle_D = \begin{cases} N & \text{si } \frac{n-m}{N} \text{ est un entier} \\ 0 & \text{sinon} \end{cases}.$$

On définit ainsi la série de Fourier discrète (*Discrete Fourier Transform (DFT)*) par

$$f(x_j) = \sum_{n=-N/2}^{N/2-1} c_n\, e^{inx_j}; \quad c_n = \frac{1}{N}\sum_{j=0}^{N-1} f(x_j)\, e^{-inx_j}.$$

3.4 Problèmes

3.4.1 Applications directes

1. Soit f la fonction de période T, définie sur l'intervalle $[0, T]$ par :

$$f(t) = \begin{cases} 1 & \text{si } 0 \le t < T/2 \\ 0 & \text{si } T/2 \le t < T. \end{cases}$$

   — Représenter f graphiquement.
   — Calculer les coefficients de Fourier a_n, b_n et c_n de f.
   — Représenter graphiquement le spectre de f.
   — Écrire la série de Fourier de f.

2. Répondre aux mêmes question si

$$f(t) = \begin{cases} t & \text{si } 0 \le t < T/2 \\ T - t & \text{si } T/2 \le t < T. \end{cases}$$

3. Répondre aux mêmes question si

$$f(t) = \sin\left(\frac{2\pi}{T}t\right)\cos\left(3\frac{2\pi}{T}t\right)$$

4. Répondre aux mêmes question si

$$f(t) = t$$

et en déduire les résultats pour

$$f(t) = t(T-t)$$

3.4.2 Calcul des coefficients de Fourier par décomposition

1. Soient g et h deux fonctions T–périodiques, α, β et γ trois réels. On note g_n et h_n les coefficients de Fourier de g et h.
   — Montrer que la fonction $f(t) = \alpha g(t) + \beta h(t) + \gamma$ est T–périodique ;
   — Montrer que les coefficients de Fourier f_n de f vérifient

$$\forall n \qquad f_n = \alpha g_n + \beta h_n + \gamma$$

2. Soit f une fonctions T–périodiques dont les coefficients de Fourier $\{f_n\}$ sont connus.
   — Montrer que la fonction $g(t) = \int_0^t f(u)du$ est T–périodique ;
   — Montrer que les coefficients de Fourier g_n de g vérifient

$$\forall n \qquad g_n = \frac{g(T) - g(0)}{-in2\pi} + \frac{T}{in2\pi} f_n$$

3. On note $f_{a,b}(t)$ la fonction en escalier T–périodique défini par deux paramètres réels a et b tels que $0 \leq a \leq 1$ et $a \leq b \leq 1$

$$f_{a,b}(t) = \begin{cases} 1 & \text{si} aT \leq t \leq aT \\ 0 & \text{sinon} \end{cases}$$

   Déteminer les coefficients de Fourier de $f_{a,b}$

4. Utiliser les trois résultats précédent pour retrouver les coefficients de Fourier des fonctions T–périoque

$$f(t) = t$$

$$f(t) = t(T-t)$$

$$f(t) = \begin{cases} t & \text{si } 0 \leq t < T/2 \\ T-t & \text{si } T/2 \leq t < T. \end{cases}$$

$$f(t) = \begin{cases} 2 & \text{si } 0 \leq t < T/2 \\ -3 & \text{si } T/2 \leq t < T. \end{cases}$$

$$f(t) = \begin{cases} 0 & \text{si } 0 \leq t < T/4 \\ t - T/4 & \text{si } T/4 \leq t < 3T/4 \\ 1 & \text{si } 3T/4 \leq t < T. \end{cases}$$

$$f(t) = \begin{cases} 0 & \text{si } 0 \leq t < T/4 \\ t - T/4 & \text{si } T/4 \leq t < 3T/4 \\ 0 & \text{si } 3T/4 \leq t < T. \end{cases}$$

3.4.3 Autour des séries de Fourier (Novembre 2012)

Jeux de coefficients

1. À partir des définitions données dans l'annexe, déterminer la relation entre les coefficients de Fourier a_n et b_n et les coefficients complexes c_n.
2. On considère le développement en série de Fourier d'une fonction impaire :
   — Que valent les coefficients de Fourier a_n ? Justifier votre réponse.
   — Que peut-on en déduire sur le caractère réel ou imaginaire des coefficients c_n ?
3. Soit f une fonction réelle, de période L, dont on connaît les coefficients de Fourier c_n :

$$c_n = \frac{1}{L}\int_0^L f(x)\exp\left(-in\frac{2\pi}{L}x\right)\,dx.$$

   On définit la primitive F de f par la relation $F(x) = \int_0^x f(u)\,du$ ce qui implique bien sûr que $F'(x) = f(x)$ et $F(0) = 0$. Les coefficients de Fourier C_n de F sont définis par la relation intégrale

$$C_n = \frac{1}{L}\int_0^L F(x)\exp\left(-in\frac{2\pi}{L}x\right)\,dx.$$

   — Déterminer la relation entre c_0 et $F(L)$.
   — En utilisant une intégration par partie, démontrer la relation suivante entre les coefficients de Fourier de f et les coefficients de Fourier de F

$$C_n = \frac{L}{in\ 2\pi}\,(c_n - c_0) \quad \text{si} \quad n \neq 0.$$

Application simple

On considère la fonction f de période L définie par morceaux sur l'intervalle $[0, L]$ par

$$f(x) = \begin{cases} -1 & \text{si} \quad 0 \leq x < L/2 \\ 1 & \text{si} \quad L/2 \leq x < L. \end{cases} \tag{3.1}$$

1. Faire une représentation graphique de la fonction f sur l'intervalle $[-L, 2L]$.
2. La fonction est-elle continue ? à variation bornée ?
3. Quelle loi de décroissance, en fonction de $|n|$, peut–on prévoir pour $|c_n$?
4. Calculer les coefficients de Fourier c_n de f, en explicitant les grandes étapes de calcul (avec des phrases).
5. Simplifier le résultat, si vous le pouvez, pour le mettre sous forme élégante.
6. Représenter graphiquement le spectre de f et commenter l'absence de certaines harmoniques.

Combinatoire

On considère la fonction réelle $f_{a,b}$ de période L, dépendant de deux paramètres réels a et b compris dans l'intervalle $[0, 1]$ avec $b \geq a$. La fonction $f_{a,b}$ est définie par morceaux sur $[0, L]$ par :

$$f(x) = \begin{cases} 0 & \text{si} \quad 0 \leq x < aL \\ 1 & \text{si} \quad aL \leq x < bL \\ 0 & \text{si} \quad bL \leq x < L. \end{cases} \tag{3.2}$$

La figure suivante donne quelques représentations graphiques de $f_{a,b}$ sur l'intervalle $[0, L]$.

1. Calculer les coefficients de Fourier de $f_{a,b}$.

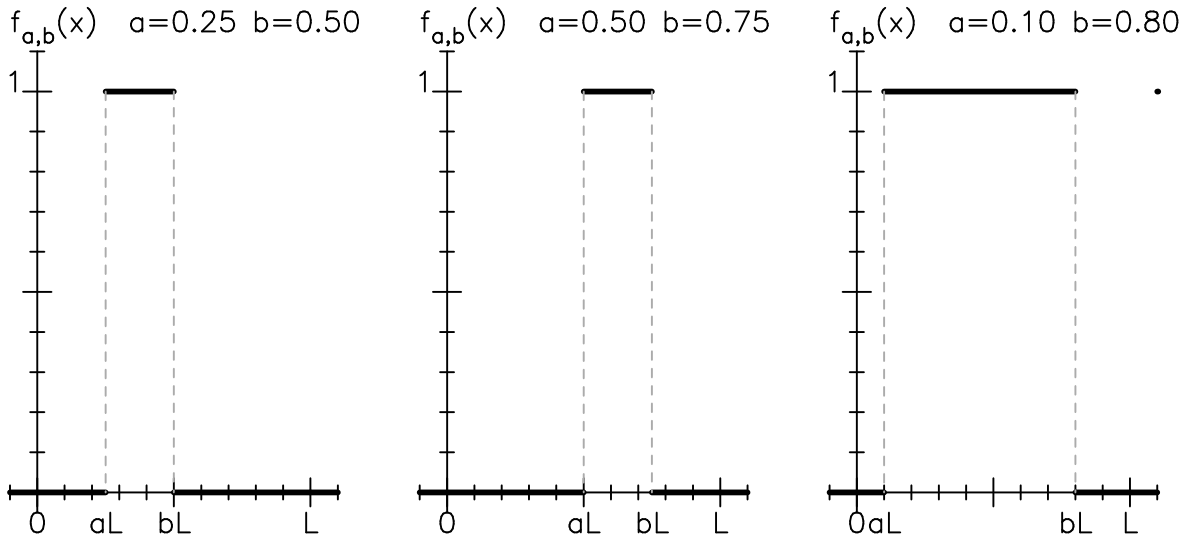


FIGURE 3.5 – Représentation graphique de $f_{a,b}$

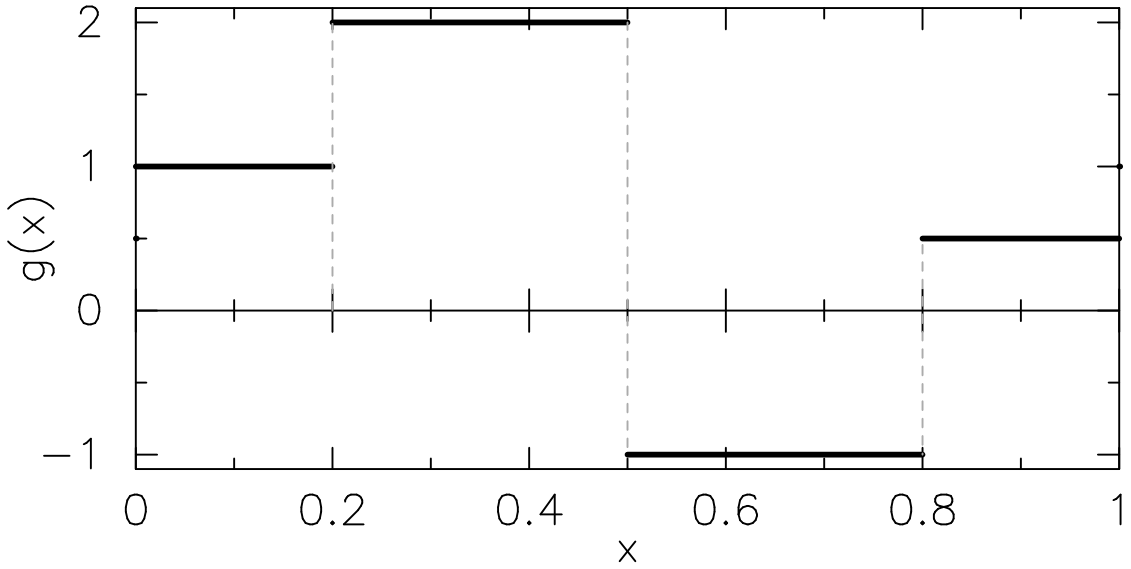


FIGURE 3.6 – Fonction en escalier de période 1

2. Montrer que la fonction f de l'exercice B-2 peut être exprimée comme une combinaison linéaire des fonctions $f_{0,1/2}$ et $f_{1/2,1}$. Utiliser cette décomposition pour retrouver les coefficients de Fourier de f en fonction des expressions des coefficients de Fourier de $f_{0,1/2}$ et $f_{1/2,1}$.
3. On considère la fonction g en escalier de période 1 représentée sur la figure 3.6. Exprimer g comme une combinaison linéaire de fonction $f_{a,b}$ et déterminer ses coefficients de Fourier.

On note $g_{a,b}$ la primitive de $f_{a,b}$ définie sur $[0, L]$ par $g_{a,b}(x) = \int_0^x f_{a,b}(u)\, du$.

1. Faire une représentation graphique de $g_{0,1/2}$, $g_{1/3,2/3}$ sur l'intervalle $[0, L]$.
2. Calculer les coefficients de Fourier de $g_{a,b}$ en utilisant, entre autres, les relations de la question B–1 3.
3. On définit la fonction h de période L par la combinaison linéaire

$$h(x) = g_{1/5,2/5}(x) - g_{3/5,4/5}(x).$$

— Faire une représentation graphique de la fonction $h(x)$ et de sa dérivée $h'(x)$ sur une période.
— Déterminer les coefficients de Fourier de h en utilisant les coefficients de Fourier des fonctions $g_{a,b}$.
— Les fonctions h et h' sont-elles continue ? à variation bornée ?
— La décroissance des modules de coefficients de Fourier de h est-elle conforme aux observations de la question précédente ?

3.4.4 Solution en série d'une équation différentielle (Février 2011)

On utilisera toujours la représentation complexe des séries de Fourier.

1. Montrer que la fonction $f(t) = |\sin(2t)|$ est périodique et déterminer sa période.
2. Représenter graphiquement la fonction $f(t)$.
3. Calculer les coefficients de Fourier de la fonction $f(t)$.

On va utiliser les résultats précédents pour déterminer une solution périodique de l'équation différentielle

$$\frac{d^4y(t)}{dt^4} + 5\frac{d^2y(t)}{dt^2} + 4y(t) = |\sin(2t)|.$$

1. Écrire la variable $y(t)$ sous la forme d'une série de Fourier de coefficients inconnus et de période $\frac{\pi}{2}$.
2. Utiliser le résultat précédent pour écrire l'expression $\frac{d^4y(t)}{dt^4} + 5\frac{d^2y(t)}{dt^2} + 4y(t)$ sous la forme d'une série de Fourier de période $\frac{\pi}{2}$.
3. En identifiant les coefficients termes à termes, déterminer les coefficients de Fourier de la solution de l'équation différentielle

3.4.5 Systèmes orthogonaux (Novembre 2012)

1. Quelle interprétation graphique peut-on donner au produit scalaire de deux fonctions.
2. Expliquer le plus clairement possible la notion de série de Fourier généralisé en précisant la méthode de calcul des coefficients, vous appuierez vos explications sur une analogie avec l'algèbre des vecteurs dans $\mathbb{R}^2$.
3. On considère une famille de fonctions définies sur $[0, \pi]$ par

$$\varphi_k(x) = \sin(kx), \quad k \in \mathbb{N}, k > 0.$$

— Représenter graphiquement φ_1, φ_2 et φ_5.
— Montrer que le système $\{\varphi_k(x)\}_{k\geq 1}$ est orthogonal sur $[0, \pi]$ pour le produit scalaire

$$\langle u, v\rangle = A\int_0^\pi u(x)\, v(x)\, dx;$$

où A est une constante réelle positive non encore définie.
— Calculer en fonction de A la norme des fonctions $\varphi_k(x)$.
— Quelle valeur de A faut-il choisir pour que la famille de fonctions soit orthonormale sur $[0, \pi]$?
— Calculer les coefficients de la série de Fourier généralisée associée à la fonction définie sur $[0, \pi]$ par $f(x) = x$.

Chapitre 4

Transformée de Fourier

Sommaire

La transformation de Fourier est un analogue de la théorie des séries de Fourier pour les fonctions non périodiques, et permet de leur associer un spectre en fréquences. On cherche ensuite à obtenir l'expression de la fonction comme « somme infinie » des fonctions trigonométriques de toutes fréquences qui forment son spectre. Une telle sommation se présentera donc sous forme d'intégrale. Série et transformation de Fourier constituent les deux outils de base de l'analyse harmonique.

4.1 Définition de la transformée de Fourier

Pour une fonction T–**périodique** f, on obtient une relation de la forme

$$f(t) = \sum_{c=-\infty}^{\infty} c_n \, e^{in\frac{2\pi}{T}t}$$

qui peut être interprétée comme la décomposition de la fonction f sur la base des fonctions T–périodiques $e^{in\frac{2\pi}{T}t}$. On remarquera que $\frac{n}{T}$ a une dimension de fréquence. Lorsque n d'écrit l'ensemble des entiers relatifs, $\frac{n}{T}$ décrit un ensemble de fréquences qui dépend de T.

Pour une fonction f **qui n'est pas périodique**, il est exclu d'utiliser une telle décomposition. On peut cependant considérer qu'un fonction qui n'est pas périodique est une fonction dont la période est infinie. Or si T est grand, l'ensemble des fréquences $\frac{n}{T}$ est un ensemble qui couvre presque toutes les fréquences possibles. On est passé d'une succession de fréquences à un ensemble continu de fréquences ; aussi quand il s'agit de faire la somme, il faut passer d'une somme discrète, au sens des séries, à une somme continue, c'est-à-dire au sens du calcul intégral.

Plus riguresement, on suppose $f(t)$ une fonction non périodique *i.e.* de période $T \to \infty$ définie par sa série de Fourier

$$f(t) = \sum_{n=-\infty}^{\infty} c_n \, e^{in\omega t} = \frac{1}{T} \sum_{n=-\infty}^{\infty} \int_{-T/2}^{T/2} f(\tau) e^{in\omega(t-\tau)} \, \mathrm{d}\tau,$$

avec $\omega = 2\pi/T$ la pulsation. En posant $\omega_n = n\,\omega = \frac{n2\pi}{T}$ et $\Delta\omega = \omega_{n+1} - \omega_n = \omega$ on écrit

$$\lim_{T\to\infty} f(t) = \lim_{\Delta\omega\to 0} \frac{\Delta\omega}{2\pi} \sum_{n=-\infty}^{\infty} \int_{-T/2}^{T/2} f(\tau) e^{in\omega(t-\tau)} \, \mathrm{d}\tau = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\tau) \, e^{-i\omega\tau} \mathrm{d}\tau \right] \, e^{i\omega t} \mathrm{d}\omega.$$

La variable continue ω se substitut aux variables discrètes n/T.

Pour définir plus rigoureusement la transformée de Fourier , il faut introduire l'espace $\mathcal{L}^1(\mathbb{R})$:

> **Espace des fonctions sommables :** *On note $\mathcal{L}^1(\mathbb{R})$ l'ensemble des fonctions f définies de $\mathbb{R}$ dans $\mathbb{R}$, continues par morceaux et telles que*
>
> $$\int_{-\infty}^{\infty} | \, f(t) \, | \, dt \quad \textit{existe.}$$

4.1.1 Définition

> **Définition :** *Soit $f = f(u) \in \mathcal{L}^1(\mathbb{R})$, on appelle transformée de Fourier de f, la fonction $\mathrm{TF}[f]$ de la variable s telle que*
>
> $$\mathrm{TF}[f(u)](s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u) e^{-\mathrm{i}su} \, \mathrm{d}u.$$
>
> *On note $\mathrm{TF}[f(u)](s) = \tilde{f}(s)$ lorsqu'il n'y a pas d'ambiguïté. La courbe $y(s) =\mid \tilde{f}(s) \mid$ est appelée spectre de f.*

Remarque : La définition de la transformée de Fourier peut différer d'une constante multiplicative dans certains ouvrages.

Notations en physique :

— Si $f(t)$ est une fonction du temps, il est usuel d'utiliser ω (pulsation) comme variable de la transformée de Fourier plutôt que s.

$$\mathrm{TF}[f(t)](\omega) = \tilde{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-\mathrm{i}\omega t} \, \mathrm{d}t.$$

— Si $f(x)$ est fonction d'une variable d'espace x, il est usuel d'utiliser k (nombre d'onde) comme variable de la transformée de Fourier :

$$\mathrm{TF}[f(x)](k) = \tilde{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-\mathrm{i}kx} \, \mathrm{d}x.$$

4.1.2 Transformée inverse

> **Théorème d'inversion :** *Soit $f(t) \in \mathcal{L}^1(\mathbb{R})$, et soit $\tilde{f}$ sa transformée de Fourier , Si $\tilde{f} \in \mathcal{L}^1(\mathbb{R})$ alors on a presque partout*
>
> $$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \tilde{f}(\omega) e^{+\mathrm{i}\,\omega t} \, \mathrm{d}\omega.$$
>
> *De même, pour une fonction spatiale $f(x)$*
>
> $$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \tilde{f}(k) e^{+\mathrm{i}\,kx} \, \mathrm{d}k.$$

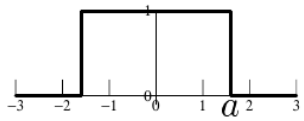
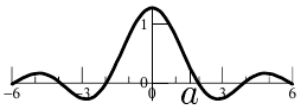
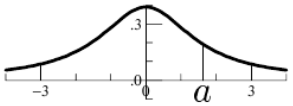
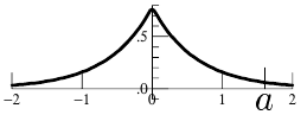
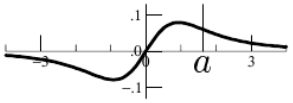
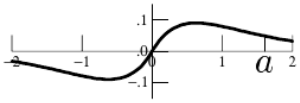
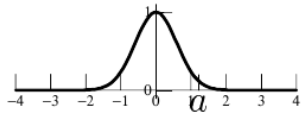
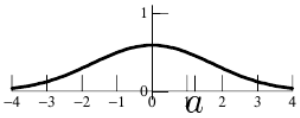
$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widehat{f}(\omega) \cdot e^{i\omega x}\, d\omega$	$\widehat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) \cdot e^{-i\omega t}\, dt$
 $\begin{cases} 1 & \text{si } \lvert x\rvert < a, \\ 0 & \text{si } \lvert x\rvert \geq a; \end{cases}$	$\sqrt{\frac{2}{\pi}}\, \frac{\sin a\omega}{\omega}$ 
 $\frac{1}{x^2 + a^2} \quad (a > 0)$	$\sqrt{\frac{\pi}{2}}\, \frac{e^{-a\lvert\omega\rvert}}{a}$ 
 $\frac{x}{(x^2 + a^2)^2} \quad (a > 0)$	$-\sqrt{\frac{\pi}{8}}\, \frac{i\omega e^{-a\lvert\omega\rvert}}{a}$ 
 $e^{-a^2x^2} \quad (a > 0)$	$\frac{1}{a\sqrt{2}}\, e^{\frac{-\omega^2}{4a^2}}$ 

TABLE 4.1 – une petite table de transformée de Fourier

4.2 Propriétés de la transformée de Fourier

4.2.1 Propriétés analytiques

Soit $f \in \mathcal{L}^1(\mathbb{R})$, alors

- $\tilde{f}$ est continue,
- $\tilde{f}$ est bornée,
- $\lim_{|s|\to\infty} \mid \tilde{f}(s) \mid = 0$.

4.2.2 Linéarité

La transformée de Fourier est une application linéaire de $\mathcal{L}^1(\mathbb{R})$ dans l'espace des fonctions. Soit f et g deux fonctions de $\mathcal{L}^1(\mathbb{R})$ et a et b deux nombres complexes

$$\mathcal{F}(a\, f + b\, g) = a\, \mathcal{F}(f) + b\, \mathcal{F}(g)$$

4.2.3 Parité et réalité

$f(t)$	$\tilde{f}(\omega)$
paire	paire
impaire	impaire
réelle	$\tilde{f}(-\omega) = \tilde{f}^*(\omega)$
imaginaire pure	$\tilde{f}(-\omega) = -\tilde{f}^*(\omega)$

Exercice : Démontrer les résultats précédents.

4.2.4 Échelle, Translation & Modulation

Échelle : $f(t) \in \mathcal{L}^1(\mathbb{R})$ *et* $a \in \mathbb{R}^*$

$$\mathrm{TF}\left[f(a\,t)\right](\omega) = \frac{1}{\mid a \mid} \mathrm{TF}\left[f(t)\right]\left(\frac{\omega}{a}\right) = \frac{1}{\mid a \mid} \tilde{f}\left(\frac{\omega}{a}\right).$$

L'effet d'une dilatation dans l'espace réel est une contraction dans l'espace de Fourier et vice–versa

Translation : $f(t) \in \mathcal{L}^1(\mathbb{R})$ *et* $\tau \in \mathbb{R}^*$

$$\mathrm{TF}\left[f(t+\tau)\right](\omega) = e^{\mathrm{i}\,\omega\,\tau}\ \mathrm{TF}\left[f\right](\omega) = e^{\mathrm{i}\,\omega\,\tau}\ \tilde{f}(\omega).$$

L'effet d'une translation dans l'espace réel est un déphasage dans l'espace de Fourier.

Modulation : $f(x) \in \mathcal{L}^1(\mathbb{R})$ *et* k_0 *réel*

$$\mathrm{TF}\left[f(x)e^{\mathrm{i}k_0 x}\right](k) = \tilde{f}(k-k_0).$$

L'effet d'une modulation dans l'espace réel est une translation dans l'espace de Fourier.

Exercice : Démontrer les résultats précédents.

4.2.5 Transformée de Fourier et dérivation

Théorème : *Soit* $f(x) \in \mathcal{L}^1(\mathbb{R})$ *continue et dérivable presque partout telle que* $f' \in \mathcal{L}^1(\mathbb{R})$, *alors*

$$\mathrm{TF}\left[\frac{\mathrm{d}f}{\mathrm{d}x}(x)\right] = \mathrm{i}\ k\ \tilde{f}(k).$$

Théorème : *Soit* $f(t)$, m *fois dérivables, telle que ses* m *premières dérivées soient continues et appartiennent toutes à* $\mathcal{L}^1(\mathbb{R})$, *alors*

$$\mathrm{TF}\left[f^{(m)}\right](\omega) = (\mathrm{i}\,\omega)^m\ \tilde{f}(\omega).$$

De ce dernier théorème, on peut déduire un résultat concernant la vitesse de décroissance vers zéro de la transformée de Fourier de f en fonction de l'ordre maximale de dérivation. En effet, si f est telle que $f^{(m)}$ soit sommable et définie presque partout, alors sa transformée de Fourier est bornée :

$$\exists M > 0, \forall\ s \in \mathbb{C} \mid (i\ s)^m\ \mathcal{F}(f)(s) \mid \leq M.$$

Donc si $k \neq 0$

$$\mid \mathcal{F}(f)(s) \mid \leq \frac{M}{\mid s \mid^m}.$$

Plus la fonction f est dérivable, plus sa transformée de Fourier décroît rapidement, pour une fonction infiniment dérivable, la transformée de Fourier décroît plus vite que n'importe qu'elle puissance de $\mid s \mid^{-m}$.

Théorème : *Soit* $f(t) \in \mathcal{L}^1(\mathbb{R})$ *telle que la fonction définie par :* $t \to t\ f(t)$ *soit sommable. Alors* $\mathrm{TF}\left[tf(t)\right](\omega)$ *est dérivable et on a*

$$\mathrm{TF}\left[\,t\,f(\,t\,)\,\right](\omega) = \mathrm{i}\,\frac{\mathrm{d}}{\mathrm{d}\omega}\mathrm{TF}\left[f(t)\right](\omega) = \mathrm{i}\,\frac{\mathrm{d}}{\mathrm{d}\omega}\tilde{f}(\omega)$$

Théorème (généralisation) : *Soit* $f(t) \in \mathcal{L}^1(\mathbb{R})$ *telle que la fonction définie par :* $t \to t^m\ f(t)$ *soit sommable. Alors* $\mathrm{TF}\left[tf(t)\right](\omega)$ *est dérivable et on a*

$$\mathrm{TF}\left[\,t^m\,f(\,t\,)\,\right](\omega) = \mathrm{i}^m\,\frac{\mathrm{d}^m}{\mathrm{d}\omega^m}\mathrm{TF}\left[f(t)\right](\omega) = \mathrm{i}^m\,\frac{\mathrm{d}^m}{\mathrm{d}\omega^m}\tilde{f}(\omega)$$

4.3 Théorèmes généraux

Théorème de Plancherel : *Soient $f(x)$ et $g(x)$ deux fonctions de $\mathcal{L}^1(\mathbb{R})$, et soient $\tilde{f}$ et $\tilde{g}$ leurs transformée de Fourier respectives, alors*

$$\int_{-\infty}^{\infty} f(x)g^*(x)\mathrm{d}x = \int_{-\infty}^{\infty} \tilde{f}(k)\tilde{g}^*(k)\mathrm{d}k.$$

Théorème de Parseval : *Soit $f \in \mathcal{L}^1(\mathbb{R})$, et soit $\tilde{f}$ sa transformée de Fourier , alors*

$$\int_{-\infty}^{\infty} \mid f(x) \mid^2 \mathrm{d}x = \int_{-\infty}^{\infty} \mid \tilde{f}(k) \mid^2 \mathrm{d}k.$$

La transformée de Fourier se généralise sans difficulté aux fonctions de plusieurs variables.

Exercice : Écrire les équations de Maxwell dans le vide dans l'espace de Fourier.

4.4 Produit de convolution

Définition : *Soient f et g deux fonctions de $\mathcal{L}^1(\mathbb{R})$, alors l'intégrale $\int f(x-x')g(x')dx'$ existe pour presque tout x et définit une fonction de $\mathcal{L}^1(\mathbb{R})$ appelée produit de convolution de f et de g et notée $f * g$*

$$f * g(x) = \int f(x-x')g(x')dx'.$$

Le produit de convolution est commutatif $f * g = g * f$.

Le produit de convolution prends une forme bien plus simple dans l'espace de Fourier !

Théorème : *Soient f et g deux fonctions sommables sur $\mathbb{R}$, alors*

$$\mathrm{TF}\left[f * g\right] = \sqrt{2\pi}\mathrm{TF}\left[f\right]\ \mathrm{TF}\left[g\right] = \sqrt{2\pi}\,\tilde{f}\,\tilde{g}$$

4.5 Distribution de Dirac

- D'une manière abusive, mais représentative, on peut définir la fonction de Dirac $\delta(x)$ par

$$\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases}$$

- On peut aussi se représenter la fonction de Dirac $\delta(x)$ par une fonction "porte" d'aire unité :

$$\Delta_a(x) = \begin{cases} \frac{1}{a} & \text{si } \mid x \mid \leq \frac{a}{2}, \\ 0 & \text{sinon.} \end{cases}$$

  dans la limite $a \to 0$: $\delta(x) = \lim_{a\to 0} \Delta_a(x)$.
- D'une manière plus rigoureuse, on définit la distribution de Dirac par

$$\int_{-\infty}^{\infty} f(x)\delta(x-a)\,dx = f(a)$$

  pour toute fonction f définie sur $\mathbb{R}$.

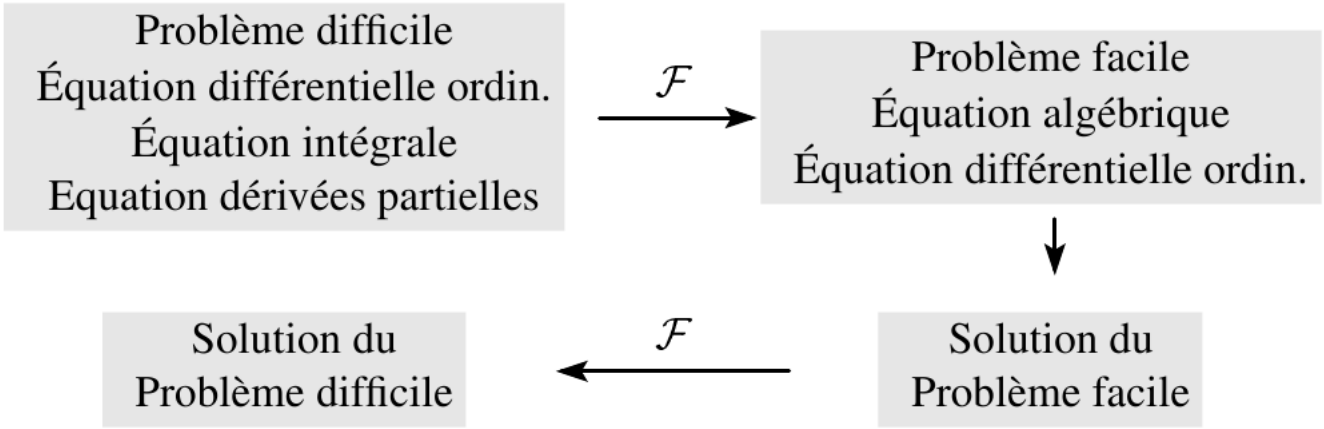
Il en découle directement que

$$\int_{-\infty}^{\infty} \delta(x)\,dx = 1.$$

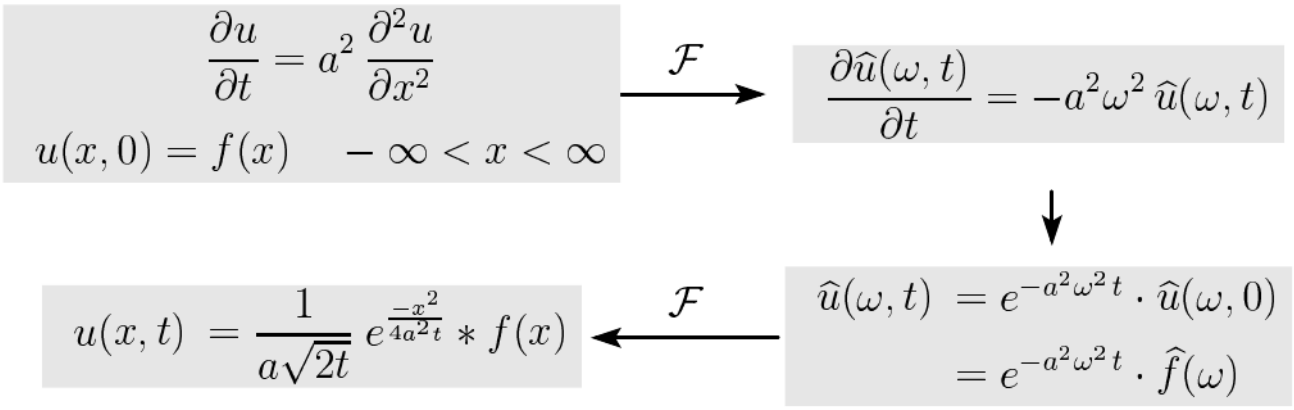
$$\mathrm{TF}\left[\delta(x)\right](k) = \tilde{\delta}(k) = \frac{1}{\sqrt{2\pi}}$$

4.6 Application aux équations différentielles

Le schéma est le suivant : la transformation de Fourier change une équation différentielle en une équation plus simple. Cette équation est résolue, puis la transformée inverse donne la solution.



Voyons avec quelle élégance la transformée de Fourier amène à la solution de l'équation de la chaleur pour un fil infini :



Sous des conditions sur $f(x)$ qui permettent de faire toutes ces opérations, la solution est alors donnée par la formule de Poisson

$$u\,(x,t) = \frac{1}{2a\sqrt{2\pi}} \int_{-\infty}^{\infty} f(s) \exp\left(-\frac{(x-s)^2}{4a^2t}\right) ds$$

4.7 Problèmes

De la série à la transformée

Soit la fonction $f(x)$ de période L définie sur l'intervalle $[-L/2, L/2]$ par :

$$f(x) = \begin{cases} 1 & \text{si } \mid x \mid \leq a \\ 0 & \text{sinon,} \end{cases}$$

avec a un paramètre réel positif inférieur à $L/2$

— Déterminer la série de Fourier de f.

— Discuter de la limite $L \to \infty$.

Applications directes

1. Soit la fonction gausienne de paramètre réel a : $f(x) = \exp(-x^2/a^2)$
   — Déterminer la transformée de Fourier de f.
   — Représenter le spectre de f.
   — Discuter de la largeur du spectre dans la limite $a \to 0$ et $a \to \infty$
2. Déterminer la transformée de Fourier de $f(t) = \exp(-a \mid t \mid), \quad a > 0.$
3. Déterminer la transformée de Fourier de $f(t) = \begin{cases} \exp(-at) & \text{si } t \geq 0, \\ 0 & \text{sinon.} \end{cases}$
4. Déterminer la transformée de Fourier $(a > 0)$ de $f(t) = \begin{cases} \exp(-at)\, e^{i\omega_0 t} & \text{si } t \geq 0, \\ 0 & \text{sinon.} \end{cases}$
5. Déterminer la transformée de Fourier de $f(t) = \exp(-t^2/a^2)\cos(\omega_0 t)$.
6. Déterminer la transformée de Fourier de $f(t) = \begin{cases} \cos(\omega_0 t) & \text{si } \mid t \mid \leq \frac{1}{2}, \\ 0 & \text{sinon.} \end{cases}$

Exercice : On considère la fonction définie par

$$f = \begin{cases} 1 & \text{si} \quad x \in]-a, a[\\ 0 & \text{sinon} \end{cases} \qquad a > 0.$$

1. Montrer que $f \in \mathcal{L}^1(\mathbb{R})$.
2. Calculer $\mathcal{F}(f)(s)$.
3. Monter que $\mathcal{F}(f)(s) \notin \mathcal{L}^1(\mathbb{R})$.

Exercice : On considère la fonction définie par

$$f = \begin{cases} \cos(\omega_0 t) & \text{si} \quad t \in]-\tau, \tau[\\ 0 & \text{sinon} \end{cases} \qquad a > 0.$$

1. Montrer que $f \in \mathcal{L}^1(\mathbb{R})$.
2. Calculer $\mathcal{F}(f)(s)$.

Exercice : On considère la fonction définie par (a>0)

$$f = \begin{cases} \exp(-x/a) & \text{si} \quad x \in [0, \infty[\\ 0 & \text{sinon} \end{cases} \qquad a > 0.$$

1. Montrer que $f \in \mathcal{L}^1(\mathbb{R})$.
2. Calculer $\mathcal{F}(f)(s)$.

Principe d'incertitude

Supposons une fonction en créneau $\Psi(t)$ de largeur Δt et d'aire unité centré autour de t_0.

1. Donner la transformée de Fourrier. On précisera le résultat en fonction de Δt et t_0.
2. Montrer qu'il existe un principe d'incertitude temps–fréquence.
3. Commenter le cas $\Delta t \to 0$.

Fonction de Dirac

Une représentation de la fonction de Dirac $\delta(x)$ est obtenue en prenant la fonction créneau

$$\Delta_a(x) = \begin{cases} \frac{1}{a} & \text{si } \mid t \mid \leq \frac{a}{2}, \\ 0 & \text{sinon.} \end{cases}$$

dans la limite $a \to 0$: $\delta(x) = \lim_{a\to 0} \Delta_a(x)$. De même, la transformée de Fourier $\delta(k) = \mathrm{TF}\{\delta(x)\}$ est donnée par $\delta(k) = \lim_{a\to 0} \mathrm{TF}\{\Delta_a(x)\}$.

1. En déduire la transformée de Forrier de $\delta(x)$.
2. Montrer en faisant la transformation inverse que l'on obtient :$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ikx}\, dk$.
3. En déduire la transformée de Fourier des fonction $f(x) = e^{-ik_0x}$, $g(x) = \cos(k_0x)$ et $g(x) = \sin(k_0x)$.

Peigne de Dirac

1. Calculer la série de Fourier de la fonction

$$F(X) = \sum_{n=-\infty}^{\infty} \delta(X - n\,a).$$

2. En déduire que la transformée de Fourier d'un peigne de Dirac $g(x) = \sum_{n=-\infty}^{\infty} \delta(x - n\,a)$ est un peigne de Dirac en k de la forme

$$\tilde{g}(k) = \frac{\sqrt{2\pi}}{a} \sum_{m=-\infty}^{\infty} \delta\left(k - m\frac{2\pi}{a}\right).$$

Produit de convolution

Soit la fonction "porte" $P(x) = \begin{cases} 1 & \text{si } \mid t \mid \leq a, \\ 0 & \text{sinon.} \end{cases}$

1. Déterminer le produit de convolution de la fonction "porte" par elle–même : $T(x) = P(x) * P(x)$.
2. En déduire la tranformée de Fourier de la fonction triangle.

Filtre de déreverbération

Voici un exemple d'utilisation de la transformée de Fourier dans le domaine temporel. On se place dans le cadre d'une étude des ondes sismique marine. Un hydrophone enregistre l'activité sismique représenté par un signal acoustique $i(t)$. De manière général, le signal électrique de sortie $s(t)$ de l'hydrophone ne sera pas le reflet exact du signal sismique $i(t)$. En effet les ondes sismiques se réfléchissent sur la surface de l'océan et sur le fond de l'océan. Le signal enregistré, $s(t)$, sera donc le reflet du signal brut $i(t)$ et de ses multiples réflexions. Pour simplifier, on suppose que l'hydrophone est placé à la surface de l'océan, on note T le temps de trajet d'un aller retour de l'onde sonore entre la surface et le fond de l'océan et r le coefficient d'absorption de l'onde lors de l'aller retour. Le signal enregistré $s(t)$ sera de la forme

$$s(t) = i(t) + ri(t - T) + r^2 i(t - 2T) + \ldots$$

1. Pouvez-vous déduire $i(t)$ de $o(t)$?
2. Calculer $S(\omega)$ la transformée de Fourier de $s(t)$ en fonction de $I(\omega)$ la transformée de Fourier de $i(t)$.
3. Quelle procédure pouvez–vous proposer pour extraire $i(t)$ de $s(t)$.
4. Montrez que le résultat dépend sensiblement de l'estimation de r et de T.

Équation de la chaleur

On se propose d'étudier l'équation de la chaleur $\frac{\partial T}{\partial t} = D\frac{\partial^2 T}{\partial x^2}$ dans une barre à une dimension supposée infinie. $T = T(x,t)$ est ici la distribution de température à la date t. On donne la distribution initiale de température $T(x, t=0) = f(x)$.

1. On appelle $T(k,t)$ la transformée de Fourier du profil de température $T(x,t)$. Que devient l'équation de la chaleur pour $T(k,t)$?
2. En déduire que $T(x,t) = \frac{1}{\sqrt{2\pi}} \int \mathrm{d}k e^{-Dk^2 t}\, e^{ikx}\, T(k, t=0)$
3. En déduire la solution formelle

$$T(x,t) = \frac{1}{2\pi} \iint \mathrm{d}k\mathrm{d}u\; e^{-Dk^2 t}\, e^{ik(x-u)}\, f(u)$$

4. Déterminer l'évolution du profil de température lorsque
   — $f(x) = \delta(x)$
   — $f(x) = T_0\, e^{-\frac{x^2}{a^2}}$

Chapitre 5

Transformée de Laplace

Sommaire

En mathématiques, la transformation de Laplace est une transformation intégrale, c'est-à-dire une opération associant à une fonction (à valeur dans $\mathbb{R}^n$ ou dans $\mathbb{C}^n$) $f(t)$ une nouvelle fonction dite transformée de $f(t)$, notée traditionnellement $F(p)$, via une intégrale. La transformation de Laplace est bijective et par usage de tables il est possible d'inverser la transformation. Le grand avantage de la transformation de Laplace est que la plupart des opérations courantes sur la fonction originale $f(t)$, telle que la dérivation, ou un décalage sur la variable t, ont une traduction (plus) simple sur la transformée $F(p)$. Ainsi la transformée de Laplace de la dérivée $f'(t)$ est simplement $pF(p) - f(0^-)$, et la transformée de la fonction « décalée » $f(t-\tau)$ est simplement $e^{-p\tau}F(p)$. Cette transformation fut introduite pour la première fois sur une forme proche de celle utilisée par Laplace en 1774, dans le cadre de la théorie des probabilités

La transformée de Laplace est proche de la transformée de Fourier qui est également utilisée pour résoudre les équations différentielles, mais contrairement à cette dernière elle tient compte des conditions initiales et peut ainsi être utilisée en théorie des vibrations mécaniques ou en électricité dans l'étude des régimes forcés sans négliger le régime transitoire. Dans ce type d'analyse, la transformée de Laplace est souvent interprétée comme un passage du domaine temps, dans lequel les entrées et sorties sont des fonctions du temps, dans le domaine des fréquences, dans lequel les mêmes entrées et sorties sont des fonctions de la « fréquence » (complexe) p. Ainsi il est possible d'analyser simplement l'effet du système sur l'entrée pour donner la sortie en terme d'opérations algébriques simples.

5.1 Définition

Transformée de Laplace : *La transformée de Laplace monolatérale d'une fonction f (éventuellement généralisée, telle que la « fonction de Dirac ») d'une variable réelle t, à support positif, est la fonction F de la variable complexe p, définie par :*

$$F(p) = \mathcal{L}\{f(t)\} = \int_{0^-}^{+\infty} e^{-pt} f(t)\, dt.$$

Les propriétés de cette transformation lui confèrent une grande utilité dans l'analyse des systèmes dynamiques linéaires. La plus intéressante de ces propriétés est que l'intégration et la dérivation sont transformées en division et multiplication par p, de la même manière que le logarithme transforme la multiplication en addition. Elle permet ainsi de ramener la résolution des équations différentielles linéaires à coefficients constants à la résolution d'équations affines (dont les solutions sont des fonctions rationnelles de p).

Remarque : la notation « s » (variable de Laplace) est souvent utilisée dans les pays anglo-saxons alors que la notation « p » est utilisée notamment en France et en Allemagne.

La transformation de Laplace est très utilisée par les ingénieurs pour résoudre des équations différentielles et déterminer la fonction de transfert d'un système linéaire. Par exemple, en électronique, contrairement à la décomposition de Fourier qui est utilisée pour la détermination du spectre d'un signal périodique ou même quelconque, elle tient compte de l'existence d'un régime transitoire précédant le régime permanent (exemple : la prise en compte de l'allure du signal avant et après la mise en marche d'un générateur de fréquence).

Il suffit en effet de transposer l'équation différentielle dans le domaine de Laplace pour obtenir une équation beaucoup plus simple à manipuler.

Par exemple, lors de l'étude d'une machine à courant continu : $e(t) = R \cdot i(t) + L \frac{di(t)}{dt}$ dans le domaine temporel devient $E(p) = R \cdot I(p) + p \cdot L \cdot I(p)$ dans le domaine de Laplace. Ceci n'est valable qu'à conditions initiales nulles ($i(0) = 0$).

Inversion L'inversion de la transformation de Laplace s'effectue par le biais d'une intégrale dans le plan complexe. À l'aide du théorème des résidus, on démontre la formule de Bromwich :

$$f(t) = \mathcal{L}^{-1}\{F(p)\} = \frac{1}{2\pi i} \int_{\gamma - i\cdot\infty}^{\gamma + i\cdot\infty} e^{pt} F(p)\, dp$$

où γ est choisi de sorte que l'intégrale soit convergente, ce qui implique que γ soit supérieur à la partie réelle de toute singularité de $F(p)$ et qu'à l'infini, $|F(p)|$ tende vers 0 au moins aussi rapidement que $\frac{1}{|p|^2}$. Lorsque cette dernière condition n'est pas satisfaite, la formule ci-dessus est encore utilisable s'il existe un entier n tel que $|p^{-n}F(p)|$ tende vers 0 aussi rapidement que $\frac{1}{|p|^2}$, c'est-à-dire lorsque, pour $|p|$ tendant vers l'infini, $|F(p)|$ est majorée par un polynôme en $|p|$. En remplaçant $F(p)$ par $p^{-n}F(p)$ dans l'intégrale ci-dessus, on trouve dans le membre de gauche de l'égalité une fonction généralisée à support positif dont la dérivée d'ordre n (au sens des distributions) est la fonction généralisée (elle aussi à support positif) cherchée.

En pratique néanmoins, la formule de Bromwich-Mellin est peu utilisée, et on calcule les inverses des transformées de Laplace à partir des tables de transformées de Laplace.

5.2 Propriétés

Le tableau de la page suivante rassemble les propriétés principales de la transformée de Laplace unilatérale

	Domaine temporel	Domaine "p"	Commentaires
Linéarité	$af(t) + bg(t)$	$aF(p) + bG(p)$	Résulte des règles de base de l'intégration.
Dérivée de la transformée	$tf(t)$	$-F'(p)$	F' est la dérivée première de F.
Dérivées d'ordre n de la transformée	$t^n f(t)$	$(-1)^n F^{(n)}(p)$	Forme plus générale, dérivéen-ième de F(p).
Dérivée première de la fonction dans le domaine temporel	$f'(t)$	$pF(p) - f(0^-)$	f est supposée dérivable, et sa dérivée est supposé tendre vers 0 exponentiellement. Peut-être obtenue par intégration par parties.
Dérivée seconde	$f''(t)$	$p^2F(p) - pf(0^-) - f'(0^-))$	f est supposé deux fois dérivable et sa dérivée seconde converger exponentiellement à l'infini.
Dérivée n-ième de f	$f^{(n)}(t)$	$p^nF(s) - p^{n-1}f(0^-) - \cdots - f^{(n-1)}(0^-)$	f est supposé n-fois dérivable, avec une dérivée n-ième à convergence exponentielle à l'infini.
Intégration de la transformée de Laplace	$\frac{f(t)}{t}$	$\int_p^\infty F(\sigma)\,d\sigma$	
Intégration	$\int_0^t f(\tau)\,d\tau = (u * f)(t)$	$\frac{1}{p}F(p)$	$u(t)$ est la fonction échelon de Heaviside. (u * f)(t) est le produit de convolution de u(t) et f(t).
Dilatation échelle de temps	$f(at)$	$\frac{1}{\lvert a\rvert}F\left(\frac{p}{a}\right)$	
Décalage sur p	$e^{at}f(t)$	$F(p-a)$	
Décalage domaine temporel	$f(t-a)u(t-a)$	$e^{-ap}F(p)$	$u(t)$ est la fonction échelon de Heaviside step function
Multiplication	$f(t)g(t)$	$\frac{1}{2\pi i}\lim_{T\to\infty}\int_{c-iT}^{c+iT} F(\sigma)G(p-\sigma)\,d\sigma$	L'intégration est effectuée le long de la ligne verticale $\Re(\sigma) = c$ qui est entièrement située dans le rayon de convergence de F.
Produit de convolution	$(f * g)(t) = \int_0^t f(\tau)g(t-\tau)\,d\tau$	$F(p) \cdot G(p)$	$f(t)$ et $g(t)$ sont étendues sur $\mathbb{R}$ pour la définition du produit de convolution.
Conjugaison complexe	$f^*(t)$	$F^*(p^*)$	
Fonction de corrélation	$f(t) \star g(t)$	$F^*(-p^*) \cdot G(p)$	
Fonction périodique	$f(t)$	$\frac{1}{1-e^{-Tp}}\int_0^T e^{-pt}f(t)\,dt$	$f(t)$ est une fonction périodique de période T telle que $f(t) = f(t+T)$, $\forall t \geq 0$. Ceci résulte de la propriété de décalage dans le domaine temporel et de la série géométrique.

5.3 Quelques transformées usuelles

Fonction	Domaine temporel $x(t) = \mathcal{L}^{-1}\{X(p)\}$	Transformée de Laplace $X(p) = \mathcal{L}\{x(t)\}$	Région de convergence
délai idéal	$\delta(t-\tau)$	$e^{-\tau p}$	
impulsion unité	$\delta(t)$	1	$\forall\ p$
retard à la n-ième puissance avec décalage fréquentiel	$\frac{(t-\tau)^n}{n!}e^{-\alpha(t-\tau)}\cdot\Upsilon(t-\tau)$	$\frac{e^{-\tau p}}{(p+\alpha)^{n+1}}$	$\Re(p) > 0$
puissance n-ième	$\frac{t^n}{n!}\cdot\Upsilon(t)$	$\frac{1}{p^{n+1}}$	$\mathrm{Re}(p) > 0$
puissance q-ième	$\frac{t^q}{\Gamma(q+1)}\cdot\Upsilon(t)$	$\frac{1}{p^{q+1}}$	$\mathrm{Re}(p) > 0$
échelon unité	$\Upsilon(t)$	$\frac{1}{p}$	$\mathrm{Re}(p) > 0$
échelon retardé	$\Upsilon(t-\tau)$	$\frac{e^{-\tau p}}{p}$	$\mathrm{Re}(p) > 0$
rampe	$t\cdot\Upsilon(t)$	$\frac{1}{p^2}$	$\mathrm{Re}(p) > 0$
retard avec décalage fréquentiel	$\frac{t^n}{n!}e^{-\alpha t}\cdot\Upsilon(t)$	$\frac{1}{(p+\alpha)^{n+1}}$	$\mathrm{Re}(p) > -\alpha$
décroissance exponentielle	$e^{-\alpha t}\cdot\Upsilon(t)$	$\frac{1}{p+\alpha}$	$\mathrm{Re}(p) > -\alpha$
approche exponentielle	$(1-e^{-\alpha t})\cdot\Upsilon(t)$	$\frac{\alpha}{p(p+\alpha)}$	$\mathrm{Re}(p) > 0$
sinus	$\sin(\omega t)\cdot\Upsilon(t)$	$\frac{\omega}{p^2+\omega^2}$	$\mathrm{Re}(p) > 0$
cosinus	$\cos(\omega t)\cdot\Upsilon(t)$	$\frac{p}{p^2+\omega^2}$	$\mathrm{Re}(p) > 0$
sinus hyperbolique	$\sinh(\alpha t)\cdot\Upsilon(t)$	$\frac{\alpha}{p^2-\alpha^2}$	$\mathrm{Re}(p) > \lvert\alpha\rvert$
cosinus hyperbolique	$\cosh(\alpha t)\cdot\Upsilon(t)$	$\frac{p}{p^2-\alpha^2}$	$\mathrm{Re}(p) > \lvert\alpha\rvert$
décroissance exponentielle d'une onde sinusoidale	$e^{-\alpha t}\sin(\omega t)\cdot\Upsilon(t)$	$\frac{\omega}{(p+\alpha)^2+\omega^2}$	$\mathrm{Re}(p) > -\alpha$
décroissance exponentielle d'une onde cosinusoidale	$e^{-\alpha t}\cos(\omega t)\cdot\Upsilon(t)$	$\frac{p+\alpha}{(p+\alpha)^2+\omega^2}$	$\mathrm{Re}(p) > -\alpha$
n-ième racine	$\sqrt[n]{t}\cdot\Upsilon(t)$	$p^{-(n+1)/n}\cdot\Gamma\left(1+\frac{1}{n}\right)$	$\mathrm{Re}(p) > 0$
logarithme	$\ln\left(\frac{t}{t_0}\right)\cdot\Upsilon(t)$	$-\frac{t_0}{p}\ [\ \ln(t_0 p)+\gamma\ ]$	$\mathrm{Re}(p) > 0$
fonction de Bessel du premier type, d'ordre n (n > -1)	$J_n(\omega t)\cdot\Upsilon(t)$	$\frac{\omega^n\left(p+\sqrt{p^2+\omega^2}\right)^{-n}}{\sqrt{p^2+\omega^2}}$	$\mathrm{Re}(p) > 0$

où

— $\Upsilon(t)$ représente la fonction de Heaviside.
— $\delta(t)$ représente la fonction de Dirac.
— $\Gamma(z)$ est la fonction Gamma.
— γ est la constante d'Euler-Mascheroni.
— t , est un nombre réel, il représente typiquement le temps, mais peut désigner n'importe quelle autre quantité.
— p est un nombre complexe.
— q est un nombre réel ($q + 1 > 0$).
— α , β , τ , et ω sont des nombres réels.
— n est un entier.

5.4 Exemples d'utilisations

On considère un circuit dit « R,C », constituée d'une résistance électrique de valeur R et d'un condensateur de capacité électrique C, placés en série. Dans tout les cas on considère que le circuit n'est placé aux bornes d'un générateur idéal de tension délivrant une tension (en général) variable $u(t)$ qu'à un instant choisi pour origine des dates, et que le condensateur est initialement déchargé. On a ainsi respectivement pour la charge $q(t)$ du condensateur et l'intensité dans le circuit $i(t) \equiv \frac{dq}{dt}$ les conditions initiales suivantes : $q(0^-) = 0$, $i(0^-) = 0$. On applique la tension $u(t)$ suivante :

$$u(t) = \begin{cases} 0, \text{ si } t < 0 \\ U_0 = cte, \text{ si } t \geq 0 \end{cases}$$

et l'équation différentielle reliant la réponse $q(t)$ à l'entrée $u(t)$ est en appliquant les lois usuelles de l'électricité :

$$U_0 \Upsilon(t) = R\frac{dq}{dt} + \frac{q(t)}{C},$$

soit encore en posant $\tau \equiv RC$:

$$\frac{CU_0}{\tau} = \frac{q(t)}{\tau} + \frac{dq}{dt}.$$

On prend la transformée de Laplace membre à membre de cette dernière équation, en notant $Q(p)$ la transformée de $q(t)$, il vient (en prenant en compte le fait que $q(0^-) = 0$) :

$$Q(p) = CU_0 \frac{\frac{1}{\tau}}{p\left(\left(\frac{1}{\tau}\right) + p\right)},$$

ce qui peut aussi s'écrire sous la forme :

$$Q(p) = H(p)\, U(p), \text{ avec } H(p) \equiv \frac{(1/\tau)}{[(1/\tau) + p]},$$

fonction de transfert du système R,C, et

$$U(p) = CU_0/p,$$

transformée de Laplace de l'entrée, que l'on peut aussitôt inverser en (on utilise la table ci-dessus avec $\alpha = 1/\tau$) :

$$q(t) = U_0 C\left[1 - e^{-t/\tau}\right].$$

La transformation de Laplace permet de s'abstraire complétement de la résolution de l'équation différentielle dans l'espace des temps par un passage dans « l'espace p ». Par ailleurs, la prise en compte des conditions initiales est effectuée lors de la transformation.

Exercice : Reprendre l'exemple précédent avec en entrée une impulsion.

Exercice : Déterminer en utilisant la transformée de Laplace, les solutions du système différentiel

$$\frac{dX}{dt} = -\omega Y(t) - kX(t) \tag{5.1}$$

$$\frac{dY}{dt} = \omega X(t) - kY(t) \tag{5.2}$$

$$\tag{5.3}$$

Exercice : Déterminer en utilisant la transformée de Laplace, les solutions de l'équation différentielle

$$\frac{d^2Y}{dt^2} + Y(t) = F(t) \tag{5.4}$$

$$Y'(0) = Y(0) = 0 \tag{5.5}$$

avec $F(t) = 1$ si $0 \geq t \geq \frac{\pi}{2}$, $F(t) = -1$ si $\frac{\pi}{2} < t \geq \pi$ et $F(t) = 0$ sinon.

Chapitre 6

Polynômes orthogonaux

Sommaire

En mathématiques, une suite de polynômes orthogonaux est une suite infinie de polynômes $p_0(x)$, $p_1(x)$, $p_2(x)$ ... à coefficients réels, dans laquelle chaque $p_n(x)$ est de degré n, et telle que les polynômes de la suite sont orthogonaux deux à deux pour un produit scalaire de fonctions donné.

6.1 Introduction

Le produit scalaire de fonctions le plus simple est l'intégrale du produit de ces fonctions, sur un intervalle borné :

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

Plus généralement, on peut introduire une « fonction poids » $\omega(x)$ dans l'intégrale (sur l'intervalle d'intégration $]a, b[$, W doit être à valeurs finies et strictement positives, et l'intégrale du produit de la fonction poids par un polynôme doit être finie ; les bornes a, b peuvent être infinies) :

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

Avec cette définition du produit scalaire, deux fonctions sont orthogonales entre elles si leur produit scalaire est égal à 0 (de la même manière que deux vecteurs sont orthogonaux (perpendiculaires) si leur produit scalaire égale zéro). On introduit alors la norme associée :

$$||f|| = \sqrt{\langle f, f\rangle}\,:$$

le produit scalaire fait de l'ensemble de toutes les fonctions de norme finie un espace de Hilbert. L'intervalle d'intégration est appelé intervalle d'orthogonalité.

6.2 Propriétés

Toute suite de polynômes p_0, p_1 ..., où chaque p_k est de degré k, est une base de l'espace vectoriel $\mathbb{R}[x]$ (de dimension infinie) de tous les polynômes. Une suite de polynômes orthogonaux est une telle

base qui est, de plus, orthogonale pour un certain produit scalaire. Ce produit scalaire étant fixé, une telle suite est presque unique (unique à produit près de ses vecteurs par des scalaires non nuls), et peut s'obtenir à partir de la base canonique $(1, x, x^2, \ldots)$ (non orthogonale en général), par le procédé de Gram-Schmidt.

Quand on construit une base orthogonale, on peut être tenté de la rendre orthonormale, c'est-à-dire telle que $\langle p_n, p_n \rangle = 1$ pour tout n, en divisant chaque p_n par sa norme. Dans le cas des polynômes, on préfère ne pas imposer cette condition supplémentaire car il en résulterait souvent des coefficients contenant des racines carrées. On préfère souvent choisir un multiplicateur tel que les coefficients restent rationnels, et donnent des formules aussi simples que possible. C'est la standardisation. Les polynômes « classiques » énumérés ci-dessous ont été ainsi standardisés ; typiquement, le coefficient de leur terme de plus haut degré ou leur valeur en un point ont été mis à une quantité donnée (pour les polynômes de Legendre, $P_n(1) = 1$). Cette standardisation est une convention qui pourrait aussi parfois être obtenue par une mise à l'échelle de la fonction poids correspondante.

Notons

$$h_n = \langle p_n,\ p_n \rangle$$

(la norme de p_n est la racine carrée de h_n). Les valeurs de h_n pour les polynômes standardisés sont énumérées dans le tableau ci-dessous. On a

$$\langle p_m,\ p_n \rangle \ = \ \delta_{mn}\, h_n;$$

où $\delta_{m,n}$ est le delta de Kronecker.

Toute suite (p_k) de polynômes orthogonaux possède un grand nombre de propriétés remarquables. Pour commencer :

> **Lemme :** *la famille $(p_0, \ldots, p_n)$ est une base de $\mathbb{R}_n[x]$, l'espace des polynômes de degrés n.*

Ceci est dû au fait que p_k est de degré k.

> **Lemme :** *Pour tous n, le polynôme p_n est orthogonal à $\mathbb{R}_{n-1}[x]$.*

Qui découle de l'orthogonalité de p_k pris deux à deux.

6.3 Relation de récurrence

> **Théorème :** *Pour toute suite de polynômes orthogonaux, il existe une relation de récurrence relativement à trois polynômes consécutifs.*

$$p_{n+1} \ = \ (a_n\, x + b_n)\ p_n \ - \ c_n\ p_{n-1}$$

Les coefficients a_n,b_n,c_n sont donnés par

$$a_n = \frac{k_{n+1}}{k_n}, \qquad b_n = a_n \left(\frac{k'_{n+1}}{k_{n+1}} - \frac{k'_n}{k_n} \right), \qquad c_n = a_n \left(\frac{k_{n-1} h_n}{k_n h_{n-1}} \right),$$

où k_j et k'_j désignent les deux premiers coefficients de p_j :

$$p_j(x) = k_j x^j + k'_j x^{j-1} + \cdots$$

Rappelons que h_j le produit scalaire de p_j par lui-même :

$$h_j \ = \ \langle p_j,\ p_j \rangle.$$

(Par convention, $c_0,\ p_{-1},\ k'_0$ sont nuls.)

Démonstration :

Avec les valeurs données pour a_n et b_n, le polynôme $(a_n\,x + b_n)p_n - p_{n+1}$ est de degré $< n$ (les termes de degrés $n+1$ et n s'éliminent). On peut donc l'exprimer sous forme d'une combinaison linéaire des éléments de la base $(p_j)_{j=0}^{n-1}$ de $\mathbb{R}_{n-1}[x]$:

$$(a_n x + b_n)p_n - p_{n+1} = \sum_{j=0}^{n-1} \mu_{n,j} p_j,$$

avec

$$h_j \mu_{n,j} = \langle (a_n x + b_n)p_n - p_{n+1}, p_j \rangle = a_n \langle x p_n, p_j \rangle$$

(car pour $j < n$, p_j est orthogonal à p_n et p_{n+1}).

De plus, de par la forme intégrale du produit scalaire,

$$\langle x p_n, p_j \rangle = \langle p_n, x p_j \rangle.$$

Pour $j < n-1$, ce produit scalaire est nul car xp_j est de degré $< n$. Pour $j = n-1$, il est égal à $\frac{h_n}{a_{n-1}}$ car (par le même raisonnement qu'au début) $a_{n-1}\,xp_{n-1} - p_n$ est de degré $< n$.

On peut conclure :

$$(a_n x + b_n)p_n - p_{n+1} = c_n p_{n-1},$$

avec

$$c_n = \mu_{n,n-1} = \frac{a_n}{h_{n-1}}\,\frac{h_n}{a_{n-1}} = a_n \left(\frac{k_{n-1} h_n}{k_n h_{n-1}} \right).$$

6.4 Existence et position de racines réelles

Le résultat suivant est fondamental pour l'analyse numérique, il justifie en grande partie l'engouement pour les famille de polynômes orthogonaux.

> **Théorème :** *Tout polynôme d'une suite de polynômes orthogonaux dont le degré n est supérieur ou égal à 1 admet n racines distinctes, toutes réelles, et situées strictement à l'intérieur de l'intervalle d'intégration.*

C'est une propriété remarquable : il est rare, pour un polynôme de degré élevé dont les coefficients ont été choisis au hasard, d'avoir toutes ses racines réelles *Démonstration :* Soit m le nombre des points où p_n change de signe à l'intérieur de l'intervalle d'orthogonalité ; notons $x_1 \ldots x_m$ ces points. Ce sont les racines de p_n d'ordre impair appartenant à l'intervalle d'orthogonalité. D'après le théorème fondamental de l'algèbre, $m \leq n$. On va montrer $m = n$. Soit $S(x) = \prod_{j=1}^{m}(x - x_j)$; c'est un polynôme de degré m qui change de signe en chaque point x_j ; $S(x)\ p_n(x)$ est donc strictement positif, ou strictement négatif, partout sur l'intervalle d'intégration sauf aux points x_j , et il en est donc de même de $S(x)p_n(x)W(x)$. Ainsi, $\langle S, p_n \rangle$, l'intégrale de ce produit, est non nul. Mais p_n est orthogonal à tous les polynômes de degré inférieur, donc le degré de S doit être n.

> **Théorème :** *Les racines des polynômes se trouvent strictement entre les racines du polynôme de degré supérieur dans la suite.*

Démonstration : On met d'abord tous les polynômes sous une forme standardisée telle que le coefficient dominant soit positif (ce qui ne change pas les racines), puis on effectue une récurrence sur n. Pour $n = 0$ il n'y a rien à démontrer. Supposons la propriété acquise jusqu'au rang n. Notons $x_1 < \ldots < x_n$ les racines de p_n et $y_0 < \ldots < y_n$ celles de p_{n+1}. La relation de récurrence donne $p_{n+1}(x_j) = -c_n p_{n-1}(xj)$ avec (d'après le choix de standardisation) $c_n > 0$. Or par hypothèse de récurrence, $(-1)^{n-j}\ p_{n-1}(x_j) > 0$. On en déduit $(-1)^{n+1-j} p_{n+1}(x_j) > 0$. En outre, $\forall x > y_n,\ p_{n+1}(x) > 0$ et $\forall x < y_0,\ (-1)^{n+1} p_{n+1}(x) > 0$. Ceci permet de conclure : $y_0 < x_1 < y_1 < \ldots < x_n < y_n$.

6.5 Équations différentielles conduisant à des polynômes orthogonaux

Une importante classe des polynômes orthogonaux provient d'une équation différentielle de Sturm-Liouville de la forme

$$Q(x)\,f'' + L(x)\,f' + \lambda\,f = 0$$

où Q est un polynôme quadratique donné et L un polynôme linéaire donné. La fonction f est inconnue, et la constante λ est un paramètre. On peut remarquer qu'une solution polynomiale est a priori envisageable pour une telle équation, les degrés des termes étant compatibles. Cependant, les solutions de cette équation différentielle ont des singularités, à moins que λ ne prenne des valeurs spécifiques. La suite de ces valeurs $\lambda_0, \lambda_1, \lambda_2 \ldots$ conduit à une suite de polynômes solutions $P_0, P_1, P_2 \ldots$ si l'une des assertions suivantes est vérifiée :

1. Q est vraiment quadratique, L est linéaire, Q a deux racines réelles distinctes, la racine de L est située entre les deux racines de Q, et les termes de plus haut degré de Q et L ont le même signe.
2. Q n'est pas quadratique, mais linéaire, L est linéaire, les racines de Q et L sont différentes, et les termes de plus haut degré de Q et L ont le même signe si la racine de L est plus petite que celle de Q, ou inversement.
3. Q est un polynôme constant non nul, L est linéaire, et le terme de plus haut degré de L est de signe opposé à celui de Q.

Ces trois cas conduisent respectivement aux polynômes de Jacobi, de Laguerre et d'Hermite. Pour chacun de ces cas :

— La solution est une suite de polynômes $P_0, P_1, P_2 \ldots$, chaque P_n ayant un degré n, et correspondant au nombre λ_n.
— L'intervalle d'orthogonalité est limité par les racines de Q.
— La racine de L est à l'intérieur de l'intervalle d'orthogonalité.
— En notant $R(x) = e^{\int_{x_0}^{x} \frac{L(t)}{Q(t)}\,dt}$, les polynômes sont orthogonaux sous la fonction poids $\omega(x) = \frac{R(x)}{Q(x)}$
— $\omega(x)$ ne peut pas s'annuler ou prendre une valeur infinie dans l'intervalle, bien qu'il puisse le faire aux extrémités.
— $\omega(x)$ peut être choisi positif sur l'intervalle (multiplier l'équation différentielle par -1 si nécessaire)

En raison de la constante d'intégration, la quantité $R(x)$ est définie à une constante multiplicative près. Le tableau ci-dessous donne les valeurs "officielles" de $R(x)$ et $\omega(x)$.

6.6 Formule de Rodrigues

Avec les hypothèses de la section précédente, $P_n(x)$ est proportionnel à $\frac{1}{\omega(x)}\,\frac{d^n}{dx^n}\,(\omega(x)[Q(x)]^n)$ équation mieux connue sous le nom de « formule de Rodrigues ». Elle est souvent écrite :

Formule de Rodrigues :

$$P_n(x) = \frac{1}{e_n \omega(x)}\,\frac{d^n}{dx^n}\,(\omega(x)[Q(x)]^n)$$

où les nombres e_n dépendent de la normalisation.

Les valeurs de e_n sont données dans le tableau plus bas.

Pour démontrer cette formule on vérifie, dans chacun des trois cas ci-dessus, que le polynôme P_n qu'elle fournit est bien un polynôme de degré n, puis, par intégrations par parties répétées, que pour tout polynôme P, $\langle \frac{1}{W}(\omega Q^n)^{(n)}, P\rangle$ est égal à $(-1)^n\langle Q^n, P^{(n)}\rangle$, donc est nul si P est de degré inférieur à n. Cette méthode montre en outre que $h_n e_n = (-1)^n n! k_n \int_a^b (Q(x))^n \omega(x) dx$.

6.7 Tableau des polynômes orthogonaux classiques

Nom et symbole conventionnel	Tchebychev, T_n	Tchebychev (seconde sorte), U_n	Legendre, P_n	Hermite (forme physique), H_n	Laguerre associé, $L_n^{(\alpha)}$	Laguerre, L_n
Limite d'orthogonalité	-1, 1	-1, 1	-1, 1	$-\infty$, ∞	0, ∞	0, ∞
Poids, $W(x)$	$(1-x^2)^{-1/2}$	$(1-x^2)^{1/2}$	1	e^{-x^2}	$x^\alpha e^{-x}$	e^{-x}
Normalisation	$T_n(1)=1$	$U_n(1)=n+1$	$P_n(1)=1$	Coefficient dominant $=2^n$	Coefficient dominant $=\frac{(-1)^n}{n!}$	Coefficient dominant $=\frac{(-1)^n}{n!}$
Carré de la norme h_n	π si n=0 $\pi/2$ sinon	$\pi/2$	$\frac{2}{2n+1}$	$2^n\, n!\sqrt{\pi}$	1	1
Coefficient dominant k_n	2^{n-1}	2^n	$\frac{(2n)!}{2^n\,(n!)^2}$	2^n	$\frac{(-1)^n}{n!}$	$\frac{(-1)^n}{n!}$
Coefficient suivant k'_n	0	0	0	0	$\frac{(-1)^{n+1}(n+^alpha)}{(n-1)!}$	$\frac{(-1)^{n+1}n}{(n-1)!}$
Q	$1-x^2$	$1-x^2$	$1-x^2$	1	x	x
L	$-x$	$-3x$	$-2x$	$-2x$	$\alpha+1-x$	$1-x$
$R(x)=e^{\int \frac{L(x)}{Q(x)}dx}$	$(1-x^2)^{1/2}$	$(1-x^2)^{3/2}$	$1-x^2$	e^{-x^2}	$x^{\alpha+1}\,e^{-x}$	xe^{-x}
Constante dans l'équation différentielle, λ_n	n^2	$n(n+2)$	$n(n+1)$	$2n$	n	n
Constante dans la formule de Rodrigues, e_n	$(-2)^n\frac{\Gamma(n+1/2)}{\sqrt{\pi}}$	$2(-2)^n\frac{\Gamma(n+3/2)}{(n+1)\sqrt{\pi}}$	$(-2)^n n!$	$(-1)^n$	$n!$	$n!$
Relation de récurrence, a_n	2	2	$\frac{2n+1}{n+1}$	2	$\frac{-1}{n+1}$	$\frac{-1}{n+1}$
Relation de récurrence, b_n	0	0	0	0	$\frac{2n+1+\alpha}{n+1}$	$\frac{2n+1}{n+1}$
Relation de récurrence, c_n	1	1	$\frac{n}{n+1}$	$2n$	$\frac{n+\alpha}{n+1}$	$\frac{n}{n+1}$

Chapitre 7

Équations différentielles ordinaires

Sommaire

Les équations différentielles constituent un outil fondamental pour la modélisation dynamique des systèmes naturels. Avant d'envisager les méthodes numériques de résolution, il importe de rappeler les développements analytiques élémentaires permettant la résolution de certaines équations.

On appelle équation différentielle une équation établissant une relation entre la variable indépendante, x la fonction inconnue $y = f(x)$ et ses dérivées $y', y", \ldots, y^{(n)}$. On appelle ordre d'une équation différentielle l'ordre de la dérivée la plus élevée contenue dans cette équation.

7.1 Équations différentielles du premier ordre

Une équation différentielle ordinaire (EDO) du premier ordre est une relation entre la dérivée première d'une fonction inconnue d'une variable réelle ou complexe et la valeur de la fonction.

Définition : *Soit y une fonction inconnue d'une variable x et y' sa dérivée première, soit F une fonction quelconque alors la définition formelle d'une équation différentielle du premier ordre est*

$$F(y, y', x) = 0.$$

Lorsque cette équation est résoluble en y', on peut la mettre sous la forme

$$y' = f(y, x).$$

Dans le cas d'une équation résoluble, on a le théorème fondamental suivant.

Théorème : *Si dans l'équation*

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

la fonction $f(x, y)$ et sa dérivée partielle $\frac{\partial f}{partialy}$ par rapport à y sont continues dans un certain domaine D du plan xOy et si (x_0, y_0) est un point de ce domaine, il existe une solution unique $y = \phi(x)$ satisfaisant à la condition $y = y_0$ lorsque $x = x_0$.

Géométriquement, ce théorème signifie qu'il existe une fonction $y = \phi(x)$ et une seule dont la courbe représentative passe par le point (x_0, y_0). La condition que la fonction y doit prendre la valeur donnée y_0 lorsque $x = x_0$ s'appelle la condition initiale.

Solution générale : *On appelle solution générale d'une équation du premier ordre une fonction*

$$y = \phi(x, C),$$

dépendant d'une constante arbitraire C et satisfaisant aux conditions suivantes

1. *elle satisfait à l'équation différentielle, quelle que soit la valeur concrète de la constante C.*
2. *quelle que soit la condition initiale $y = y_0$ lorsque, $x = x_0$on peut trouver une valeur $C = C_0$ telle que la fonction $y = \phi(x, C_0)$ vérifie la condition initiale donnée.*

Il arrive que la recherche de la solution générale d'une équation différentielle conduise à une égalité de la forme

$$\Phi(y, x, C)$$

qui ne puisse être résolue en y. Une telle égalité donnant implicitement la solution générale est appelée **intégrale générale** de l'équation différentielle.

On appelle **solution particulière** toute fonction $y = \phi(x, C_0)$ déduite de la solution générale. La relation $\Phi(y, x, C_0)$ est dite une **intégrale particulière**.

Résoudre (ou intégrer) une équation différentielle consiste à :

1. chercher sa solution générale ou son intégrale générale
2. chercher la solution particulière satisfaisant aux conditions initiales

c'est trouver une fonction g satisfaisant la relation définissant l'EDO dans un domaine à préciser et vérifiant une condition particulière $g(X_0) = g_0$.

7.1.1 Équation à variable séparable

Les équations différentielles ordinaires (EDO) du premier ordre à variable séparable constituent une sous-classe remarquable dans la mesure où une méthode générale de résolution existe.

Définition : *Une équation différentielle du premier ordre à variables séparables est une EDO pouvant se mettre sous la forme :*

$$b(y)\,\frac{d\,y(x)}{dx} = a(x),$$

où a et b sont des fonctions données et y la fonction inconnue.

Il est facile d'intégrer une telle équation lorsque a et b sont continues. En effet, soit $y = \phi(x)$ l'une de ses solutions et posons

$$A(x) = \int a(x)\mathrm{d}x \quad \text{et} \quad B(y) = \int b(y)\mathrm{d}xy.$$

En vertu de la continuité de a et b, les deux fonctions $A(x)$ et $B\left(\phi(x)\right)$ sont dérivables et admettent respectivement pour dérivées $a(x)$ et $b\left[\phi(x)\right] = \phi'(x)$. L'équation étant satisfaite, par hypothèse, quand on prend $y = \phi(x)$, on a :

$$\frac{d\,B\left(\phi(x)\right)}{dx} = \frac{d\,A\left(x\right)}{dx}.$$

Il existe une constante C telle que

$$B\left(\phi(x)\right) = A\left(x\right) + C.$$

La fonction $y = \phi(x)$ satisfait donc l'équation :

$$B(y) = A(x) + C,$$

qui définit l'ensemble des solutions de l'équation différentielle.

Exemple : L'équation du premier ordre

$$\frac{dy}{dx} = -\frac{y}{x}$$

est à variable séparable avec $b(y) = y^{-1}$ et $a(x) = -x^{-1}$. Un calcul élémentaire conduit à $A(x) = -\ln(x)$ et $B(y) = \ln(y)$. En posant $C = \ln c$, on démontre que l'ensemble des solutions est de la forme $y(x) = c/x$.

7.1.2 Équation homogène du premier ordre

Définition : *On dit qu'une équation différentielle du premier ordre es homogène si, pour $x \neq 0$, elle peut s'écrire :*

$$y' = f\left(\frac{y}{x}\right),$$

où f est une fonction définie dans un certain intervalle.

La résolution d'une équation homogène se mène à l'aide du paramètre $u = y/x$ dont la différentiation conduit à $dy = x\,du + u\,dx$. En récrivant l'équation différentielle sous la forme $dy = f\left(\frac{y}{x}\right)dx$ on en déduit :

$$\left[f(u) - u\right]dx = x\,du,$$

qui est une équation à variables séparables. La résolution se mène en deux temps

1. Si il existe un paramètre α tel que $f(\alpha) = \alpha$ alors la courbe $y = \alpha x$ est une solution particulière de l'équation différentielle,
2. En désignant par $\phi(u)$ l'intégrale $\int \frac{du}{f(u)-u}$ on obtient la solution générale suivante

$$x = \lambda e^{\phi(u)}, \qquad y = \lambda u e^{\phi(u)},$$

où λ est une constante arbitraire.

Exemple : L'équation du premier ordre

$$\frac{dy}{dx} = -\frac{y}{x}$$

est homogène. En posant $u = y/x$ et $f(u) = -u$, on obtient $\phi(u) = -\int 1/2udu = \ln u^{-1/2}$ et donc $y = \lambda u^{1/2}$ soit en substituant $y = c/x$.

7.1.3 Équations linéaires du premier ordre

Définition : *Un équation différentielle du premier ordre est dite linéaire quand elle est linéaire par rapport à la fonction inconnue y et à sa dérivée. Une telle équation peut toujours s'écrire :*

$$a(x)\, y' + b(x)\, y = c(x).$$

Dans la suite, nous supposerons a, b et c définies et continues sur un certain intervalle I, sur lequel a ne possède pas de racine.

On résout de manière générale l'équation par la méthode de variation des constantes. Celle-ci consiste à se ramener, par un changement de fonction variable, à un problème de calcul de primitive. On écrit la solution générale de l'équation homogène associée $ay' + by = 0$

$$y(x) = Ke^{-A(x)}, K \in \mathbb{R},$$

on prend pour nouvelle fonction variable la fonction k définie par la relation

$$y(x) = k(x)e^{-A(x)},$$

ce qui explique la formulation imagée : on fait « varier la constante », en fait on la remplace par une fonction.

En reportant dans l'équation initiale, on obtient une équation équivalente à l'équation initiale, mais portant sur k

$$a(x)k'(x) = c(x)e^{A(x)}.$$

En notant B une primitive de la fonction $\frac{ce^A}{a}$, l'ensemble des solutions est

$$k(x) = B(x) + C.$$

La solution générale s'écrit alors sous la forme

$$y(x) = (B(x) + C)e^{-A(x)}.$$

Soit finalement

$$f = \exp\left(-\int \frac{b(x)}{a(x)}\mathrm{d}x\right)\left\{C + \int \frac{c(x)}{a(x)} \exp\left(\int \frac{b(x)}{a(x)}\mathrm{d}x\right)\mathrm{d}x\right\}$$

De nouveau, il faut réaliser un calcul de primitive, ce qui peut empêcher de donner l'expression de la solution à l'aide des fonctions usuelles.

7.1.4 Autres équations remarquables

Il existe des méthodes générales de résolution pour les équations suivantes

Équations de Bernoulli Ce sont des équations de la forme

$$y' + P(x)\, y + Q(x)\, y^r = 0$$

où r est un réel quelconque, P et Q deux fonctions définies et continues sur un intervalle I. Le changement de variable $u = y^{1-r}$ permet d'obtenir une équation linéaire.

Équation de Lagrange et Clairaut On appelle une équation de Lagrange et Clairaut toute équation différentielle de la forme :

$$y = x\, f(y') + g(y')$$

où f et g sont deux fonctions définies et différentiables sur un intervalle I.

Équation aux différentielles totales Une équation différentielle de la forme

$$a(x,y)\, dx + b(x,y)\, dy = 0$$

est une équation aux différentielles totales si $\frac{\partial a}{\partial y} = \frac{\partial b}{\partial x} \quad \forall x, y$. Il existe alors une fonction $W(x,y)$ telle que $dW(x,y) = a(x,y)\, dx + b(x,y)\, dy = 0$ définissant une intégrale générale. Le calcul de W s'effectue en deux quadratures.

7.2 Équations différentielles du second ordre

Une équation différentielle ordinaire (EDO) du second ordre est une relation entre les dérivés première et seconde d'une fonction inconnue d'une variable réelle ou complexe et la valeur de la fonction.

> **Définition :** *Soit y une fonction inconnue d'une variable x, y' et $y"$ ses dérivées première et seconde, soit F une fonction quelconque alors la définition formelle d'une équation différentielle du premier ordre est*
>
> $$F\left(y, y', y", x\right) = 0.$$
>
> *Lorsque cette équation est résoluble en $y"$, on peut la mettre sous la forme*
>
> $$y" = f\left(y, y', x\right).$$

7.2.1 Quelques types d'équations différentielles du second ordre se ramenant à des équations du premier ordre.

$F(x, y', y") = 0$ se ramène à l'équation du premier ordre $F(x, z, z')$ en posant $z(x) = y'(x)$. Si la solution générale z peut être obtenue, la solution y s'obtient en une quadrature

$$y(x) = \int z(x)dx + C$$

avec C une constante d'intégration.

$F(y, y', y") = 0$ se ramène à l'équation du premier ordre $F(y, p(y), p'(y)p(y))$ en posant $y'(x) = p(y)$. Si la solution générale p peut être obtenue, la solution y s'obtient en résolvant

$$y'(x) = p(y)$$

qui est à variables séparables.

$y" = f(y)$ S'intègre en deux temps par la multiplication membre à membre par y'. En notant $F(y) = 2\int f(y)dy$ on obtient une intégrale première

$$y'^2 = F(y) + \lambda$$

dont l'intégration est immédiate $x = \int \pm \frac{dy}{\sqrt{F(y)+\lambda}} + \mu$ avec λ et μ deux constantes d'intégration.

7.2.2 Équation différentielle linéaire du second ordre

Les équations différentielles linéaires d'ordre deux sont des équations différentielles de la forme :

$$a(x)\,y'' + b(x)\,y' + c(x)\,y = d(x)$$

où a, b, c et d sont des fonctions continues. Toutes ne peuvent être résolues explicitement, cependant beaucoup de méthodes existent pour résoudre celles qui peuvent l'être, ou pour faire l'étude qualitative des solutions à défaut. Parmi les plus simples à résoudre sont les équations à coefficients constants (où a, b, c sont des constantes). Le qualificatif de linéaire indique qu'il est possible d'appliquer des procédés de superposition de solutions, et d'exploiter des résultats d'algèbre linéaire. Un rôle particulier est dévolu aux équations différentielles homogènes (où d = 0).

Équation différentielle homogène à coefficients constants de la forme $ay'' + by' + cy = 0$ où a, b et c sont des réels, a non nul.

On cherche des solutions sous forme exponentielle, c'est-à-dire telles que $y(x) = e^{\lambda x}$. Une telle fonction sera solution de l'équation différentielle si et seulement si λ est solution de $a\lambda^2 + b\lambda + c = 0$. Cette équation est appelée équation caractéristique de l'équation différentielle.

Comme pour toute équation du second degré, trois cas se présentent selon le signe du discriminant Δ.

1. Si $\Delta > 0$ L'équation possède deux solutions λ_1 et λ_2. L'équation possède au moins deux fonctions exponentielles solutions $y_1(x) = e^{\lambda_1 x}$ et $y_2(x) = e^{\lambda_2 x}$. On démontre que ces deux solutions engendrent l'ensemble des solutions. C'est-à-dire que l'ensemble des solutions sont les fonctions définies sur $\mathbb{R}$ par $y(x) = C_1 y_1(x) + C_2 y_2(x)$ où C_1 et C_2 sont deux réels quelconques.
2. Si $\Delta = 0$: L'équation ne possède qu'une seule solution λ. On démontre alors que l'ensemble des solutions sont les fonctions y définies sur $\mathbb{R}$ par $y(x) = (Ax + B)e^{\lambda x}$ où A et B sont des réels quelconques.
3. Si $\Delta < 0$: L'équation ne possède pas de solutions réelles, mais deux solutions complexes conjuguées que l'on peut écrire $\Lambda_\pm = \rho \pm \omega$ avec ρ et ω réels. L'ensemble des solutions est formé des fonctions définies sur $\mathbb{R}$ par $y(x) = e^{\rho x}(A\cos(\omega x) + B\sin(\omega x))$, où A et B sont deux réels quelconques. On peut aussi écrire cette solution sous la forme : $y(x) = \alpha e^{\rho x}\cos(\omega x + \phi)$ avec α et ϕ deux réels quelconques.

La détermination de A et B (ou α et ϕ) se fait par la donnée de deux informations sur y.

Équation différentielle à coefficients constants de la forme $ay'' + by' + cy = d(x)$ où a, b, et c sont des réels, a non nul et $d(x)$ une fonction donnée.

L'équation obtenue en remplaçant d par la fonction nulle est appelée équation homogène associée à l'équation différentielle ; on la suppose résolue. Il suffit alors de trouver une solution de l'équation avec second membre : y_0, pour les connaître toutes. En effet, les solutions de l'équation différentielle sont les fonctions $y_0 + g$ où g est une solution générale de l'équation homogène associée.

Si le second membre d est la somme de plusieurs fonctions d_1 et d_2 : $ay'' + by' + cy = d_1 + d_2$, on peut chercher une solution particulière $s1$ de l'équation différentielle de second membre d_1 : $ay'' + by' + cy = d_1$, puis une solution particulière $s2$ de l'équation différentielle de second membre d_2 : $ay'' + by' + cy = d_2$. La somme de ces deux solutions particulières $s = s1 + s2$ est solution particulière de l'équation de départ.

Si d est une fonction polynôme ou trigonométrique, on cherchera alors une solution particulière en suivant les règles suivantes :

— Si d est un polynôme de degré n alors l'équation admet une solution particulière de la forme $y : t \mapsto t^p Q(t)$ où Q est un polynôme de degré n, et p peut prendre trois valeurs :

1. si 0 n'est pas racine de l'équation caractéristique alors $p = 0$
2. si 0 est racine simple de l'équation caractéristique alors $p = 1$
3. si 0 est racine double de l'équation caractéristique alors $p = 2$

— Si d est de la forme $d : t \mapsto e^{mt}P(t)$ où P est un polynôme de degré n et m un complexe alors on cherchera une solution particulière de l'équation différentielle de la forme $y : t \mapsto t^p e^{mt}Q(t)$ où Q est un polynôme de degré n et p prend trois valeurs :

1. si m n'est pas racine de l'équation caractéristique alors $p = 0$
2. si m est racine simple de l'équation caractéristique alors $p = 1$
3. si m est racine double de l'équation caractéristique alors $p = 2$

— si $d(x) = A\cos(\omega x + \phi) + B\sin(\omega x + \phi)$ une cherche une solution particulière sous la forme d'une combinaison linéaire de $\cos(\omega x + \phi)$ et $\sin(\omega x + \phi)$.

7.2.3 Méthode de Frobenius

La méthode Frobenius est une façon de trouver une solution en série infinie pour une équation différentielle ordinaire du deuxième ordre de la forme

$$x^2 \frac{d^2 y}{d x^2} + p(x)\, x^2 \frac{d y}{d x} + q(x)\, y = 0$$

aux alentours du point singulier régulier, $x = 0$. La méthode est inutile si $p(x)/x$ et $q(x)/x^2$ sont analytique en $x = 0$. Un développement en série entière standard sera suffisant dans ce cas.

La recherche de solution utilise un développement en série de puissance de la forme particulière

$$y(x) = x^r \sum_{k=0} \infty A_k\, x^k, \qquad A_0 \neq 0.$$

Trouver une solution de l'équation différentielle consiste à déterminer l'ensemble des coefficients A_k et vérifier la convergence de la série.

Après le calcul des dérivées,

$$y'(x) = \sum_{k=0}^{\infty} (k + r) A_k\, x^{k+r-1}, \qquad y''(x) = \sum_{k=0}^{\infty} (k + r - 1)(k + r) A_k\, x^{k+r-2},$$

la substitution dans l'équation différentielle conduit à :

$$0 = \left[r(r-1) + p(x)r + q(x)\right] A_0\, x^r + \sum_{k=1}^{\infty} \left[(k + r - 1)(k + r) + p(x)(k + r) + q(x)\right] A_k\, x^{k+r}.$$

L'égalité à zéro implique la nullité de tous les coefficients des monômes z^{r+k} du membre de droite. L'expression quadratique de r

$$r\,(r-1) + p\,(0)\, r + q\,(0) = I(r)$$

est appelé "polynôme indiciel", c'est le coefficient de la plus petite puissance de x. L'expression générale du coefficient de x^{r+k} correspond à

$$I(k + r)\ A_k + \sum_{j=0}^{k-1} \left[(j + r)p(k - j) + q(k - j)\right]\ A_j,$$

La nullité de ces coefficients implique une relation de récurrence entre les coefficients A_k

$$\frac{1}{-I(k + r)} \sum_{j=0}^{k-1} \left[(j + r)p(k - j) + q(k - j)\right] A_j = A_k.$$

Les fonctions définies en série

$$U_r(x) = x^r \sum_{k=0} \infty\ A_k\, x^k$$

sont solution de l'équation

$$x^2 U_r(x)'' + p(x) x U_r(x)' + q(x) U_r(x) = I(r) x^r.$$

En choisissant pour r une racine du polynôme indiciel, on obtient une solution de l'équation différentielle.

Exemple : L'application de la méthode de Frobenius à l'équation différentielle

$$x^2 f'' - x f' + (1 - x) f = 0,$$

conduit à la relation

$$0 = (r(r-1) - r + 1) A_0 x^{r-2} + \sum_{k=1}^{\infty} \left(((k+r)(k+r-1) - (k+r) + 1) A_k - A_{k-1} \right) x^{k+r-2}.$$

Le polynôme indiciel $I(r) = r(r-1) - r + 1$ à une racine double $r = 1$, ce qui conduit à la relation de récurrence

$$A_k = \frac{A_{k-1}}{k^2}.$$

En posant $A_0 = 1$, on obtient $A_k = \frac{1}{(k!)^2}$ ce qui garantie la convergence de la série dans $\mathbb{R}$.

7.2.4 Théorie de Sturm-Liouville

La théorie de Sturm-Liouville étudie le cas particulier des équations différentielles linéaires de la forme

$$-\frac{d}{dx}\left[p(x)\frac{dy}{dx}\right] + q(x) y = \lambda w(x) y,$$

dans laquelle le paramètre λ fait partie comme la fonction y des inconnues.

Un problème de Sturm–Liouville (S–L) est dit régulier si $p(x) > 0$, $w(x) > 0$, et $p(x)$, $p'(x)$, $q(x)$, et $w(x)$ sont des fonctions continues sur un intervalle fini $[a, b]$, ayant des conditions aux bords de la forme :

$$\begin{cases} \alpha_1 y(a) + \alpha_2 y'(a) = 0 & (\alpha_1^2 + \alpha_2^2 > 0), \\ \beta_1 y(b) + \beta_2 y'(b) = 0 & (\beta_1^2 + \beta_2^2 > 0), \end{cases}$$

Les principaux résultats sont les suivants :

1. Les valeurs propres $\lambda_1, \lambda_2, \lambda_3, \ldots$ sont réelles et ordonnées telles que

$$\lambda_1 < \lambda_2 < \lambda_3 < \cdots < \lambda_n < \cdots \to \infty;$$

2. A chaque valeur propre λ_n correspond une unique fonction propre $y_n(x)$ (à une constante de normalisation près. Les fonctions $y_n(x)$ ont exactement $n - 1$ zéros dans l'intervalle $[a, b]$. La fonction propre $y_n(x)$ est appelée la n–ième solution du problème de Stourm–Liouville régulier.
3. Les fonctions propres normalisées forment une base orthonormée.

$$\int_a^b y_n(x) y_m(x) w(x) \, \mathrm{d}x = \delta_{mn},$$

dans l'espace de Hilbert.

Par multiplication par un facteur intégrant convenable, toute équation différentielle linéaire d'ordre deux peut être mise sous la forme d'une équation de Sturm-Liouville.

Exemple : L'équation $-\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} = \lambda u$ avec u et λ inconnus et pour conditions aux bords $u(0) = u(\pi) = 0$ admet pour la famille $\{u_n(x) = \sin(nx), \quad \lambda_n = -n^2$ comme base orthogonale de solutions.

7.3 L'équation de Bessel

Les solutions de l'équation de Bessel font apparaître de nouvelles fonctions mathématiques, les fonctions de Bessel, que l'on introduit dans ce paragraphe. Considérons l'équation différentielle de Bessel

$$x^2y"(x) + xy'(x) + \left(x^2 - n^2\right) y(x) = 0$$

avec n un entier non–négatif. Une particularité de l'équation de Bessel est que le coefficient de $y"$ s'annule en un point où on s'intéresse à la solution (on dit que l'on a une équation différentielle avec une singularité).

La recherche de solution de l'équation de Bessel utilise la méthode de Frobenius. En posant

$$y(x) = \sum_{j=0}^{\infty} c_j x^{j+\alpha} \quad \text{avec} \quad c_0 \neq 0,$$

on obtient

$$\sum_{j=0}^{\infty} c_j(j+\alpha)(j+\alpha-1)x^{j+\alpha} + \sum_{j=0}^{\infty} c_j(j+\alpha)x^{j+\alpha} + (x^2-n^2)\sum_{j=0}^{\infty} c_j x^{j+\alpha} = 0.$$

Une comparaison des coefficients donne

$$\begin{gathered} c_0\left(\alpha^2 - n^2\right) = 0 \\ c_1\left((1+\alpha)^2 - n^2\right) = 0 \\ c_j\left((j+\alpha)^2 - n^2\right) + c_{j-2} = 0, \quad \text{pour} \quad j \geq 2. \end{gathered}$$

Comme $c_0 \neq 0$, la première équation implique $\alpha^2 = n^2$, ce qui fixe la valeur de α. La possibilité $\alpha = -n$ (pour $n > 0$) est moins intéressante, car elle implique des solutions $y(x)$ possédant une singularité en $x = 0$ appelée fonction de Bessel de deuxième espèce. On poursuit donc le calcul avec l'autre possibilité $\alpha = n$. On trouve pour $c_1 = 0$ et une relation de récurrence entre les c_j

$$c_j\ j(j+2n) + c_{j-2} = 0, \quad \text{pour} \quad j \geq 2.$$

Le fait que $c_1 = 0$ implique que les coefficients impairs sont tous nuls. En posant $j = 2k$ on trouve le résultat

$$c_{2k} = \frac{(-1)^k}{4^k k! \Pi_{i=1}^{k}(i+n)} c_0.$$

Avec le choix $c_0 = 1/(2^n n!)$, la fonction $y(x)$ devient

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

Cette fonction définie par une série entière s'appelle la **fonction de Bessel** d'indice n. Le critère du quotient

$$\left|\frac{c_{2k}}{c_{2(k+1)}}\right| = 4(k+1)(k+1+n) \to \infty$$

montre que la série converge pour tout $x \in \mathbb{R}$.

Symétrie de fonction de Bessel

$$J_{-n}(x) = (-1)^n J_n(x)$$

$$J_n(-x) = (-1)^n J_n(x)$$

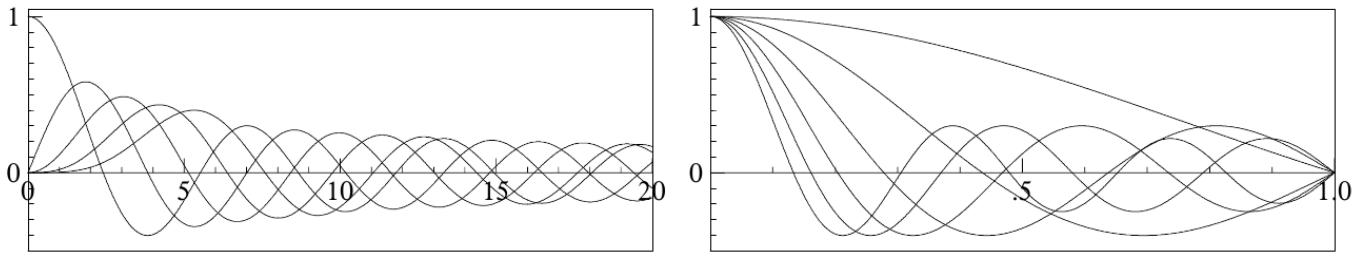


FIGURE 7.1 – Fonctions de Bessel $J_n(x)$ (gauche), $J_0(j_{0,k}x)$ (droite)

Relations de récurrence

On démontre les relations, très utiles, suivantes

$$J_{n+1}(x) = \frac{nJ_n(x)}{x} - J'_n(x)$$
$$J_{n+1}(x) + J_{n-1}(x) = \frac{2n}{x}J_n(x)$$
$$J_{n+1}(x) - J_{n-1}(x) = -2J'_n(x)$$

Fonction génératrice

Les fonctions de Bessel sont parfois définies par l'intermédiaire d'une série de Laurent, correspondant à la fonction génératrice :

$$e^{(x/2)(z-1/z)} = \sum_{n=-\infty}^{\infty} J_n(x)z^n, \quad \text{avec} z \in \mathbb{C}$$

Deux cas particuliers importants sont obtenus en posant $x = e^{i\phi}$

$$e^{ix\sin\phi} = \sum_{n=-\infty}^{\infty} J_n(x)e^{in\phi};$$

et $x = ie^{i\phi}$

$$e^{iz\cos\phi} = \sum_{n=-\infty}^{\infty} i^n J_n(z)e^{in\phi}.$$

Ainsi, les fonctions de Bessel permettent d'obtenir les coefficients de Fourrier pour une modulation de fréquence.

Intégrales de Bessel

Pour les valeurs entières de $\alpha = n$, les fonctions de Bessel peuvent être représentées par des intégrales :

$$J_n(x) = \frac{1}{\pi}\int_0^{\pi} \cos(n\tau - x\sin\tau)\,\mathrm{d}\tau.$$

ou encore par :

$$J_n(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-\mathrm{i}\,(n\tau - x\sin\tau)}\,\mathrm{d}\tau.$$

Relation avec les polynômes de Laguerre

Notant L_k le k–ème polynôme de Laguerre, les fonctions de Bessel peuvent être exprimées ainsi :

$$\frac{J_\alpha(x)}{\left(\frac{x}{2}\right)^\alpha} = \frac{e^{-t}}{\alpha!}\sum_{k=0}^{\infty} \frac{L_k^{(\alpha)}\left(\frac{x^2}{4t}\right)}{\binom{k+\alpha}{k}}\frac{t^k}{k!},$$

où l'expression de droite ne dépend pas de t.

Comportement asymptotique

Pour les grandes valeurs de l'argument, les fonctions de Bessel se comportent comme des cosinus décroissants

$$J_n(x) \to \sqrt{2}\pi x \cos\left(x - \frac{1}{2}n\pi - \frac{1}{4}\pi\right) \quad \text{quand} \quad |x| \to \infty.$$

Zéros réels des fonctions de Bessel

Pour n entier, les fonctions de Bessel on une infinité de zéros, c–à–d. l'équation

$$J_n(x) = 0$$

a une infinité de solutions dans $\mathbb{R}$. Comme les fonctions de Bessel sont paires ou impaires, on numérote les zéros à partir des valeurs positives et on note $j_{n,s}$ le s–ème zéro positif ou nul de la n–ème fonction de Bessel *i.e.*

$$J_n(j_{n,s}) = 0, \qquad \forall s.$$

On montre que les zéros des différentes fonctions de Bessel sont entrelacés :

$$j_{n,1} < j_{n+1,1} < j_{n,2} < j_{n+1,2} < j_{n,3} \dots$$

Théorème d'orthogonalité des fonctions de Bessel : *Soient $j_{n,1} < j_{n,2} < j_{n,3} < \dots$ les zéros positifs de $J_n(x)$, alors*

$$\int_0^1 J_n\left(j_{n,k}\, r\right) J_n\left(j_{n,l}\, r\right) r dr = \begin{cases} 0 & \text{pour} \quad k \neq l \\ \frac{1}{2}\left(J_n'(j_{n,k})\right)^2 & \text{pour} \quad k = l \end{cases}$$

Les fonctions de Bessel $f_k(r) = J_n\left(j_{n,k}\, r\right)$ sont orthogonales sur l"intervalle $[0,1]$ par rapport à la fonction de poids $\omega(r) = r$.

Les séries de Fourier–Bessel

Une conséquence du théorème d'orthogonalité des fonctions de Bessel est le développement d'une fonction $f(x)$ en série de Fourier-Bessel

Définition : *Soit $f(x)$ une fonction à valeurs réelles définie sur un intervalle $[0,b]$. On nome série de Fourier–Bessel le développement de f suivant*

$$f(x) = \sum_{k=0}^{\infty} c_k J_n\left(j_{n,k}\frac{x}{b}\right).$$

Les coefficients sont alors

$$c_k = \frac{2}{\left(d\, J_n'(j_{n,l})\right)^2} \int_0^1 f(x) J_n\left(j_{n,k} r\right) x\, dx$$

7.4 Système dynamique différentiel

Depuis les travaux d'Isaac Newton (1687), l'idée que l'évolution temporelle d'un système physique quelconque est bien modélisée par une équation différentielle (ou ses généralisations à la théorie des champs, les équations aux dérivées partielles) est admise. Cette modélisation différentielle s'est depuis étendue avec succès à d'autres disciplines comme la chimie, la biologie, l'économie, ...On considère typiquement un système différentiel du premier ordre du type :

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = f(x(t), t)$$

où la fonction f définit le système dynamique étudié (pour un système à n degrés de liberté, il s'agit à proprement parler d'un champ de vecteurs à n dimensions, c'est-à-dire, d'un point de vue prosaïque, un ensemble de n fonctions scalaires).

Distinguons les systèmes dynamiques linéaires des systèmes dynamiques non linéaires. Les solutions d'un système dynamique linéaire forment un espace vectoriel, ce qui permet l'utilisation de l'algèbre linéaire et simplifie considérablement l'analyse. La transformée de Laplace permet de transformer les équations différentielles en des équations algébriques simplifiant considérablement la résolution.

Des **systèmes dynamiques non linéaires** peuvent faire preuve de comportements complètement imprévisibles, qui peuvent même sembler aléatoires (alors qu'il s'agit de systèmes parfaitement déterministes). Cette imprédictibilité est appelée chaos. La branche des systèmes dynamiques qui s'attache à définir clairement et à étudier le chaos s'appelle la théorie du chaos. Elle décrit qualitativement les comportements à long terme des systèmes dynamiques. Dans ce cadre, on ne met pas l'accent sur la recherche de solutions précises aux équations du système dynamique (ce qui, de toute façon, est souvent sans espoir), mais plutôt sur la réponse à des questions comme « Le système convergera-t-il vers un état stationnaire à long terme, et dans ce cas, quels sont les états stationnaires possibles ? » ou « Le comportement à long terme du système dépend-il des conditions initiales ? ».

Un objectif important est la description des points fixes, ou états stationnaires, du système ; ce sont les valeurs de la variable pour lesquelles elle n'évolue plus avec le temps. Certains de ces points fixes sont attractifs, ce qui veut dire que si le système parvient à leur voisinage, il va converger vers le point fixe.

> **Point fixe :** *On nomme points fixes, ou points stationnaires, d'un système dynamique différentiel*
>
> $$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = f(x(t), t),$$
>
> *l'ensemble des points x^* tels que :*
>
> $$f(x^*, t) = 0 \quad \forall t$$

Un système dynamique ayant un de ses points fixes x^* comme condition initiale a donc une évolution temporelle triviale : $\forall t > 0 \quad x(t) = x^*$. La question est alors de déterminer l'évolution temporelle des conditions initiales voisines d'un point fixe

— Si au cours du temps la solution "se rapproche" du point fixe, celui-ci est dit stable
— Si au cours du temps la solution "s'éloigne" du point fixe, celui-ci est dit instable

La stabilité d'un point fixe x^* s'étudie par une **analyse de stabilité linéaire**. Pour cela

1. on pose le changement de variable $x(t) = x^* + \varepsilon u(t)$ avec $\varepsilon \ll 1$ et u bornée pour ce placer au voisinage du point fixe.
2. on "linéarise" l'équation différentielle en effectuant un développement de Taylor de f en fonction du petit paramètre ε.
3. La résolution des l'équation linéarisée détermine la stabilité
   — si $u \to 0$ quand $t \to \infty$, la solution au voisinage du point fixe tends vers celui–ci, l'approximation linéaire est justifiée et le point fixe est stable
   — si $|u| \to \infty$ quand $t \to \infty$, la solution au voisinage du point fixe s'éloigne de celui–ci, l'approximation linéaire n'est plus justifié et le point fixe est instable

L'idée maîtresse des systèmes dynamiques est l'étude des changements de comportement (bifurcation) en fonction de la variation d'un paramètre physique.

7.5 Exercices

7.5.1 Résolution des équations différentielles

1. Donner l'ensemble des solutions de l'équation différentielle suivante

- $\frac{dN}{dt} + \delta N = 0$
- $\frac{dN}{dt} - \delta N = K$
- $\frac{dN}{dt} - \delta N = \exp(t/\tau)$
- $\frac{dN}{dt} + \delta N = \sin(\omega t)$
- $\frac{dN}{dt} + \ln(t/\tau)N = 0$
- $\frac{dN}{dt} - \sin(\omega t) = 0$
- $\frac{dy}{dt} = \sin(\omega t)y$
- $\frac{dy}{dx} = y\sin(x)$
- $\frac{dy}{dt} = \cos(\omega t)y$
- $\frac{dy}{dt} = \sin(\omega t)y^2$
- $\frac{dy}{dt} = \cos(\omega t)(1+y^2)$
- $\frac{dy}{dt} = \cos(y)^2 t$
- $\frac{dy}{dt} = \exp(-\lambda t)y$
- $\frac{dy}{dt}\exp(-y) = t$
- $\frac{dy}{dt} = \exp(y)t$
- $\frac{dy}{dt} = a^y t$
- $\frac{dy}{dt}\sin(y^2)y = t$
- $\frac{d^2z}{dt^2} = -GM_T\frac{1}{z^2}$
- $\frac{ld^2\theta}{dt^2} = g\sin(\theta)$
- $m\frac{d^2z}{dt^2} + \omega_0^2 z = 0$
- $m\frac{d^2z}{dt^2} - \omega_0^2 z = 0$
- $\frac{d^2z}{dt^2} + \omega_0^2 z = F_0\cos(\omega t)$
- $\frac{d^2z}{dt^2} + 2\gamma\frac{dz}{dt} + \omega_0^2 z = 0$
- $\frac{d^2z}{dt^2} + 2\gamma\frac{dz}{dt} + \omega_0^2 z = F_0\exp(-t/\tau)$
- $\frac{d^2z}{dt^2} + 2\gamma\frac{dz}{dt} + \omega_0^2 z = F_0\cos(\omega t)$
- $\frac{d^2z}{dt^2} + 2\gamma\frac{dz}{dt} + \omega_0^2 z = F_0\cos(\omega t)\exp(-t/\tau)$
- $\frac{d^2z}{dt^2} + 2\gamma\frac{dz}{dt} + \omega_0^2 z = F_0 t^2\exp(-t/\tau)$
- $\frac{d^2z}{dt^2} + 2\gamma\frac{dz}{dt} + \omega_0^2 z = F_0\left(1+\frac{t}{\tau}\right)\cos(\omega t)$

2. Donner l'ensemble des solutions de l'équation différentielle suivante $\frac{dN(t)}{dt} + \frac{N(t)}{\tau} = f(t)$ en utilisant $F(t) = \int_0^t f(u)\exp(u/\tau)du$.
3. Donner l'ensemble des solutions de l'équation suivante : $\left[\frac{dy(t)}{dt}\right]^2 = (1+y^2)^2$.
4. Donner l'ensemble des solutions de l'équation suivante : $\frac{dx(t)}{dt} = \sqrt{y+t}$.
5. Donner en fonction des paramètres $(m,k) \in \mathbb{R}^{*+}$ les solutions de l'équation $m\frac{d^2y(t)}{dt^2} + ky(t) = 0$ vérifiant $y(0) = 0$ et $y(L) = 0$.

7.5.2 Équations de Bernoulli

Soit $P(x)$ et $Q(x)$ deux polynômes et r un réel, on appelle équations de Bernoulli les équations de la forme :

$$\frac{dy}{dx} + P(x)y + Q(x)y^r = 0$$

1. Quelles sont les valeurs de r pour lesquels l'intégration utilise les techniques vues précédemment ?
2. A quelle condition peut–on diviser chaque membre de l'équation par y^r ?
3. Montrer que dans ce cas il existe un changement de variable permettant de résoudre l'équation en utilisant les techniques vues précédemment.
4. Appliquer cette méthode pour $P(x) = 1$, $Q(x) = -1$ et $r = 2$.

Application des équations de Bernoulli à l'oscillateur amorti Voici l'équation du mouvement d'un oscillateur harmonique dont l'amortissement est proportionnel au carré de la vitesse :

$$\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} + \omega_0^2 x = 0$$

1. Montrer que si l'on pose $u = \frac{dy}{dt}$, l'équation du mouvement se ramène à une équation différentielle du premier ordre par rapport à la variable x.
2. Montrer que cette équation est une équation de Bernoulli et donner les solutions $u(y)$
3. Trouver la trajectoire si les conditions initiales sont $x(0) = \frac{1}{2\gamma}$ et $x'(0) = 0$.

7.5.3 Analyse de stabilité linéaire

Exercice : Analyser les bifurcations des systèmes suivant en recherchant les points fixes et an étudiant leur stabilité en fonction des paramètres,

1. bifurcation Selle–nœud : $\dot{x} = \mu - x^2$
2. bifurcation transcritique $\dot{x} = \mu x - x^2$

3. bifurcation fourche super critique $\dot{x} = \alpha x(\mu - x^2)$
4. bifurcation fourche sous critique $\dot{x} = \alpha x(\mu + x^2)$
5. Modele proie–prédateur $\begin{cases} \frac{du}{dt} = & ku - uv \\ \frac{dv}{dt} = & -\gamma v + \gamma uv \end{cases}$

Exercice : On veut étudier le comportement du système dynamique

$$\begin{aligned} \frac{dx}{dt} &= -\omega y + \alpha x\left(\rho^2 - x^2 - y^2\right) \\ \frac{dy}{dt} &= \omega x + \alpha y\left(\rho^2 - x^2 - y^2\right) \end{aligned}$$

avec $x(t)$ et $y(t)$ des variables réelles.

1. Effectuer le changement de variable $\{x(t), y(t)\} = \{r(t)\cos\theta(t), r(t)\sin\theta(t)\}$. Montrer que les équations pour les nouvelles variables $\{r(t), \theta(t)\}$ sont de la forme :

$$\begin{aligned} \frac{dr}{dt} &= \alpha r(\rho^2 - r^2) \\ \frac{d\theta}{dt} &= \omega \end{aligned}$$

2. Résoudre l'équation sur θ.
3. Déterminer pour la variable $r(t)$ les points fixes r^*.
4. Étudier la stabilité des points fixes r^* en fonction des paramètres.

7.5.4 Analyse de stabilité (Fevrier 2011)

On veut étudier de manière qualitative l'évolution temporelle de la variable $x(t)$ lorsqu'elle est déterminée par l'équation différentielle

$$\frac{dx}{dt} = \alpha - x^2$$

où α est un paramètre positif supposé constant dans un premier temps.

Recherche du point fixe :

3 - Montrer que cette équation admet une solution constante $x(t) = x_0$ à déterminer ; x_0 est appelé le point fixe de l'équation différentielle et il dépend de la valeur du paramètre α.

Étude de la stabilité du point fixe :
Pour déterminer qualitativement l'évolution temporelle de $x(t)$, on souhaite étudier la dynamique au voisinage du point fixe x_0. Pour cela on pose le changement de variable $x(t) = x_0 + \varepsilon x_1(t)$.

4 - Déterminer l'équation différentielle régissant l'évolution temporelle de la nouvelle variable $x_1(t)$.

5 - Linéariser cette équation différentielle lorsque $|\varepsilon| \ll |x_0|$.

6 - Montrez que le point fixe est stable en résolvant l'équation linéarisé.

7 - Déterminer le temps caractéristique τ de retour au point fixe défini comme $x_1(\tau) = 0,1x_1(0)$.

8 - Représenter l'évolution temporelle qualitative des solutions de l'équation $\frac{dx}{dt} = \alpha - x^2$ ayant pour condition initiale $x(0) = 0,1x_0$ et $x(0) = 10x_0$.

Évolution adiabatique :
On suppose que la variable x est initialement au point fixe $x(0) = x_0$ et qu'un opérateur externe fait lentement varier le paramètre α en fonction du temps : $\alpha(t) = \alpha_0 + \mu t$.

9 - Justifier qualitativement que lorsque $|\mu\tau| \ll 1$ la variable x continuellement être égale à la valeur du point fixe correspondant à la valeur instantanée de $\alpha(t)$.

7.5.5 Analyse de stabilité (Novembre 2010)

On veut étudier de manière qualitative l'évolution temporelle de la variable $x(t)$ lorsqu'elle est déterminée par l'équation différentielle

$$\frac{dx}{dt} = \alpha - x^3$$

où α est un paramètre positif supposé constant.

Recherche du point fixe : Montrer que cette équation admet une solution constante $x(t) = x_0$ à déterminer ; x_0 est appelé le point fixe de l'équation différentielle et il dépend de la valeur du paramètre α.

Étude de la stabilité du point fixe : Pour déterminer qualitativement l'évolution temporelle de $x(t)$, on souhaite étudier la dynamique au voisinage du point fixe x_0. Pour cela on pose le changement de variable $x(t) = x_0 + \varepsilon x_1(t)$.

1. Déterminer l'équation différentielle régissant l'évolution temporelle de la nouvelle variable $x_1(t)$.
2. Linéariser cette équation différentielle lorsque $|\varepsilon| \ll |x_0|$.
3. Montrez que le point fixe est stable en résolvant l'équation linéarisée.

Chapitre 8

Équations aux dérivées partielles

Sommaire

Contrairement aux équations différentielles ordinaires, les équations aux dérives partielles ont comme inconnue une fonction de plusieurs variables et l'équation contient des dérivées partielles. Les séries de Fourier ont été les premiers outils pour leurs solutions. Dans ce chapitre, on focalise sur la méthode résolution des équations et non sur leurs établissements.

8.1 Méthode de séparation des varibles sur deux exemples simples

8.1.1 Équation des ondes — corde vibrante

Initié par Taylor (1713) et John Bernoulli (1728), le problème de la corde vibrante fut l'un des principaux champs de recherche et de disputes (d'Alembert, Clairaut, Euler, Danièle Bernoulli, Lagrange) au XVIIIe siècle. Il représente la première EDP (équation aux dérivées partielles) de l'histoire. On s'intéresse au problème suivant : trouver la l'évolution temporelle de la déformation d'une corde élastique tendue de longueur L et d'extrémités fixes ; connaissant à l'instant initial la déformation de la corde et la vitesse en chaque point de la corde. En notant $u(x,t)$ la déformation de la corde à la position x et à l'instant t, le problème se modélise de la façon suivante :

$$\begin{aligned} \frac{\partial^2 u(x,t)}{\partial t^2} &= c^2 \frac{\partial^2 u(x,t)}{\partial x^2} && \text{pour } 0 < x < L,\ t \geq 0 \\ u(x,0) = f(x),\ \frac{\partial u}{\partial x}(x,0) &= g(x) && \text{pour } 0 < x < L \text{ (conditions initiales)} \\ u(0,t) = 0,\ u(L,t) &= 0 && \text{pour } t \geq 0 \text{ (conditions aux bords)} \end{aligned}$$

Méthode de séparation des variables

Cette méthode, adoptée par Fourier à de nombreuses reprises, souvent appelée méthode de Fourier, est appliquée pour la première fois par d'Alembert à la corde vibrante en 1750. Elle consiste à rechercher la solution $u(x,t)$ *comme le produit d'une fonction* $X(x)$*, ne dépendant que de* x*, et d'une fonction* $T(t)$*, ne dépendant que de* t*,* c–à–d.

$$u(x,t) = X(x) \cdot T(t).$$

Ainsi l'équation de la corde devient

$$X(x)\,\frac{d^2T(t)}{dt^2} = c^2 \frac{d^2X(x)}{dx^2}\,T(t).$$

La deuxième idée consiste à diviser cette relation par $X(x)\cdot T(t)$ pour obtenir

$$\frac{1}{c^2}\,\frac{1}{T(t)}\frac{d^2T(t)}{dt^2} = \frac{1}{X(x)}\frac{d^2X(x)}{dx^2} = -\lambda,$$

où la conclusion est la suivante : *ce qui est à gauche, ne dépend pas de* x*, ce qui est à droite, ne dépend pas de* t*, donc le tout doit être une constante, que l'on appelle* $-\lambda$. On a obtenu deux équations différentielles ordinaires, que l'on traite l'une après l'autre ; commençons par

$$\frac{dX(x)}{dx^2} + \lambda X(x) = 0 \quad \Rightarrow \quad X(x) = A\cdot\cos(\sqrt{\lambda}x) + B\cdot\sin(\sqrt{\lambda}x).$$

Les deux conditions aux bords donnent $A = 0$ pour $x = 0$ et $\sin(\sqrt{\lambda}L) = 0$ pour $x = L$. Ainsi il faut que $\sqrt{\lambda}L = k\pi$, et on obtient pour la fonction $X(x)$,

$$\lambda = \lambda_k = \left(\frac{k\pi}{L}\right)^2, \quad X(x) = \sin(\frac{k\pi x}{L}), \quad k = 1,2,3,\ldots$$

L'autre équation donne

$$\frac{dT(t)}{dt^2} + c^2\lambda T(t) = 0 \quad \Rightarrow \quad T(t) = C\cdot\cos\left(c\sqrt{\lambda_k}\,t\right) + D\cdot\sin\left(c\sqrt{\lambda_k}\,t\right).$$

On obtient ainsi, pour chaque valeur de $k = 1,2,3,\ldots$, une solution. Toutes ces solutions peuvent être additionnées, car notre équation est **linéaire**. Ainsi la solution générale devient par superposition

$$u(x,t) = \sum_{k\geq 1} \sin\left(k\pi\frac{x}{L}\right)\left(C_k\cdot\cos\left(k\pi\frac{ct}{L}\right) + D_k\cdot\sin\left(k\pi\frac{ct}{L}\right)\right)$$

Pour déterminer les coefficients C_k et D_k, il reste à utiliser les conditions initiales. Ainsi en posant $t = 0$, on obtient

$$f(x) = \sum_{k\geq 1} C_k \sin\left(k\pi\frac{x}{L}\right) \;\Rightarrow\; C_k = \frac{2}{L}\int_0^L f(x)\,\sin\left(k\pi\frac{x}{L}\right)dx$$

$$g(x) = \sum_{k\geq 1} \frac{k\pi c}{L} D_k \sin\left(k\pi\frac{x}{L}\right) \;\Rightarrow\; D_k = \frac{2}{ak\pi}\int_0^L g(x)\,\sin\left(k\pi\frac{x}{L}\right)dx$$

Synthèse de la méthode

1. Séparation des variables
2. Résolution des équations séparées : on obtient toutes **les** solutions satisfaisants aux conditions aux bords
3. Utilisation des séries de Fourier pour trouver **la** solution satisfaisant aux conditions initiales.

Exercice : **Corde frappée** Déterminer la solution de la corde vibrante si $f(x) = 0$ et $g(x) = \alpha$ si $\frac{L}{n} - d \leq x \leq \frac{L}{n} + d$ et $g(x) = 0$ sinon.

Exercice : **Corde pincée (clavecin)** Déterminer la solution de la corde vibrante si $g(x) = 0$ et $f(x) = 2\alpha\frac{x}{L}$ si $x \leq \frac{L}{2}$ et $f(x) = 2\alpha\frac{L-x}{L}$ si $x \geq \frac{L}{2}$.

8.1.2 L'équation de la chaleur

Voici donc le problème au coeur du livre qui est à l'origine des séries de Fourier, la *Théorie analytique de la Chaleur* de Joseph Fourier. Un premier manuscrit présenté par Fourier à l'Académie en 1807, et un second en 1811, ont rencontré une vive opposition de la part du Comité. Ce livre a finalement été publié en 1822, après la mort de Lagrange en 1813 qui s'était violemment opposé aux idées avant–gardistes de Fourier, .

L'équation

On suppose une tige homogène de longueur L dans laquelle on étudie l'évolution spatio–temporelle de la température $u(x,t)$. Les conditions aux bords peuvent varier d'un problème à l'autre, mais les plus simples à traiter sont celle utilisées ci–après.

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} = \kappa \frac{\partial^2 u(x,t)}{\partial x^2} \qquad & \textit{pour } 0 < x < L,\ t \geq 0 \\ u(x,0) = f(x) \qquad & \textit{pour } 0 < x < L \text{ (conditions initiales)} \\ u(0,t) = 0,\ u(L,t) = 0 \qquad & \textit{pour } t \geq 0 \text{ (conditioins aux bords)} \end{aligned}$$

Méthode de séparation des variables

Comme dans le paragraphe précédent, on cherche la solution sous la forme

$$u(x,t) = X(x) \cdot T(t).$$

Ainsi l'équation de la chaleur devient

$$X(x)\ \frac{dT(t)}{dt} = \kappa \frac{d^2X(x)}{dx^2}\ T(t).$$

Une division par $X(x) \cdot T(t)$ sépare les variables et permets de conclure à l'existence d'une constante λ telle que

$$\frac{1}{\kappa}\ \frac{1}{T(t)} \frac{dT(t)}{dt} = \frac{1}{X(x)} \frac{d^2X(x)}{dx^2} = -\lambda.$$

De nouveau, on est confronté à deux équations différentielles ordinaires : celle pour $X(x)$ est identique à celle du paragraphe précédent et on obtient donc la même solution. L'autre équation donne $T'(t) = -\kappa\lambda T(t)$ avec comme solution $T(t) = C \cdot e^{-\kappa\lambda t}$. Finalement la solution générale est de la forme

$$u(x,t) = \sum_{k \geq 1} C_k \cdot \sin\left(k\pi \frac{x}{L}\right) \exp\left(-\kappa \left(\frac{k\pi}{L}\right)^2 t\right).$$

Si l'on pose ici $t = 0$, on obtient, à l'aide de la condition initiale, les formules

$$f(x) = u(x,0) = \sum_{k \geq 1} C_k \sin\left(k\pi \frac{x}{L}\right) \quad \Rightarrow \quad C_k = \frac{2}{L} \int_0^L f(x)\ \sin\left(k\pi \frac{x}{L}\right) dx$$

ce qui complète le problème.

8.2 Le problème de Dirichlet pour l'équation de Laplace

L'équation du potentiel (ou équation de Laplace), en deux dimensions, est

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0.$$

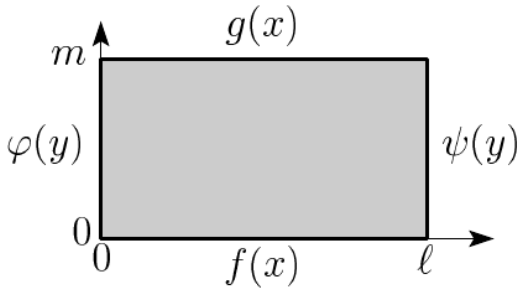
Tirant son origine de l'astronomie, cette équation a une grande importance dans presque toute la physique.

Problème de Dirichlet. : *soit donné un domaine U de frontière ∂U et une fonction $f(x,y)$ définie sur ∂U. Trouver $u(x,y)$ avec*

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0 \text{ dans } U \quad et \quad u(x,y) = f(x,y) \text{ sur } \partial U.$$

8.2.1 Le problème de Dirichlet pour un rectangle.

Considérons le rectangle $0 \le x \le l$ et $0 \le y \le m$ et exprimons les conditions aux bords par quatre fonctions données $f(x)$, $g(x)$ pour $0 \le x \le l$ et $\phi(y)$, $\psi(y)$ pour $0 \le y \le m$. Trouver la solution qui satisfasse ces conditions.



Solution On suppose d'abord $\phi(y) = \psi(y) = 0$. Posons, pour séparer les variables $u(x,y) = X(x) \cdot Y(y)$, ce qui donne

$$\frac{X''(x)}{X(x)} = -\frac{Y''(y)}{Y(y)} = -\lambda.$$

Comme $X(0) = X(l) = 0$, on a pour X et λ les mêmes solutions que pour la partie spatiale de la corde vibrante :

$$\lambda = \lambda_k = \left(\frac{k\pi}{l}\right)^2, \quad X(x) = \sin\frac{k\pi x}{l}, \quad k = 1, 2, 3, \ldots$$

L'équation pour Y donne pour $Y(y)$ les fonctions hyperboliques :

$$Y(y) = c_1 \cosh\sqrt{\lambda}y \; + \; c_2 \sinh\sqrt{\lambda}y.$$

Pour faciliter le traitement des conditions aux bords $y = 0$ et $y = m$, on choisit comme base $\sinh\sqrt{\lambda}y$ (qui est nul en $y = 0$) et $\sinh\sqrt{\lambda}(m-y)$ (qui est nul en $y = m$). Ainsi, on a, de nouveau par superposition, la solution générale

$$u(x,y) = \sum_{k=1}^{\infty} \sin\frac{k\pi x}{l}\left(c_k \cosh\frac{k\pi y}{l} \; + \; d_k \sinh\frac{k\pi(m-y)}{l}\right).$$

En utilisant les conditions initiales $y = 0$ et $y = m$, on trouve

$$d_k = \frac{2}{l \cdot \sinh\frac{k\pi m)}{l}} \int_0^l f(x) \; \sin\left(k\pi\frac{x}{l}\right) dx, \quad c_k = \frac{2}{l \cdot \sinh\frac{k\pi m)}{l}} \int_0^l g(x) \; \sin\left(k\pi\frac{x}{l}\right) dx.$$

Similairement, on calcul une solution $v(x,y)$, nulle sur les droites horizontales et satisfaisant $\phi(y)$, $\psi(y)$ pour $0 \le y \le m$. La solution finale est alors $u(x,y) + v(x,y)$.

8.2.2 Le problème de Dirichlet pour le disque.

Un deuxième cas où le problème de Dirichlet peut être résolu à l'aide des séries de Fourier est celui d'un disque. Soit D le disque unitaire. On choisit des coordonnées polaires r, ϕ. Pour la fonction $v(r,\phi)$ l'équation de Laplace devient

$$\frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\frac{\partial v}{\partial r} + \frac{1}{r^2}\frac{\partial^2 v}{\partial \phi^2} = 0.$$

La condition au bord est exprimée par $v(1,\phi) = f(\phi)$. Déterminer la solution de l'équation et montrer qu'elle peut s'écrire sous la forme

$$v(r,\phi) = \frac{1}{2\pi}\int_0^{2\pi} \frac{1-r^2}{1-2r\cos(\phi-\psi)+r^2} f(\psi) d\psi.$$

8.3 Équation de Poisson

L'équation de Poisson joue un rôle important en
— en **électrosatique** : définit le champs electrique induit par une distribution de charges connues,
— en **gravitation** : définit le champs de gravitation induit par une distribution de masse connues,
— en **thermique** : définit la distribution de température induite par un terme de source,
— en **mécanique des fluides** : définit l'écoulement d'unfluide visqueux induit par un gradient de pression.

> **Équation de Poisson :** *Soit Ω un domaine de frontière $\partial\Omega$, u une fonction inconnue sur Ω, ρ une fonction connue sur Ω et g une fonction connue définie sur $\partial\Omega$. Résoudre un problème de Poisson sur Ω revient à déterminer l'inconnue u telle que*
>
> $$\begin{aligned} \Delta u &= \rho \qquad sur\ \Omega \\ u &= g \qquad sur\ \partial\Omega, \end{aligned}$$
>
> *où Δ est l'opérateur différentiel Laplacien.*

La résolution d'un problème de Poisson passe par la résolution du problème d'Helmholtz associé :

> **Problème de Helmholtz associé :** *Trouver U une fonction inconnue sur Ω et Λ un réel inconnue tels que*
>
> $$\begin{aligned} \Delta U + \lambda^2 U &= 0 \qquad sur\ \Omega \\ U &= g \qquad sur\ \partial\Omega. \end{aligned}$$

La résolution du problème de Hemlholtz associé donne une infinité de couple $\{U_n, \lambda_n\}$ où les fonctions U_n forment une base orthogonale de solution vis a vis du produit scalaire

$$(f, g) = \int_\omega f\, g^\dagger \, \mathrm{d}\Omega.$$

Ce problème résolu, la solution du problème de Poisson est triviale :

1. On décompose la fonction source ρ sur la base des solutions du problème de Helmholtz

   $$\rho = \sum_n r_n\, U_n,$$

   les coefficients r_n se détermninent par un calcul intégral : $r_n = \frac{(U_n,\rho)}{(U_n,U_n)}$.
2. On décompose la fonction inconnue u sur la base des solutions du problème de Helmholtz

   $$u = \sum_n x_n\, U_n,$$

3. On introduit les deux décomposition dans l'équation de Poisson

   $$\Delta u = \sum_n x_n\, \lambda_n^2\, U_n = \sum_n r_n\, U_n$$

   que l'on projette scalairement sur U_m pour déterminer l'ensemble des coefficients x_m :

   $$x_m = \frac{r_m}{\lambda_m^2}.$$

ce qui complète le problème.

8.4 Équation des ondes — membrane circulaire

On suppose une membrane circulaire élastique dont on veut étudier la vibration transversale comme une fonction de trois variables $u(x, y, t)$. On suppose pour simplifier que la déformation est nulle au bord de la membrane. La mise en équation du problème se ramène à la résolution mathématique suivante

$$\left\{\begin{array}{rcccc} \frac{\partial^2 u}{\partial t^2} & = & 0 & \text{pour} & (x,y)\in\Omega \qquad t>0 \\ u(x,y,0) & = & f(x,y) & \text{pour} & (x,y)\in\Omega \quad \text{condition initiale, position} \\ \frac{\partial u}{\partial t}(x,y,0) & = & g(x,y) & \text{pour} & (x,y)\in\Omega \quad \text{condition initiale, vitesse} \\ u(x,y,t) & = & 0 & \text{pour} & (x,y)\in\partial\Omega \quad \text{condition au bord} \end{array}\right.$$

où, pour une membrane circulaire $\Omega = \{(x, y); x^2 + y^2 < 1\}$

8.4.1 Séparation des variables

Première séparation des variables Comme pour l'équation des ondes dans une dimension, on sépare le temps t des variables d'espace et on cherche des solutions de la forme $u(x, y, t) = T(t) \cdot v(x, y)$. Ceci conduit à

$$\frac{1}{a^2}\frac{T''(t)}{T(t)} = \frac{1}{v(x,y)}\,\Delta v(x,y) = -\lambda^2.$$

L'équation différentielle pour $T(t)$ donne les oscillations en temps

$$T(t) = A\cos(a\Lambda t) + B\sin(a\Lambda t)$$

et pour $v(x, y)$ on obtient le **problème de Helmholtz**

$$\begin{array}{rcll} \Delta v + \lambda^2 v & = & 0 & sur \quad \Omega \\ v & = & 0 & sur \quad \partial\Omega \end{array}$$

Si Ω est le disque unité, on peut résoudre ce problème. Il est naturel d'introduire des coordonnées polaires $x = r\cos\phi$, $y = r\sin\phi$ et de considérer la fonction $w(r, \phi) = v(r\cos\phi, r\sin\phi)$. En exprimant le laplacien Δv en terme de w, le problème devient

$$\frac{\partial^2 w}{\partial r^2} + \frac{1}{r}\frac{\partial w}{\partial r} + \frac{1}{r^2}\frac{\partial^2 w}{\partial \phi^2} + \lambda^2 w = 0$$
$$w(1,\phi) = 0, \quad w(r,0) = w(r,2\pi), \quad \frac{\partial w}{\partial \phi}(r,0) = \frac{\partial w}{\partial \phi}(r,2\pi),$$

où la limite $\lim_{r\to 0} w(r, \phi)$ est constante et indépendante de ϕ.

Deuxième séparation des variables. Pour résoudre l"équation aux dérivées partielles, on pose $w(r, \phi) = R(r) \cdot \Phi(\phi)$ ce qui donne

$$\frac{1}{R(r)}\left(r^2\, R"r) + r\; R'(r) + \lambda^2 r^2\; R(r)\right) = -\frac{\Phi"(\phi)}{\Phi(\phi)} = C.$$

La condition de périodicité, $\Phi(0) = \Phi(2\pi)$ et $\Phi'(O) = \Phi'(2\pi)$, implique que $C = n^2$ avec un entier n. Les solutions pour $\Phi(\phi)$ sont alors

$$\Phi_n(\phi) = a_n\cos n\phi \;+\; b_n\sin n\phi.$$

Pour la fonction $R(r)$ on obtient l'équation différentielle

$$r^2\, R"r) + r\; R'(r) + \left(\lambda^2 r^2 - n^2\right)\; R(r) = 0$$

avec les conditions aux bords

$$R(1) = 0 \quad \text{et} \quad R(0) = \left\{\begin{array}{cc} 0 & \text{si} \quad n > 0 \\ \text{une valeur finie} & \text{si} \quad n = 0. \end{array}\right.$$

Pour $\Lambda > 0$, la transformation $x = \lambda r$ et $y(x) = R(r)$ permet d'éliminer le paramètre λ de l'équation différentielle. On obtient ainsi l'**équation de Bessel**.

8.4.2 Solution de l'équation des ondes (membrane circulaire)

Pour tout $\lambda > 0$, la fonction $R(r) = J_n(\lambda r)$ est solution de l'équation différentielle radiale et satisfait la condition pour $R(0)$. Pour satisfaire la condition $R(1) = 0$, il faut que λ soit un zéro de J_n. La fonction

$$R(r) = J_n(j_{n,k} r)$$

satisfait alors l'équation radiale et la condition au bord. Par conséquent, pour $n \geq 0$ et pour tout $k \geq 1$,

$$w(r, \phi) = J_n(j_{n,k} r) \left(a_{n,k} \cos n\phi + b_{n,k} \sin n\phi\right)$$

est une solution de l'équation de Helmholtz en coordonnées polaires qui satisfait les conditions au bord. La solution générale $u(x, y, t)$ de l'équation des ondes est donc

$$\begin{array}{rcl} u\left(r\cos\phi, r\sin\phi, t\right) & = & \sum_{n=0}^{\infty} \sum_{n=1}^{\infty} J_n(j_{n,k} r) \left(a_{n,k} \cos n\phi + b_{n,k} \sin n\phi\right) \cos a j_{n,k} t) \\ & + & \sum_{n=0}^{\infty} \sum_{n=1}^{\infty} J_n(j_{n,k} r) \left(c_{n,k} \cos n\phi + d_{n,k} \sin n\phi\right) \in a j_{n,k} t). \end{array}$$

Il reste à déterminer les coefficients $a_{n,k}$, $b_{n,k}$, $c_{n,k}$ et $d_{n,k}$ afin de satisfaire les conditions initiales pour la position et la vitesse. Pour ce calcul on utilise l'orthogonalité des fonctions de Bessel.

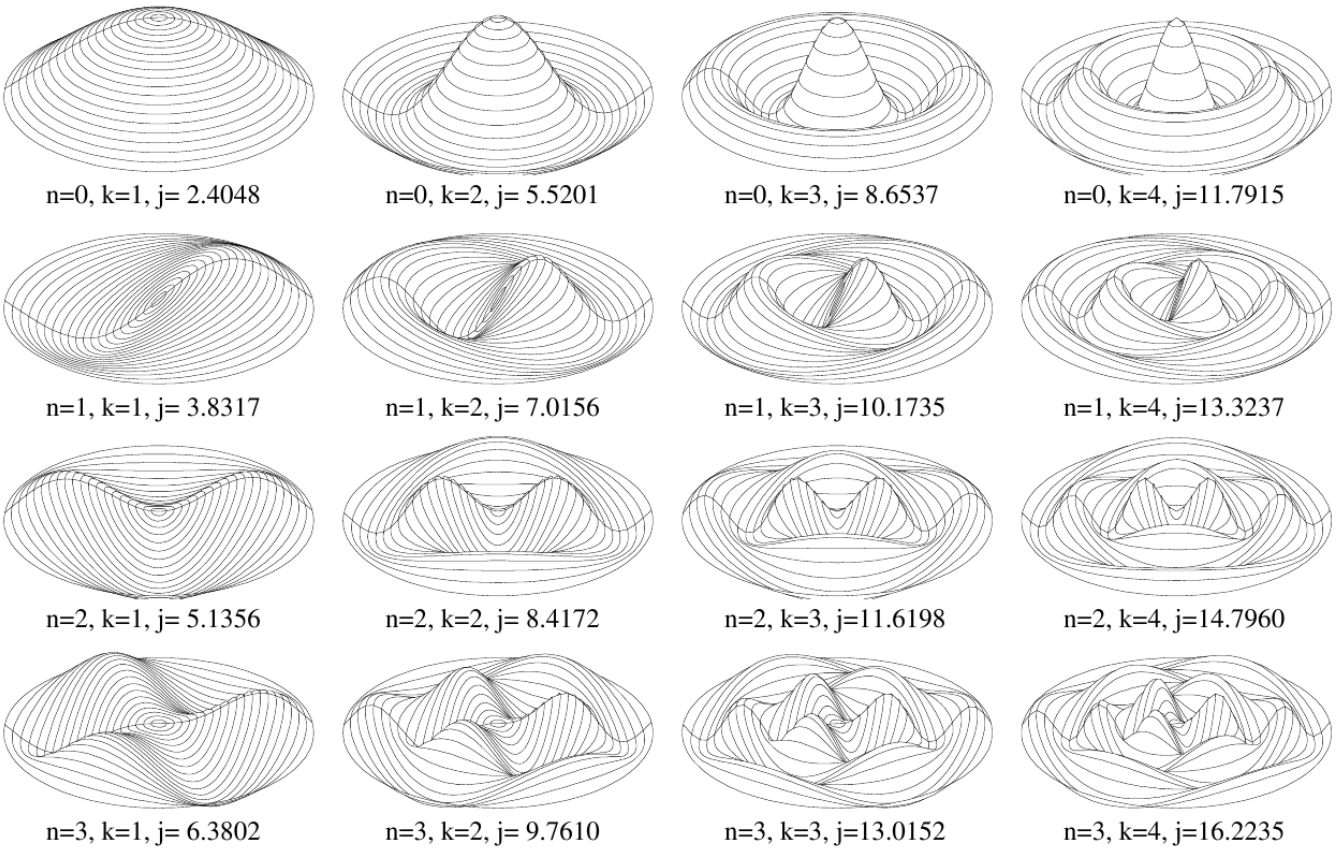


FIGURE 8.1 – Fonctions propres $J_n(j_{n,k} r) \cos n\phi$ du Laplacien sur le Disque.

8.4.3 Satisfaire les conditions initiales

Commençons par celle pour la position

$$u(x, y, 0) = f(x, y).$$

On travaille bien sûr en coordonnées polaires et on développe la fonction en série de Fourier trigonométrique la partie angulaire.

$$f(r\cos\phi, r\sin\phi) = \sum_{n=0}^{\infty} A_n(r) \cos n\phi + \sum_{n=0}^{\infty} B_n(r) \sin n\phi.$$

Les coefficients de Fourier sont calculés par les relations

$$\begin{array}{rcl} A_0(r) & = & \frac{1}{2\pi}\int_0^{2\pi} f(r\cos\phi, r\sin\phi)d\phi \\ A_{n>1}(r) & = & \frac{1}{\pi}\int_0^{2\pi} f(r\cos\phi, r\sin\phi)\cos n\phi\, d\phi \\ B_{n>1}(r) & = & \frac{1}{2\pi}\int_0^{2\pi} f(r\cos\phi, r\sin\phi)\sin n\phi\, d\phi \end{array}$$

En comparant avec la solution $u(r\cos\phi, r\sin\phi, 0)$, on obtient

$$A_n(r) = \sum_{k=1}^{\infty} a_{n,k} J_n\left(j_{n,k} r\right), \quad B_n(r) = \sum_{k=1}^{\infty} b_{n,k} J_n\left(j_{n,k} r\right).$$

Il reste à multiplier ces relations par $rJ_n\left(j_{n,l}r\right)$ et à intégrer sur $[0,1]$ pour obtenir, grâce au théorème d'orthogonalité,

$$a_{nl} = \frac{2}{\left(J_n'(j_{n,l})\right)^2}\int_0^1 A_n(r) J_n\left(j_{n,l}r\right) r\, dr, \quad b_{nl} = \frac{2}{\left(J_n'(j_{n,l})\right)^2}\int_0^1 B_n(r) J_n\left(j_{n,l}r\right) r\, dr.$$

On retrouve ici le développement en série de Fourier–Bessel des fonctions $A_n(r)$ et $B_n(r)$.

Un calcul analogue permet de déterminer les coefficients $c_{n,k}$ et $d_{n,k}$ à partir des coefficients de Fourier $C_n(r)$ et $D_n(r)$ de la fonction $g(r\cos\phi, r\sin\phi)$ (condition initiale pour la vitesse. Avec les valeurs des coefficients, on a complètement résolu le problème sur le disque unité.

8.5 Problèmes

8.5.1 Écoulement potentiel (Janvier 2011)

Contexte physique du problème

Lorsque l'écoulement d'un fluide incompressible autour d'un solide peut être qualifié de "potentiel" (on ne détaillera pas ici les approximations requises), la vitesse en tout point du fluide peut être dérivée à partir d'une fonction scalaire Φ appelée "potentiel des vitesses" par la relation suivante

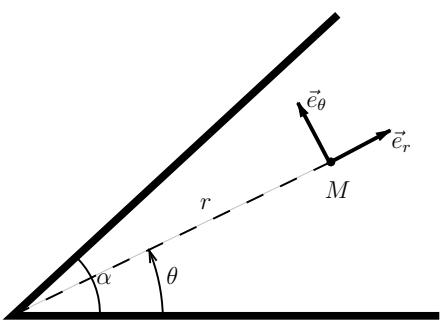
$$\overrightarrow{v} = \overrightarrow{\text{grad}}\ \Phi$$

Les solutions de l'écoulement sont obtenues en résolvant l'équation de Laplace pour le potentiel des vitesses :

$$\Delta\ \Phi = 0.$$

Les conditions aux bords de cet écoulement sont imposées par le fait que le fluide ne peut pénétrer dans le solide qui borne son écoulement, et que donc la vitesse du fluide au point de contact avec le solide doit être tangente à la paroi du solide.

On cherche ici à déterminer l'ensemble des solutions du champ de vitesse caractérisant l'écoulement potentiel d'un fluide entre deux parois planes formant un angle α entre elles. Étant donnée la configuration du problème, on utilisera les coordonnées polaires pour le potentiel des vitesses $\Phi(r,\theta)$. On donne pour mémoire les expressions du laplacien et du gradient dans ce système de coordonnée :



$$\Delta\Phi(r,\theta) = \frac{\partial^2\Phi(r,\theta)}{\partial r^2} + \frac{1}{r}\frac{\partial\Phi(r,\theta)}{\partial r} + \frac{1}{r^2}\frac{\partial^2\Phi(r,\theta)}{\partial\theta^2}$$

$$\overrightarrow{\text{grad}}\Phi(r,\theta) = \frac{\partial\Phi(r,\theta)}{\partial r}\overrightarrow{e}_r + \frac{1}{r}\frac{\partial\Phi(r,\theta)}{\partial\theta}\overrightarrow{e}_\theta$$

Résolution mathématique

Mathématiquement cela revient à résoudre le problème suivant.

$$\frac{\partial^2\Phi(r,\theta)}{\partial r^2}+\frac{1}{r}\frac{\partial\Phi(r,\theta)}{\partial r}+\frac{1}{r^2}\frac{\partial^2\Phi(r,\theta)}{\partial\theta^2}=0 \tag{8.1}$$

avec les conditions aux bords suivantes

$$\frac{\partial\Phi}{\partial\theta}(r,\theta=0)=0;\quad \frac{\partial\Phi}{\partial\theta}(r,\theta=\alpha)=0 \tag{8.2}$$

1 - Utiliser la méthode de séparation des variables (on posera $\Phi(r,\theta)=f(r)\times g(\theta)$) pour montrer que la résolution de l'équation de Laplace se ramène à la résolution de deux équations différentielles ordinaires portant sur les fonctions f et g. Détailler et commenter les étapes importantes du calcul.

2 - Résoudre l'équation différentielle sur f en cherchant des solutions de la forme $f(r)=r^a$ ou a est une constante à déterminer.

3 - Résoudre l'équation différentielle portant sur g

4 - Utiliser les conditions aux bords (2) pour montrer que les solutions de l'équation de Laplace (1) sont de la forme

$$\Phi_k(r,\theta)=r^{k\pi/\alpha}\cos\left(k\frac{\pi}{\alpha}\theta\right);\quad k\in\mathbb{N}$$

où k est un indice entier.

Les questions suivantes sont indépendantes.

5 - Un utilisant la famille de solution Φ_k, déterminer la forme générale de la solution pour le potentiel des vitesses et pour le vecteur vitesse.

6 - Montrer que la famille des fonctions Φ_k forme une famille de fonctions othogonales deux à deux si on introduit le produit scalaire suivant

$$(\Phi_k,\Phi_l)=\int_0^\alpha\int_0^\infty \Phi_k(r,\theta)\;\Phi_l(r,\theta)\;r\;dr\;d\theta \tag{8.3}$$

7 - On suppose maintenant connue une solution particulière $\Psi(r,\theta)$ de l'équation (1) avec les conditions aux bords (2). On souhaite décomposer cette solution dans la base des fonctions Φ_k. Indiquer (sans les effectuer) les calculs à mener pour parfaire cette décomposition à l'aide du produit scalaire (3).

8.5.2 Problème de thermique en coordonnées polaire (Janvier 2012)

Thermique d'une couronne

On considère une pièce métallique occupant une couronne circulaire de rayons intérieur et extérieur r_1 et r_2 avec $0<r_1<r_2$. Pour étudier la thermique de cette pièce, on cherche une solution de l'équation de la chaleur bi–dimensionnelle écrite en coordonnées polaires ($x=r\cos\theta, y=r\sin\theta$) :

$$\frac{\partial^2T(r,\theta)}{\partial r^2}+\frac{1}{r}\frac{\partial T(r,\theta)}{\partial r}+\frac{1}{r^2}\frac{\partial^2T(r,\theta)}{\partial\theta^2}=0. \tag{8.4}$$

Par application de la méthode de séparation des variables, on cherche T sous la forme $T(r,\theta)=R(r)\;\Theta(\theta)$ où R et Θ sont les nouvelles inconnues du problème. la symétrie circulaire impose que la fonction Θ soit 2π–périodique.

1 - Déterminer, en fonction d'un paramètre $\nu^2\geq 0$, deux équations différentielles ordinaires que doivent vérifier R et Θ pour que T soit une solution de l'équation (8.4).

On considère dans un premier temps le cas $\nu = 0$ et on notera R_0 et Θ_0 les solutions obtenues.

2 - Quelle est la solution pour Θ_0, compte–tenu de la périodicité ?

3 - Montrer que $R_0(r)$ peut se mettre sous la forme suivante : $R_0(r) = A_0 \ln\left(\frac{r}{r_1}\right) + B_0$.

On considère maintenant le cas $\nu \neq 0$ et on notera R_ν et Θ_ν les solutions obtenues.

4 - Quelle est la solution pour Θ_ν.

5 - En déduire les valeurs de ν compatibles avec la périodicité.

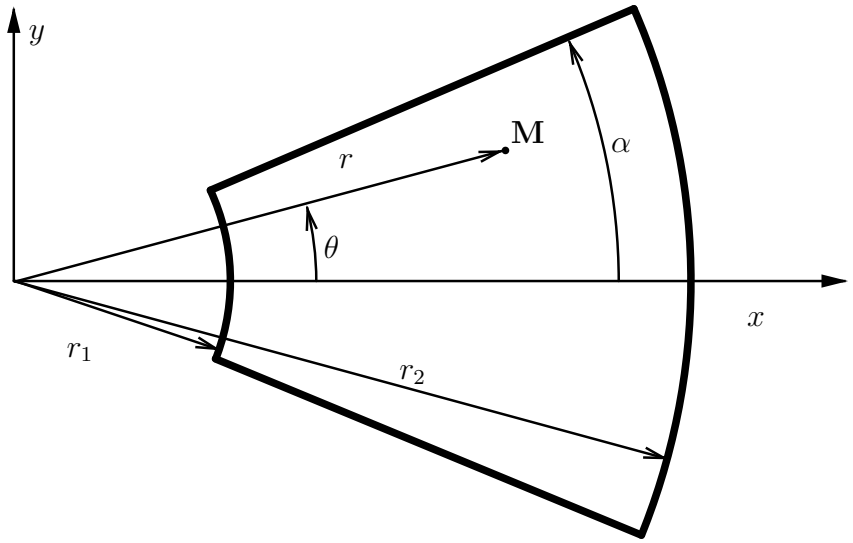
6 - Quelle est la solution pour R_ν ?

On considère la condition aux limites $T(r = r_1, \theta) = 0$ et $T(r = r_2, \theta) = f(\theta)$.

7 - Déterminer la solution du problème si $f(\theta) = \cos(\theta)$.

Thermique d'une portion de couronne

On considère une pièce métallique occupant un secteur angulaire d'ouverture 2α d'une couronne circulaire de rayons intérieur et extérieur r_1 et r_2 avec $0 < r_1 < r_2$ (voir figure).



1 - Montrer qu'en introduisant le changement de variable $u = \ln r$, l'équation de la chaleur bi–dimensionnelle (8.4) est équivalente à

$$\frac{\partial^2 T(u,\theta)}{\partial u^2} + \frac{\partial^2 T(u,\theta)}{\partial \theta^2} = 0. \tag{8.5}$$

Par application de la méthode de séparation des variables, on cherche T sous la forme $T(u, \theta) = U(u)\,\Theta(\theta)$ où U et Θ sont les nouvelles inconnues du problème.

2 - Déterminer en fonction d'un paramètre ν, deux équations différentielles ordinaires que doivent vérifier U et Θ pour que T soit une solution de l'équation (8.5).

3 - Exprimer les solutions de chacune des équations dans les trois cas suivant : $\nu = 0$, $\nu < 0$, $\nu > 0$.

4 - Déterminer en fonction d'un paramètre ν, deux équations différentielles ordinaires que doivent vérifier U et Θ pour que T soit une solution de l'équation (8.5).

On pose $u_1 = \ln(r_1)$ et $u_2 = \ln(r_2)$ et on cherche les solutions de cette équation soumises aux conditions suivantes

$$T\,(u = u_1, \theta) = 0; \quad T\,(u = u_2, \theta) = 0; \quad T\,(u, \theta = \pm\alpha) = f(u); \tag{8.6}$$

où $f(r)$ est une distribution connue de température. Par raison de symétrie, on suppose de plus que $\forall \theta \in [-\alpha, \alpha] \quad T(r, \theta) = T(r, -\theta)$ **5 -** Montrer que les conditions de température nulle en u_1 et en

u_2 imposent des solutions pour U de la forme :

$$U_k(u,\theta) = B_k \ \sin\left(k\pi\frac{u-u_1}{u_2-u_1}\right).$$

6 - Montrer que la condition de symétrie de la température impose des solutions pour Θ de la forme :

$$\Theta_k(u,\theta) = C_k \ \cosh\left(\frac{k\pi}{u_2-u_1}\theta\right).$$

7 - Montrer que la solution générale de (8.5) satisfaisant les conditions au bords (8.6) est de la forme :

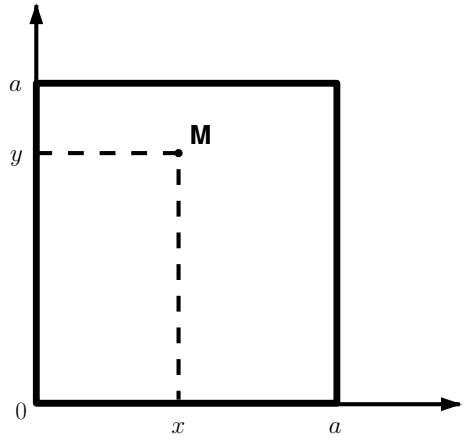
$$T(u,\theta) = \sum_{k=0}^{k=\infty} A_k \ \sin\left(k\pi\frac{u-u_1}{u_2-u_1}\right)\cosh\left(\frac{k\pi}{u_2-u_1}\theta\right). \tag{8.7}$$

8 - Déterminer, en fonction de f, les coefficients A_k.

8.5.3 Équation de la chaleur avec un terme de source (Janvier 2013)

Position du problème (10 min)

On considère le problème de la diffusion thermique dans une plaque carré de coté a. On note x et y les coordonnées cartésienne d'un point M de la plaque et t le temps. La température au point M est notée $T\,(t,x,y)$. Les bords de la plaque sont maintenus à une température nulle 0° C. On suppose enfin que la plaque est munie d'un dispositif de chauffage interne, modélisé par un terme de source constant $S\,(x,y)$ connue.



L'équation de diffusion de la chaleur en présence du terme de source S s'écrit

$$\frac{\partial}{\partial t}T\,(t,x,y) = \kappa\ \left(\frac{\partial^2 T(t,x,y)}{\partial x^2} + \frac{\partial^2 T(t,x,y)}{\partial y^2}\right) + \epsilon S\,(x,y)\,, \tag{8.8}$$

où κ et ϵ sont deux constantes physiques positives qui dépendent de la matière de la plaque.

Les conditions au bords de la plaque sont imposées :

$$\begin{cases} T(t,x=0,y) &= 0 \quad \text{pour } 0\le y\le a;\ t\ge 0\\ T(t,x=a,y) &= 0 \quad \text{pour } 0\le y\le a;\ t\ge 0\\ T(t,x,y=0) &= 0 \quad \text{pour } 0\le x\le a;\ t\ge 0\\ T(t,x,y=a) &= 0 \quad \text{pour } 0\le x\le a;\ t\ge 0 \end{cases} \tag{8.9}$$

On suppose qu'à l'instant initial, la température de la plaque est homogène à 0°C. La condition initiale de température est donc

$$T(t=0,x,y) = 0 \quad \text{pour } 0\le x\le a;\ 0\le y\le a. \tag{8.10}$$

Le terme de source S va conduire à une élévation de température dans la plaque. La distribution de température au temps long va converger vers une distribution stationnaire notée $T_\infty(x,y)$. L'équation de diffusion de la chaleur8.8 étant linéaire, on peut décomposer la recherche de solution en deux parties en posant

$$T\left(t,x,y\right)=T_t\left(t,x,y\right)+T_\infty\left(x,y\right)$$

ou

— $T_\infty\left(x,y\right)$ représente la distribution stationnaire de température
— $T_t\left(t,x,y\right)$ représente le transitoire.

La condition initiale (8.10) impose la condition initiale suivante pour $T_t\left(t,x,y\right)$:

$$T_t\left(t,x,y\right)=-T_\infty\left(x,y\right).$$

On peut ainsi décomposer la recherche de solution en deux problèmes :

— **Problème 1 :** solution asymptotique

$$\left\{\begin{array}{rcll}\kappa\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)T_\infty\left(x,y\right)+\epsilon S\left(x,y\right) & = & 0 & \\ T_\infty(x=0,y) & = & 0 & \text{pour } 0\leq y\leq a;\ t\geq 0\\ T_\infty(x=a,y) & = & 0 & \text{pour } 0\leq y\leq a;\ t\geq 0\\ T_\infty(x,y=0) & = & 0 & \text{pour } 0\leq x\leq a;\ t\geq 0\\ T_\infty(x,y=a) & = & 0 & \text{pour } 0\leq x\leq a;\ t\geq 0\end{array}\right. \tag{8.11}$$

— **Problème 2 :** solution transitoire

$$\left\{\begin{array}{rcl}\frac{\partial}{\partial t}T_t\left(t,x,y\right) & = & \kappa\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)T_t\left(t,x,y\right)\\ T_t(t,x=0,y) & = & 0\quad \text{pour } 0\leq y\leq a;\ t\geq 0\\ T_t(t,x=a,y) & = & 0\quad \text{pour } 0\leq y\leq a;\ t\geq 0\\ T_t(t,x,y=0) & = & 0\quad \text{pour } 0\leq x\leq a;\ t\geq 0\\ T_t(t,x,y=a) & = & 0\quad \text{pour } 0\leq x\leq a;\ t\geq 0\\ T_t\left(t,x,y\right) & = & -T_\infty\left(x,y\right).\end{array}\right. \tag{8.12}$$

L'objectif de ce problème est la résolution de ces deux problèmes pour un terme de source donnée. La résolution de ces deux problèmes passe par la solution du même problème de Helmholtz qui fait l'objet de la première partie.

Partie 1 : Résolution du problème de Helmholtz associé (1 heure)

Soit $U(x,y)$ une fonction inconnue de deux variables et λ un réel inconnu. On propose de résoudre le problème suivant :

$$\left\{\begin{array}{rcll}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)U\left(x,y\right)+2\lambda^2\,U\left(x,y\right) & = & 0 & \\ U(x=0,y) & = & 0 & \text{pour } 0\leq y\leq a;\ t\geq 0\\ U(x=a,y) & = & 0 & \text{pour } 0\leq y\leq a;\ t\geq 0\\ U(x,y=0) & = & 0 & \text{pour } 0\leq x\leq a;\ t\geq 0\\ U(x,y=a) & = & 0 & \text{pour } 0\leq x\leq a;\ t\geq 0\end{array}\right. \tag{8.13}$$

Par application de la méthode de séparation des variables, on cherche U sous la forme $U(x,y)=X(x)\,Y(y)$ où X et Y sont les nouvelles inconnues du problème.

1 - Détailler les étapes de calcul permettant d'achever la séparation des variables sous la forme

$$\frac{1}{X(x)}\frac{d^2X(x))}{dx^2}+\lambda^2=-\frac{1}{Y(y)}\frac{d^2Y(y))}{dy^2}-\lambda^2=\mu.$$

Expliquer en particulier comment et pourquoi la constante μ a été introduite.

2 - Déterminer les conditions aux bords pour les fonctions X et Y.

3 - En déduire que les conditions aux bords imposent $-\lambda^2 \leq \mu \leq \lambda^2$.

4 - Vérifier que les fonctions

$$U_{n,m}(x,y) = \sin\left(n\,\pi\frac{x}{a}\right)\,\sin\left(m\,\pi\frac{y}{a}\right) \quad (n,m) \in \mathbb{N}^2 \tag{8.14}$$

sont solutions du problème de Helmholtz (8.13) pour $2\lambda^2 = \lambda_{n,m}^2 = (n^2+m^2)\,\frac{\pi^2}{a^2}$.

On introduit le produit scalaire suivant :

$$(f,g) = \int_0^a \int_0^a f(x,y)g(x,y)\,dxdy. \tag{8.15}$$

5 - Montrer que les solutions $U_{n,m}$ définies par (8.14) sont orthogonales au sens du produit scalaire (8.15).

6 - Calculer les carrés de normes $h_{n,m} = (U_{n,m}, U_{n,m})$ des solutions $U_{n,m}$ définies par (8.14)

On admettra pour la suite que la famille des solutions $\{U_{n,m}\}$ forme une base complète de l'espace des fonctions définies pour $(x,y) \in [0,a]^2$.

Partie 2 : Résolution du problème 1 (1 heure)

Le terme de source $S(x,y)$ du problème 1 peut être décomposé dans la base des fonctions solutions de (8.13) définies par (8.14). On pose donc

$$S(x,y) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} s_{n,m}\,U_{n,m}(x,y). \tag{8.16}$$

7 - En utilisant l'orthogonalité des solutions $U_{n,m}$ vis à vis du produit scalaire (8.15), montrer que

$$s_{n,m} = \frac{(U_{n,m}, S)}{h_{n,m}} \tag{8.17}$$

On cherche la solution du problème (8.11) sous la forme d'une série de Fourier généralisée basée sur les solutions de (8.13).

$$T_\infty(x,y) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \alpha_{n,m}\,U_{n,m}(x,y). \tag{8.18}$$

8 - En utilisant les décompositions (8.16) et (8.18), ainsi que l'orthogonalité des solutions $U_{n,m}$ vis à vis du produit scalaire (8.15), montrer que les coefficients $\alpha_{n,m}$ sont définis par

$$\alpha_{n,m} = \frac{\epsilon}{\kappa}\frac{s_{n,m}}{\lambda_{n,m}^2},$$

ce qui complète le problème 1.

Applications

Calculer les coefficients $\alpha_{n,m}$ pour les fonctions de sources suivantes

9 - Source uniforme

$$S(x,y) = s \quad \text{pour} \quad 0 \leq x \leq a,\ 0 \leq y \leq a. \tag{8.19}$$

10 - Source centrée

$$\begin{cases} S(x,y) = s & \text{si} \quad \frac{a}{2} - b \leq x \leq \frac{a}{2} + b,\ \frac{a}{2} - b \leq y \leq \frac{a}{2} + b. \\ S(x,y) = 0 & \text{sinon} \end{cases}, \tag{8.20}$$

avec $0 < b < \frac{a}{2}$.

Partie 3 : Résolution du problème 2 (30 min)

Pour résoudre le problème 2 il convient de séparer les variables de temps et d"espace. On pose

$$T_t(t,x,y) = \Theta(t)\ E(x,y).$$

11 - Pratiquer la séparation des variables proposées et déterminer le signe de la constante à introduire pour obtenir un comportement temporel exponentiellement décroissant.

12 - Noter que la solution de la dépendance spatiale est, à une constante près, la solution trouvée pour le problème (8.13). En conséquence, poser $E(x,y) = U_{n,m}(x,y)$ et montrer que

$$\Theta(t) = \exp\left(-\kappa\lambda_{n,m}^2 t\right)$$

13 - En déduire que la solution générale du problème 2 est de la forme

$$T_t(t,x,y) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \beta_{n,m}\, U_{n,m}(x,y)\ \exp\left(-\kappa\lambda_{n,m}^2 t\right).$$

14 - Utiliser la condition initiale pour montrer que

$$\beta_{n,m} = -\alpha_{n,m}.$$

Synthèse

L'ensemble des résultats des parties précédentes démontre que la solution du problème est

$$\left\{\begin{array}{rl} (f,g) & = \int_0^a \int_0^a f(x,y)g(x,y)\,dxdy. \\ U_{n,m}(x,y) & = \sin\left(n\,\pi\frac{x}{a}\right)\ \sin\left(m\,\pi\frac{y}{a}\right) \quad (n,m)\in\mathbb{N}^2 \\ \lambda_{n,m}^2 & = (n^2+m^2)\,\frac{\pi^2}{a^2} \\ h_{n,m} & = (U_{n,m},U_{n,m}) \\ T\,(t,x,y) & = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\alpha_{n,m}\left(1-\exp\left(-\kappa\lambda_{n,m}^2 t\right)\right)\ U_{n,m}(x,y) \\ \alpha_{n,m} & = \frac{\epsilon}{\kappa}\frac{(U_{n,m},S)}{h_{n,m}\lambda_{n,m}^2} \end{array}\right. \tag{8.21}$$

Chapitre 9

Analyse vectorielle

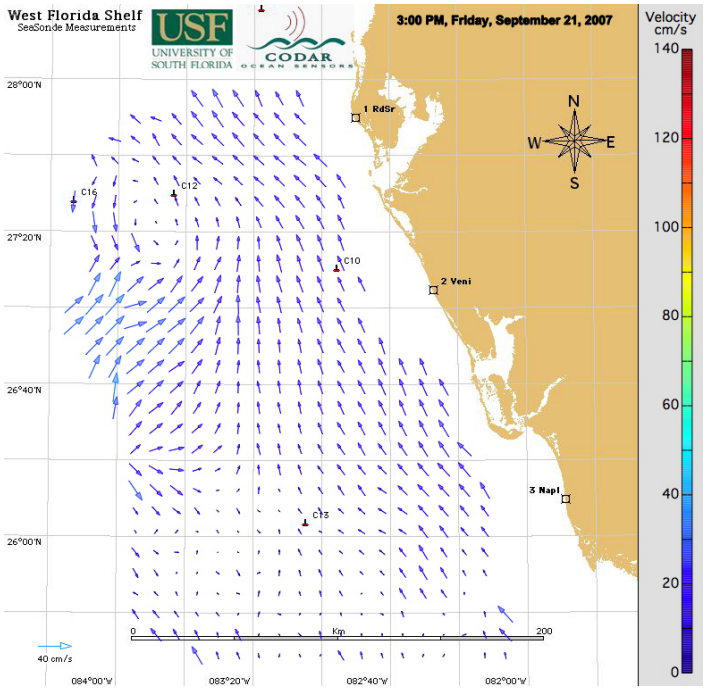
Sommaire

Dans ce chapitre, nous étudions l'analyse des champs vectoriels en (re)définissant les intégrales curvilignes, les intégrales de surface, établissant les liens entre ces intégrales et les intégrales simple, double et triple.

9.1 Les champs de vecteurs

Les vecteurs de la figure de gauche représentent la vitesse des courants marins en intensité, orientation et sens en différents points au large de la Floride. C'est un exemple de *champ vectoriel de vitesses.*

Un autre type de champ vectoriel, appelé champ de force, est celui qui associe à chaque point d'une région un vecteur représentatif d'une force. Par exemple, le champ gravitationnel.



9.1.1 Définition

Définition : *Soit E un sous ensemble de l'espace $\mathbb{R}^3$. Un* ***champ vectoriel*** *défini sur $\mathbb{R}^3$ est une fonction $\overrightarrow{F}$ qui fait correspondre à chaque point (x, y, z) de E un vecteur de dimension trois $\overrightarrow{F}(x, y, z)$.*

Un champ vectoriel $\overrightarrow{F}$ défini sur $\mathbb{R}^3$ peut être exprimé en fonction de ses composantes P, Q et R comme suit :

$$\overrightarrow{F}(x, y, z) = P(x, y, z)\ \overrightarrow{e}_x + Q(x, y, z)\ \overrightarrow{e}_y + R(x, y, z)\ \overrightarrow{e}_z,$$

où $(\overrightarrow{e}_x, \overrightarrow{e}_y, \overrightarrow{e}_z)$ est une base cartésienne de $\mathbb{R}^3$. Comme dans le cas des fonctions de plusieurs variables, on peut parler de continuité des champs vectoriels et démontrer de $\overrightarrow{F}$ est continue si ses fonctions composantes P, Q et R sont continues. On identifie parfois un point (x, y, z) avec son vecteur position $\overrightarrow{r} = (x, y, z)$ et on écrit $\overrightarrow{F}(\overrightarrow{r})$ au lieu de $\overrightarrow{F}(x, y, z)$.

En fonction de la géométrie du problème considéré, on peut être amené à utiliser d'autres systèmes de coordonnées

$$\overrightarrow{F}(r, \theta, z) = P(r, \theta, z)\ \overrightarrow{e}_r + Q(r, \theta, z)\ \overrightarrow{e}_\theta + R(r, \theta, z)\ \overrightarrow{e}_z,$$

$$\overrightarrow{F}(\rho, \theta, \phi) = P(\rho, \theta, \phi)\ \overrightarrow{e}_\rho + Q(\rho, \theta, \phi)\ \overrightarrow{e}_\theta + R(\rho, \theta, \phi)\ \overrightarrow{e}_\phi.$$

9.1.2 Les champs de gradient

Soit f une fonction de trois variables

Définition : *La dérivé de f en (x_0, y_0, z_0) dans la direction du vecteur unitaire $\overrightarrow{u} = (a, b, c)$ est*

$$f'_{\overrightarrow{u}}(x_0, y_0, z_0) = \lim_{h \to 0} \frac{f(x_0 + ha, y_0 + hb, z_0 + hc) - f(x_0, y_0, z_0)}{h}$$

Définition : *Le gradient d'une fonction f dans une base (x, y, z) est le vecteur dont les composantes sont les dérivées de la fonction selon les vecteurs de base*

$$\overrightarrow{grad}\, f = f'_{(1,0,0)}\ \vec{e}_x + f'_{(0,1,0)}\ \vec{e}_y + f'_{(O,0,1)}\ \vec{e}_z$$

ou encore

$$\overrightarrow{grad}\, f(x,y,z) = \frac{\partial f}{\partial x}\ \vec{e}_x + \frac{\partial f}{\partial y}\ \vec{e}_y + \frac{\partial f}{\partial z}\ \vec{e}_z$$

De là, $\overrightarrow{\text{grad}}\, f$ est manifestement un champ vectoriel sur $\mathbb{R}^3$ et s'appelle un **champ de gradient.**

Un champ vectoriel $\vec{F}$ est dit **champ vectoriel conservatif** s'il est le gradient d'une certaine fonction scalaire, c–à–d. , s'il existe une fonction f telle que $\overrightarrow{\text{grad}}\, f = \vec{F}$. Dans cette situation f est appelée un **fonction potentiel** de $\vec{F}$.

Remarque : Tous les champs vectoriels ne sont pas concervatifs, mais de tels champs interviennent fréquemment en physique.

Propriété : *La dérivée d'une fonction dans une direction $\vec{u}$ s'exprime simplement en terme de gradient comme*

$$f'_{\vec{u}} = \overrightarrow{grad}\, f \ \cdot\ \vec{u}.$$

9.1.3 Les lignes de champ

Les lignes de champ sont une famille de courbes de l'espace qui sont tangentes en tout point au champ de vecteur.

Supposons un champ vectoriel $\vec{F}$ défini sur $\mathbb{R}^3$ peut être exprimé en fonction de ses composantes P, Q et R comme suit :

$$\vec{F}(x,y,z) = P(x,y,z)\ \vec{e}_x + Q(x,y,z)\ \vec{e}_y + R(x,y,z)\ \vec{e}_z,$$

Supposons un déplacement élémentaire $d\vec{r} = dx\ \vec{e}_x + dy\ \vec{e}_y + dz\ \vec{e}_z$ le long d'une ligne de champ, la colinéarité de $\vec{F}$ et $d\vec{r}$ s'exprime par la nullité du produit vectoriel

$$\vec{F} \wedge d\vec{r} = \vec{0}$$

ce qui conduit aux relations

$$\frac{dx}{P} = \frac{dy}{Q} = \frac{dz}{R}.$$

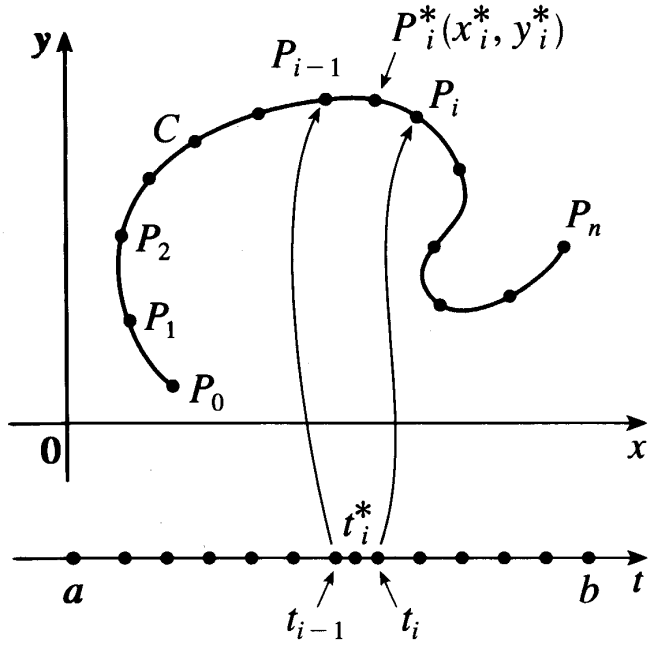
9.2 Les intégrales curvilignes

Dans cette section, nous allons définir une intégrale semblable à une intégrale simple, à ceci près que l'intégration porte sur une courbe C, au lieu d'un intervalle $[a, b]$. De telles intégrales sont appelées *intégrales curvilignes*.

Supposons une courbe plane C, définie par les équation paramétriques

$$x = x(t) \quad y = y(t) \quad a \leq t \leq b$$

où, de façons équivalente, par l'équation vectorielle $\overrightarrow{r}(t) = x(t)\ \overrightarrow{e}_x + y(t)\ \overrightarrow{e}_y$. La courbe C est supposée lisse, Ce qui signifie de $\overrightarrow{r}'$ est continue et $\overrightarrow{r}' \neq \overrightarrow{0}$. Si l'intervalle $[a, b]$ de variation du paramètre est divisé en n sous-intervalles $[t_{i-1}, t_i]$ d'égale longueur et si $x_i = x(t_i)$ et $y_i = y(t_i)$, alors l'arc C est subdivisé en n sous–arcs de longueur Δs_i (voir la figure ci–contre) sur chacun desquels est choisi un point $P^*\,(x_i^*, y_i^*)$ (qui est l'image d'un certain t_i^* appartenant à $[t_{i-1}, t_i]$.



On considère maintenant une fonction de deux variables f dont le domaine de définition comprend la courbe C, on calcule la valeur de f au point (x_i^*, y_i^*), on la multiplie par la longueur Δs_i du sous–arc et on fait la somme de ces produits

$$\sum_{i=1}^{n} f\,(x_i^*, y_i^*)\,\Delta s_i.$$

Cela ressemble à une somme de Riemann. On en prend la limite pour n tendant vers l'infini et par analogie avec l'intégrale simple, on adopte la définition suivante

Définition : *Soit f définie sur un courbe lisse C donnée par les équations paramétriques $x = x(t) \quad y = y(t) \quad a \leq t \leq b$. L'**intégrale curviligne de f le long de** C est*

$$\int_C f(x,y)ds = \lim_{n\to\infty} \sum_{i=1}^{n} f\,(x_i^*, y_i^*)\,\Delta s_i. \tag{9.1}$$

pourvu que cette limite existe.

On peut démonter que si f est continue la limite existe toujours et que l'intégrale curviligne est calculable par la relation

$$\int_C f(x,y)ds = \int_a^b f\,(x(t), y(t))\,\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\,dt \tag{9.2}$$

La valeur de l'intégrale ne dépend pas de la paramétrisation de la courbe pourvu que l'arc ne soit parcouru qu'une seule fois lorsque t varie de a à b.

Ces résultats se généralisent sans difficulté à trois dimensions et peuvent s'écrire de manière compacte

$$\int_C f(\overrightarrow{r})ds = \int_a^b f\,(\overrightarrow{r})\,||\overrightarrow{r}'(t)||dt.$$

9.3 Les intégrales curvilignes d'un champ vectoriel

On envisage maintenant un champ vectoriel

$$\overrightarrow{F}(x,y,z) = P(x,y,z)\ \overrightarrow{e}_x + Q(x,y,z)\ \overrightarrow{e}_y + R(x,y,z)\ \overrightarrow{e}_z,$$

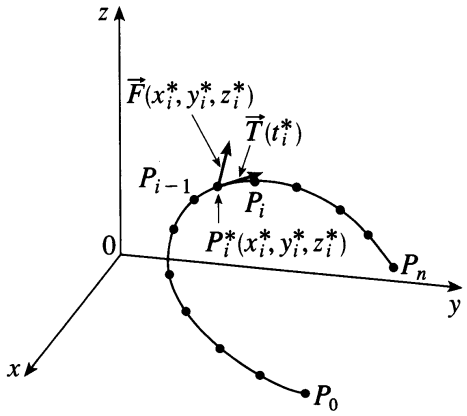
continu sur $\mathbb{R}^3$ et une courbe lisse C paramétrée par une fonction vectorielle $\overrightarrow{r}(t)$ pour $a \leq t \leq b$.

L'intégrale curviligne du champ vectoriel est l'intégrale curviligne de la projection du champ vectoriel le long de C. On introduit donc $\overrightarrow{T}(t)$ le vecteur unitaire tangent à la courbe C au point $\overrightarrow{r}(t)$.

En reprenant, les notations du paragraphe précédent l'intégrale curviligne du champ vectoriel $\overrightarrow{F}$ correspond à la limite de la somme

$$\int_C \overrightarrow{F}\cdot\overrightarrow{T}\ ds = \lim_{n\to\infty}\sum_{i=1}^{n}\overrightarrow{F}\,(x_i^*,y_i^*,z_i^*)\cdot\overrightarrow{T}\,(x_i^*,y_i^*,z_i^*)\,\Delta s_i,$$

lorsque le nombre de sous–arc n tend vers l'infini.



Si la courbe C est décrite par l'équation vectorielle $\overrightarrow{r}(t) = x(t)\overrightarrow{e}_x + y(t)\overrightarrow{e}_y + z(t)\overrightarrow{e}_z$, alors le vecteur unitaire tangent a pour expression

$$\overrightarrow{T} = \frac{\overrightarrow{r}'(t)}{||\overrightarrow{r}'(t)||}$$

et l'intégrale curviligne prend donc la forme

$$\int_C \overrightarrow{F}\cdot\overrightarrow{T}\ ds = \int_a^b \left(\overrightarrow{F}\,(\overrightarrow{r}(t))\cdot\frac{\overrightarrow{r}'(t)}{||\overrightarrow{r}'(t)||}\right)||\overrightarrow{r}'(t)||dt = \int_a^b \overrightarrow{F}\,(\overrightarrow{r}(t))\cdot\overrightarrow{r}'(t)\ dt.$$

Cette dernière forme est écrite brièvement $\int_C \overrightarrow{F}\cdot\overrightarrow{dr}$.

Définition : *Soit $\overrightarrow{F}$ un champ vectoriel continu défini sur une courbe lisse C par une fonction vectorielle $\overrightarrow{r}(t)$, $a \leq t \leq b$. Alors l'**intégrale curviligne de $\overrightarrow{F}$ le long de** C est*

$$\int_C \overrightarrow{F}\cdot d\overrightarrow{r} = \int_a^b \overrightarrow{F}\,(\overrightarrow{r}(t))\cdot\overrightarrow{r}'(t)\ dt = \int_C \overrightarrow{F}\cdot\overrightarrow{T}ds. \qquad (9.3)$$

Quand on se sert de cette définition, on se souvient que $\overrightarrow{F}\,(\overrightarrow{r}(t)))$ n'est qu'une abréviation de $\overrightarrow{F}(x(t),y(t),z(t))$ et que pour calculer $\overrightarrow{F}\,(\overrightarrow{r}(t)))$, il n'y a qu'à remplacer x par $x(t)$, y par $y(t)$, z par $z(t)$ dans l'expression de $\overrightarrow{F}(x,y,z)$. On remarque aussi qu'il est correct d'écrire $d\overrightarrow{r} = \overrightarrow{r}'(t)dt = dx\overrightarrow{e}_x + dy\overrightarrow{e}_y + dz\overrightarrow{e}_z$ ou dx, dy et dz sont des déplacements élémentaires liés puisque $d\overrightarrow{r}$ est un déplacement élémentaire **le long de** C.

9.4 Le théorème fondamentale pour les intégrales curvilignes

Rappelons en premier lieu le théorème fondamentale du calcul intégrale

$$\int_a^b F'(x)dx = F(b) - F(a);$$

où F' est une fonction continue sur $[a, b]$.

En pensant au vecteur gradient $\overrightarrow{\text{grad}}\, f$ d'une fonction comme à une sorte de dérivée de f, le théorème suivant peut être considéré comme une version du Théorème fondamentale pour les intégrales curviligne

> **Théorème :** *Soit C une courbe lisse donnée par la fonction vectorielle $\overrightarrow{r}(t)$ avec $a \leq t \leq b$. soit f une fonction différentiable dont le vecteur gradient est continu sur C. Alors*
>
> $$\int_C \overrightarrow{grad}\, f \cdot d\overrightarrow{r} = f\left(\overrightarrow{r}(b)\right) - f\left(\overrightarrow{r}(a)\right)$$

Ainsi, pour calculer l'intégrale curviligne le long de C d'un champ conservatif (champ de gradient) il suffit de connaître la valeur du potentiel aux extrémités de C.

Démonstration :

$$\begin{aligned}
\int_C \overrightarrow{\text{grad}}\, f \cdot d\overrightarrow{r} &= \int_a^b \overrightarrow{\text{grad}}\, f\left(\overrightarrow{r}(t)\right) \cdot \overrightarrow{r}'(t)dt \\
&= \int_a^b \left(\frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} + \frac{\partial f}{\partial z}\frac{dz}{dt}\right) dt \\
&= \int_a^b \frac{d}{dt}\left(f\left(\overrightarrow{r}(t)\right) dt\right) \\
&= f\left(\overrightarrow{r}(b)\right) - f\left(\overrightarrow{r}(a)\right)
\end{aligned}$$

La dernière étape découle du théorème fondamental du calcul intégrale.

Indépendance du chemin On suppose que C_1 et C_2 sont deux courbes lisses (appelées aussi chemins) qui toutes deux relient le point A, dit point initial, au point B, dit point terminal. On sait qu'en générale $\int_{C_1} \overrightarrow{F} \cdot d\overrightarrow{r} \neq \int_{C_2} \overrightarrow{F} \cdot d\overrightarrow{r}$. Mais, par ailleurs, une des conséquence du Théorème fondamentale est que

$$\int_{C_1} \overrightarrow{\text{grad}}\, f \cdot d\overrightarrow{r} = \int_{C_2} \overrightarrow{\text{grad}}\, f \cdot d\overrightarrow{r},$$

à condition que $\overrightarrow{\text{grad}}\, f$ soit continu. Autrement dit, l'intégrale curviligne d'un champ conservatif ne dépend que du point initial et du point final d'une courbe. En générale, si $\overrightarrow{F}$ est un cahmp vectoriel continu sur un domaine D, on dit que l'intégrale curviligne $\int_C \overrightarrow{F} \cdot d\overrightarrow{r}$ est **indépendante du chemin** si $\int_{C_1} \overrightarrow{F} \cdot d\overrightarrow{r} = \int_{C_2} \overrightarrow{F} \cdot d\overrightarrow{r}$, quels que soient les chemins C_1 et C_2 dans D qui relient les même points. Dans ce vocabulaire, on peut dire que *les intégrales curvilignes des champs vectoriels conservatifs sont indépendantes du chemin.*

Une courbe est dite **fermée** si sont point terminal coïncide avec son point initial. On suppose une intégrale curviligne indépendante du chemin et un chemin fermé quelconque C. En décomposant le chemin C en deux sous chemin C_1 et C_2, on peut écrire

$$\int_C \overrightarrow{F} \cdot d\overrightarrow{r} = \int_{C_1} \overrightarrow{F} \cdot d\overrightarrow{r} + \int_{C_2} \overrightarrow{F} \cdot d\overrightarrow{r} = \int_{C_1} \overrightarrow{F} \cdot d\overrightarrow{r} - \int_{-C_2} \overrightarrow{F} \cdot d\overrightarrow{r} = 0,$$

puisque C_1 et $-C_2$ partent et finissent aux mêmes points. On en déduit le théorème suivant

Théorème : *L'intégrale* $\int_C \overrightarrow{F} \cdot d\overrightarrow{r}$ *est indépendante du chemin d'intégration dans* D *si et seulement si* $\int_C \overrightarrow{F} \cdot d\overrightarrow{r} = 0$ *sur tout chemin fermé* C *de* D.

Le théorème suivant affirme que *seul* un champ vectoriel qui est indépendant du chemin est conservatif. On suppose que D est **ouvert**, ce qui signifie que pour chaque point P de D, il existe une boule centrée en P entièrement inclus dans D ; de cette façon, D ne contient aucun point de sa frontière. De plus il est supposé que D est **connexe**. Cela signifie que deux point quelconques de D peuvent être reliés par un chemin entièrement contenu dans D.

Théorème : *Soit* $\overrightarrow{F}$ *un champ vectoriel continu sur une région ouverte et connexe* D. *Si* $\int_C \overrightarrow{F} \cdot d\overrightarrow{r}$ *est indépendante du chemin dans* D, *alors* $\overrightarrow{F}$ *est un champ conservatif sur* D ; *autrement dit, il existe une fonction* f *telle que* $\overrightarrow{F} = \overrightarrow{grad}\, f$.

Une question subsiste : comment déterminer si un champ vectoriel est conservatif ? Voyons cela dans $\mathbb{R}^2$. On suppose un champ vectoriel conservatif $\overrightarrow{F} = P\overrightarrow{e}_x + Q\overrightarrow{e}_y$, où P et Q ont des dérivées partielles premières continues. Alors, il existe une fonction f telle que $\overrightarrow{F} = \overrightarrow{\text{grad}}\, f$, c–à–d. $P = \partial f/\partial x$ et $Q = \partial f/\partial y$. De la il vient

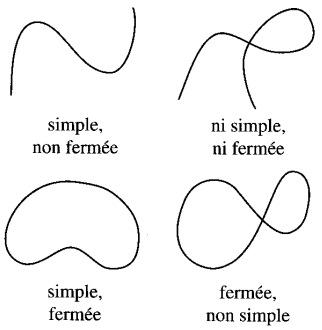
$$\frac{\partial P}{\partial y} = \frac{\partial f}{\partial y \partial x} = \frac{\partial Q}{\partial x}.$$

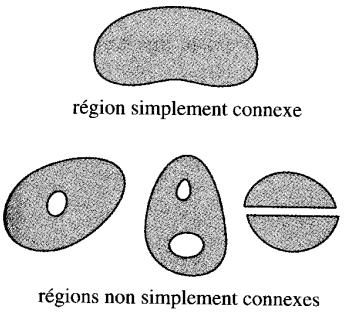
Théorème : *Si* $\overrightarrow{F}(x,y) = P(x,y)\overrightarrow{e}_x + Q(x,y)\overrightarrow{e}_y$ *est un champ vectoriel conservatif, où* P *et* Q *ont des dérivées partielles d'ordre 1, continue sur un domaine* D, *alors, sur tout* D, *on a*

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}$$

La réciproque, qui est le résultat qui nous intéresse, n'est vraie que sur un type particulier de région. En vue d'expliquer cela, on a besoin du concept de **courbe simple**, qui signifie que la courbe ne se recoupe nulle part entre ses extrémités (voir figure).

Le théorème précédent exigeait déjà une région ouverte connexe. Le théorème suivant se montre plus exigeant encore. Une région du plan **simplement connexe** est une région D telle que toute courbe simple fermée dans D n'entoure que des points de D (voir les figures ci–dessous).





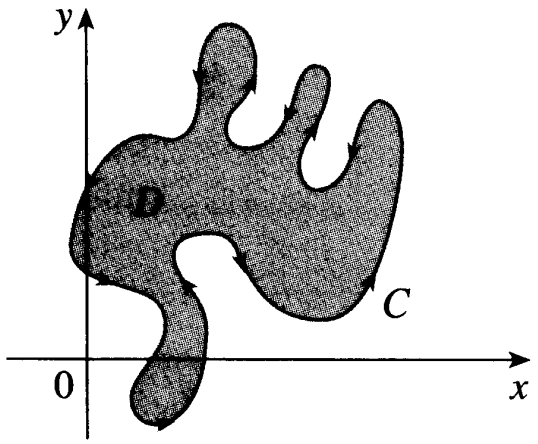
Théorème : *Soit* $\overrightarrow{F}(x,y) = P(x,y)\overrightarrow{e}_x + Q(x,y)\overrightarrow{e}_y$ *un champ vectoriel défini sur une région simplement connexe* D. *On suppose que* P *et* Q *ont des dérivées partielles d'ordre 1 continues sur* D *et que*

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x} \quad \text{sur tout } D.$$

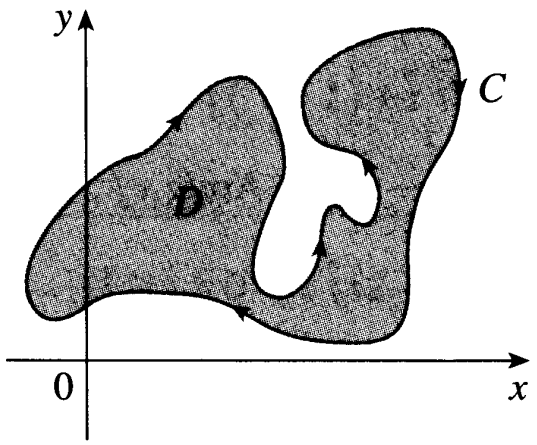
Alors $\overrightarrow{F}$ *est conservatif.*

9.5 Le théorème de Green

Le Théorème de Green établit une relation entre une intégrale curviligne le long d'un contour simple fermée C et une intégrale double sur le domaine D du plan délimité par C. L'énoncé du théorème de Green est basé sur un sens de parcours conventionnel : sera considéré **dans le sens positif** le parcours *en sens inverse des aiguilles d'une montre* de C. Si C est décrite par la fonc tion vectorielle $\overrightarrow{r}(t)$, $a \leq t \leq b$, la région D est donc toujours à gauche du point $\overrightarrow{r}(t)$ lorsqu'il parcourt C (voir figure)



a) Sens positif



b) Sens négatif

Théorème de Green : *Soit C une courbe plane simple, fermée, lisse par morceaux, orientée dans le sens positif et soit D le domaine délimité par C. Si P et Q sont pourvues de dérivées partielles continues dans une région ouverte qui contient D, alors*

$$\oint_C P\,dx + Q\,dy = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA \tag{9.4}$$

On utilise la notation $\oint_C P\,dx + Q\,dy$ pour indiquer que l'intégrale curviligne est calculée selon le sens positif de la courbe fermée C. On rencontre aussi la notation ∂D pour désigner la frontière du domaine D orientée dans le sens positif, cela donne au théorème de Green l'écriture suivant

$$\oint_{\partial D} P\,dx + Q\,dy = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA$$

Le Théorème de Green est à regarder comme l'homologue du Théorème fondamental du calcul intégral pour les intégrales doubles.

Exercice : Calculez $\int_C x^4\,dx + xy\,dy$, où C est la courbe triangulaire composée des trois segments de $(0,0)$ à $(0,1)$, de $(0,1)$ à $(1,0)$ et de $(1,0)$ à $(0,0)$.

9.6 La divergence et le rotationnel

Dans cette section sont définies deux opérations applicables à des champs vectoriels et qui jouent un rôle primordiale dans les applications de l'analyse vectorielle à l'étude de l'écoulement des fluides, de l'électricité et du magnétisme.

9.6.1 Le rotationnel

Définition : *Si* $\overrightarrow{F}(x,y,z) = P(x,y,z)\overrightarrow{e}_x + Q(x,y,z)\overrightarrow{e}_y + R(x,y,z)\overrightarrow{e}_z$ *est un champ vectoriel défini sur* $\mathbb{R}^3$ *et si les dérivées partielles de* P, Q *et* R *existent, alors le rotationnel de* $\overrightarrow{F}$ *est le champ de vectoriel défini sur* $\mathbb{R}^3$ *par*

$$\overrightarrow{rot}\,\overrightarrow{F} = \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\right)\overrightarrow{e}_x + \left(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\right)\overrightarrow{e}_y + \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\overrightarrow{e}_z$$

Afin de faciliter la mémorisation, on peut récrire l'expression à l'aide de l'opérateur différentiel $\overrightarrow{\nabla}$

$$\overrightarrow{\nabla} = \overrightarrow{e}_x\,\frac{\partial}{\partial x} + \overrightarrow{e}_y\,\frac{\partial}{\partial y} + \overrightarrow{e}_z\,\frac{\partial}{\partial z}.$$

Appliqué à une fonction scalaire f, $\overrightarrow{\nabla}$ a pour effet de produire le gradient de f. On peut envisager le rotationnel de $\overrightarrow{F}$ comme le produit vectoriel

$$\overrightarrow{\text{rot}}\,\overrightarrow{F} = \overrightarrow{\nabla} \wedge \overrightarrow{F}$$

Théorème : *Si* f *est une fonction de trois variables qui a des dérivées secondes partielles continues, alors*

$$\overrightarrow{rot}\left(\overrightarrow{\nabla} f\right) = \overrightarrow{0}.$$

Théorème : *Si* $\overrightarrow{F}$ *est un champ vectoriel défini sur tout* $\mathbb{R}^3$*, dont les fonctions composantes ont des dérivées partielles continues et si* $\overrightarrow{rot}\,\overrightarrow{F} = \overrightarrow{0}$*, alors* $\overrightarrow{F}$ *est un champ conservatif.*

9.6.2 La divergence

Définition : *Si* $\overrightarrow{F}(x,y,z) = P(x,y,z)\overrightarrow{e}_x + Q(x,y,z)\overrightarrow{e}_y + R(x,y,z)\overrightarrow{e}_z$ *est un champ vectoriel défini sur* $\mathbb{R}^3$ *et si les dérivées partielles de* P, Q *et* R *existent, alors la* ***divergence*** *de* $\overrightarrow{F}$ *est la fonction de trois variables définie par*

$$div\,\overrightarrow{F} = \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}.$$

En utilisant $\overrightarrow{\nabla}$ *on peut encore écrire*

$$div\,\overrightarrow{F} = \overrightarrow{\nabla} \cdot \overrightarrow{F}.$$

Il est important de retenir que le rotationnel est un champ vectoriel tandis que la divergence est un champ scalaire.

Théorème : *Si* $\overrightarrow{F}(x,y,z) = P(x,y,z)\overrightarrow{e}_x + Q(x,y,z)\overrightarrow{e}_y + R(x,y,z)\overrightarrow{e}_z$ *est un champ vectoriel défini sur* $\mathbb{R}^3$ *et si les dérivées secondes partielles de* P, Q *et* R *sont continue, alors*

$$div\left(\overrightarrow{rot}\,\overrightarrow{F}\right) = 0$$

9.6.3 Les formes vectorielles du théorème de Green

On suppose que le domaine D du plan, sa frontière C et les fonctions P et Q satisfont aux hypothèses du Théorème de Green. On considère alors le champs vectoriel $\overrightarrow{F} = P\ \overrightarrow{e}_x + Q\ \overrightarrow{e}_y$ comme un champ vectoriel de $\mathbb{R}^3$ ayant une troisième composante identiquement nulle. Il en suit que le rotationnel de $\overrightarrow{F}$ a pour expression

$$\overrightarrow{\text{rot}}\ \overrightarrow{F} = \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\ \overrightarrow{e}_z$$

On peut donc écrire le Théorème de Green sous la forme :

$$\oint_C \overrightarrow{F} \cdot d\overrightarrow{r} = \iint_D \overrightarrow{\text{rot}}\ \overrightarrow{F} \cdot \overrightarrow{e}_z\ dA$$

On peut aussi établir une expression analogue pour la composant normale de $\overrightarrow{F}$ en introduisant le vecteur $\overrightarrow{n}$ unitaire et orienté selon la normale extérieur du contour C

$$\overrightarrow{n}(t) = \frac{\overrightarrow{r}'(t)}{||\overrightarrow{r}'(t)||}.$$

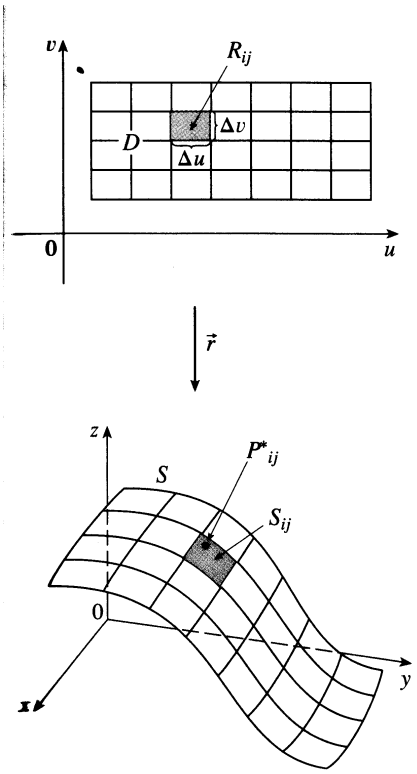
$$\oint_C \overrightarrow{F} \cdot \overrightarrow{n} ds = \iint_D \text{d}iv\ \overrightarrow{F}\ dA$$

Cette version dit que l'intégrale curviligne de la composante normale de $\overrightarrow{F}$ le long de C est égale à l'intégrale double de la divergence de $\overrightarrow{F}$ sur le domaine D délimité par C.

9.7 Les intégrales de surface

Les intégrales de surface sont à l'aire d'une surface ce que les intégrales curvilignes sont à l'abscisse curviligne. On suppose de f est une fonction de trois variables définie sur un domaine qui inclut une surface S. On va définir l'intégrale de surface de f sur S de telle sorte que, au cas où $f(x, y, z) = 1$, la valeur de l'intégrale de surface soit égale à l'aire de S.

9.7.1 Les surfaces paramétrées



On suppose la surface S décrite par la fonction vectorielle

$$\overrightarrow{r}(u, v) = x(u, v)\ \overrightarrow{e}_x + y(u, v)\ \overrightarrow{e}_y + z(u, v)\ \overrightarrow{e}_z \quad (u, v) \in D.$$

On fait l'hypothèse que les paramètres varient dans un domaine rectangulaire qu'on divise en sou-rectangle R_{ij} de dimension Δu et Δv. Alors, la surface est divisée en éléments d'aire $\Delta S_{i,j}$. On calcul la valeur de f en un point P^*_{ij} de chaque éléments de surface, on multiplie cette valeur par l'aire ΔS_{ij} et on forme la somme de Riemann

$$\sum_{i=0}^{m}\sum_{j=0}^{n} \Delta P^*_{ij}\ S_{ij}.$$

On prend ensuite la limite lorsque le nombre d'éléments de surface croît sans borne et on défini l'**intégrale de surface de f sur S**

$$\iint_s f(x, y, z)\ dS = \lim_{n,m\to\infty} \sum_{i=0}^{m}\sum_{j=0}^{n} \Delta P^*_{ij}\ S_{ij}.$$

Afin de calculer l'intégrale de surface, on remplace l'aire de l'élément de surface ΔS_{ij} par celle, qui lui est approximativement égale, d'un parallélogramme du plan tangent :

$$\Delta S_{ij} = ||\overrightarrow{r}'_u \wedge \overrightarrow{r}'_v||\Delta u \Delta v$$

où

$$\overrightarrow{r}'_u = \frac{\partial x}{\partial u}\ \overrightarrow{e}_x + \frac{\partial y}{\partial u}\ \overrightarrow{e}_y + \frac{\partial z}{\partial u}\ \overrightarrow{e}_z \quad \text{et} \quad \overrightarrow{r}'_v = \frac{\partial x}{\partial v}\ \overrightarrow{e}_x + \frac{\partial y}{\partial v}\ \overrightarrow{e}_y + \frac{\partial z}{\partial v}\ \overrightarrow{e}_z.$$

sont les vecteurs tangents en un sommet de S_{ij}. Si les composantes sont continues et si les vecteurs $\overrightarrow{r}'_u$ et $\overrightarrow{r}'_v$ ne sont ni nuls ni parallèles à l'intérieur de D, on peut montrer, même si D n'est pas rectangulaire, que

$$\iint_S f(x, y, z)\ dS = \iint_D f\left(\overrightarrow{r}(u, v)\right) ||\overrightarrow{r}'_u \wedge \overrightarrow{r}'_v||dA.$$

Remarque : Toute surface S d'équation $z = f(x, y)$ peut–être vue comme une surface paramétrée décrite par

$$\overrightarrow{r}(u, v) = u\ \overrightarrow{e}_x + v\ \overrightarrow{e}_y + f(u, v)\ \overrightarrow{e}_z \quad (u, v) \in D.$$

9.7.2 Les surfaces orientables

Avant de définir les intégrales de surface d'un champ vectoriel, il est nécessaire d'écarter les surfaces non orientables dont le ruban de Möbius est un exemple. Une surface orientable est une surface présentant deux faces distincts. S'il est possible de choisir un vecteur unitaire normal $\overrightarrow{n}$ de telle sorte que $\overrightarrow{n}$ varie continûment sur S, alors S est appelé une surface orientable et le choix de $\overrightarrow{n}$ dote S d'une orientation. Toute surface orientable à deux orientations possibles.

Au cas ou S est une surface lisse orientable décrite paramétriquement par une fonction vectorielle $\overrightarrow{r}(u,v)$, alors elle est automatiquement dotée d'une orientation par le vecteur unitaire normal

$$\overrightarrow{n} = \frac{\overrightarrow{r}'_u \wedge \overrightarrow{r}'_v}{||\overrightarrow{r}'_u \wedge \overrightarrow{r}'_v||}$$

et de l'orientation opposé $-\overrightarrow{n}$.

Dans le cas d'une surface fermée, c–à–d. une surface qui constitue la cloison d'une région solide E, il est convenu que l'orientation positive est celle du vecteur normal dirigé vers l'extérieur de E

Exercice : Déterminer le vecteur normal d'une sphère de rayon a.

9.7.3 Les intégrales de surface de champs de vecteurs

Définition : *Si $\overrightarrow{F}$ est un champ vectoriel défini sur une surface orientée S de vecteur unitaire normal $\overrightarrow{n}$, alors l'**intégrale de $\overrightarrow{F}$ sur** S est*

$$\iint_S \overrightarrow{F} \cdot d\overrightarrow{S} = \iint_S \overrightarrow{F} \cdot \overrightarrow{n}\, dS.$$

*Cette intégrale est aussi appelée le **flux** du champ $\overrightarrow{F}$ à travers S.*

Si c'est une fonction vectorielle $\overrightarrow{r}(u,v)$ qui décrit S, et D le domaine paramétré, alors

$$\iint_S \overrightarrow{F} \cdot d\overrightarrow{S} = \iint_D \overrightarrow{F} \cdot (\overrightarrow{r}'_u \wedge \overrightarrow{r}'_v)\, dA.$$

Exercice : Calculer le flux du champ vectoriel $\overrightarrow{F}(x,y,z) = x\overrightarrow{e}_x + y\overrightarrow{e}_y + z\overrightarrow{e}_z$ à travers la sphère de rayon unité.

9.8 Les théorèmes de Stokes et d'Ostrogradskii

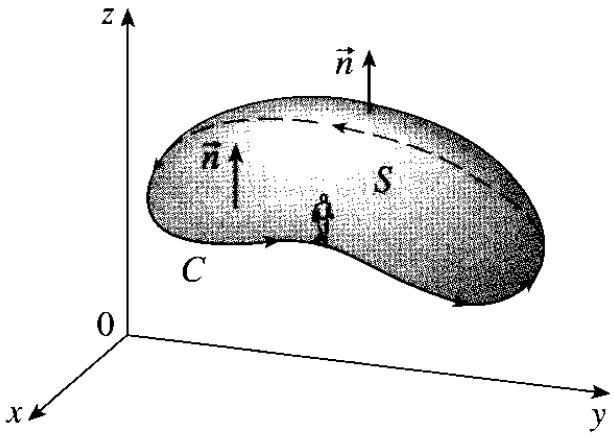


FIGURE 1

Le Théorème de Stokes peut être vu comme une version du Théorème de Green de dimension supérieure. Alors que le Théorème de Green met en relation une intégrale double sur un domaine du plan D avec une intégrale curviligne autour de sa courbe frontière, le Théorème de Stokes lie une intégrale de surface sur une surface S (à priori non plane) à une intégrale curviligne autour de la courbe frontière de S (qui est une courbe de l'espace). La figure ci–contre montre une surface orientée par un vecteur unitaire normal $\overrightarrow{n}$. L'orientation de la surface S détermine le **sens positif de la courbe frontière** C, indiqué par des flèches dans la figure. En effet, si vous marchez sur C dans le sens positif, la tête dans la direction de $\overrightarrow{n}$, la surface sera toujours à votre gauche.

Théorème de Stokes : *Soit S une surface orientée, lisse par morceaux, dont le bord est une courbe C simple et fermée, lisse par morceaux, orientée positivement. Soit $\overrightarrow{F}$ un champ de vecteurs dont les composantes ont des dérivées partielles continues sur une région ouverte de $\mathbb{R}^3$ qui contient S. Alors*

$$\oint_C \overrightarrow{F} \cdot d\overrightarrow{r} = \iint_S \overrightarrow{rot}\, \overrightarrow{F} \cdot d\overrightarrow{S}.$$

Dans la section F, le Théorème de Green a été écrit sous la forme vectorielle

$$\oint_C \overrightarrow{F} \cdot \overrightarrow{n}\, ds = \iint_D \mathrm{d}iv\ \overrightarrow{F}\ dA,$$

où C est la frontière, orientée positivement, du domaine D. Le Théorème d'Ostrogradskii (encore appelé Théorème flux—divergence) généralise cette idée en mettant un relation le flux d'un champ de vecteur a travers une surface fermée et l'intégrale volumique de la divergence du champ de vecteur

Théorème d'Ostrogradskii : *Soit E une région solide simple et soit S la surface fermée frontière de E, orientée positivement (vers l'extérieur). Soit $\overrightarrow{F}$ un champ de vecteurs dont les composantes ont des dérivées partielles continus sur une région ouverte de $\mathbb{R}^3$ qui contient E. Alors*

$$\iint_S \overrightarrow{F} \cdot d\overrightarrow{S} = \iiint_E div\ \overrightarrow{F}\, dV$$

Deuxième partie

Analyse numérique élémentaires

Chapitre 10

Compléments d'algèbre linéaire

Sommaire

L'étude d'un système physique requiert en général la collection de plusieurs données (position, vitesse, pression, température...) pour caractériser de manière univoque l'état du système. Les lois physiques sont les relations entre ces différentes quantités qui permettent de prédire le comportement futur du système connaissant l'état initial. La représentation et la manipulation de cette collection de données obligent à travailler dans un espace à grand nombre de dimensions alors que notre esprit est habitué à l'espace "matériel" à trois dimensions. Cependant, les outils utilisés pour appréhender l'espace matériel (mesure, distance, projection, orientation dans l'espace...) se généralisent aux espaces de plus grande dimension et permettent de travailler dans ces espaces "exotiques".

Le but de ce chapitre est de rappeler les définitions et les utilisations de ces différents outils, en focalisant sur la projection et la métrique qui sont parmi les outils les plus précieux du physicien.

10.1 Espace vectoriel

Supposons que l'état d'un système physique soit défini de manière univoque par une collection de N nombres réels $(x_1, x_2, \ldots, x_n)$. L'état du système est donc caractérisé dans un espace à n-dimensions. Par analogie à l'espace "matériel" à 3D, on appelle **vecteur** une telle collection de données et les membres x_i de la collection se nomment **composantes** du vecteur. On notera $\overrightarrow{u} = (x_1, x_2, \ldots, x_n)$ un vecteur de l'espace. Dans la pratique les x_i ne sont pas toujours réels, il peuvent être des entiers des chaînes de caractère, etc.

On va supposer pour l'instant que l'on peut faire au moins deux types d'opérations sur les vecteurs de notre espace : l'addition de deux vecteurs et la multiplication par un scalaire. Remarquez que l'on pourrait inventer bien d'autres opérations

— **Addition.** On définit l'addition de deux vecteurs comme l'addition composante par composante. Soit $\overrightarrow{u} = (x_1, x_2, \ldots, x_n)$ et $\overrightarrow{v} = (y_1, y_2, \ldots, y_n)$ deux vecteurs de notre espace à N-dimensions, on note $\overrightarrow{w}$ le résultat de l'addition de $\overrightarrow{u}$ à $\overrightarrow{v}$:

$$\overrightarrow{w} = \overrightarrow{u} + \overrightarrow{v} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)\,.$$

— **Multiplication.** La multiplication d'un vecteur $\overrightarrow{u} = (x_1, x_2, \ldots, x_N)$ par un nombre réel λ correspond à la multiplication de chacune de ses composantes par λ :

$$\overrightarrow{q} = \lambda\overrightarrow{u} = (\lambda x_1, \lambda x_2, \ldots, \lambda x_n)$$

Bien entendu, pour que ces opérations soit bien définies, il est nécessaire que leurs résultats soient contenus dans l'espace de départ, c'est-à-dire que $\overrightarrow{w}$ et $\overrightarrow{q}$ appartiennent au même espace que $\overrightarrow{u}$. Cela est vérifié si les composantes, comme λ, peuvent prendre toutes les valeurs réelles ou réelles positives, mais pas si les valeurs doivent être prises dans $[0, 2]$. On dit que ces lois sont internes.

Un espace muni de ces deux opérations se nomme un **espace vectoriel sur** $\mathbb{R}$ si les $x_i \in \mathbb{R}$. Notre espace matériel est un espace vectoriel dont les vecteurs sont les vecteurs positions utilisés en géométrie et en mécanique.

En appliquant les propriétés bien connues de l'addition de la multiplication des réelles, on redémontre sans peine les relations suivantes

— $\overrightarrow{u} + \overrightarrow{v} = \overrightarrow{v} + \overrightarrow{u}$
— $\lambda(\overrightarrow{u} + \overrightarrow{v}) = \lambda\overrightarrow{u} + \lambda\overrightarrow{v}$
— $(\overrightarrow{u} + \overrightarrow{v}) + \overrightarrow{w} = \overrightarrow{u} + (\overrightarrow{v} + \overrightarrow{w})$

Si l'on peut définir la soustraction de deux vecteurs $\overrightarrow{u} - \overrightarrow{v} = \overrightarrow{u} + (-1)\overrightarrow{v}$, la division, elle, n'existe pas.

Les résultats précédents se généralisent sans difficulté si les composantes sont des nombres complexes, on parlera alors d'**espace vectoriel sur** $\mathbb{C}$

10.2 Distance et norme

Distance

En mathématiques, une métrique ou distance est une fonction qui définit la distance entre les éléments d'un ensemble. Un ensemble muni d'une distance est appelé un **espace métrique**.

> **Définition :** *Une distance sur un ensemble E est une fonction de E^2 dans l'ensemble $\mathbb{R}$ des nombres réels : $d : E \times E \to \mathbb{R}$ qui satisfait les conditions suivantes pour tous x, y, z dans E*
>
> 1. $d(x, y) = 0$ *si et seulement si* $x = y$ *(identité des indiscernables),*
> 2. $d(x, y) = d(y, x)$ *(symétrie),*
> 3. $d(x, z) \leq d(x, y) + d(y, z)$.

Des propriétés 1, 2 et 3 il découle que d est à valeurs positives : pour tous x et y dans X, $d(x, y) \geq 0$.

Ces conditions expriment les notions intuitives du concept de distance. Par exemple, que la distance entre des points distincts est strictement positive et que la distance de x à y est la même que la distance de y à x. L'inégalité triangulaire signifie que la distance parcourue directement entre x et z, n'est pas plus grande que la distance à parcourir en partant d'abord de x vers y puis de y vers z.

Exemple : La distance triviale (ou encore distance discrète ou métrique discrète) : sur un ensemble non vide, on décide que la distance entre deux points distincts est 1 ($d(x, y) = 1$) pour tout x différent de y et 0 tout x égal à y ($d(x, x) = 0$).

Norme

En géométrie, la norme est une extension de la valeur absolue des nombres aux vecteurs. Elle permet de mesurer la longueur commune à toutes les représentations d'un vecteur dans un espace, mais définit aussi une distance entre deux vecteurs invariante par translation. La norme usuelle dans le plan ou l'espace est dite euclidienne.

D'autres normes sont très utilisées sur les espaces vectoriels de dimension finie ou infinie, appelés alors espaces vectoriels normés. Elles sont notamment très importantes en analyse fonctionnelle pour obtenir des majorations, exprimer la différentiation sur les espaces de fonctions d'une ou plusieurs variables réelles ou complexes, calculer estimations et approximations.

Définition : *Une norme sur un espace E est une application $\mathcal{N}$ sur E à valeurs réelles positives et satisfaisant les hypothèses suivantes :*

- *séparation :* $\forall x \in E,\ \mathcal{N}(x) = 0 \Rightarrow x = 0_E$;
- *homogénéité :* $\forall (\lambda, x) \in \mathbb{K} \times E,\ \mathcal{N}(\lambda \cdot x) = |\lambda| \mathcal{N}(x)$;
- *sous-additivité (appelé également Inégalité triangulaire) :* $\forall (x, y) \in E^2,\ \mathcal{N}(x + y) \leq \mathcal{N}(x) + \mathcal{N}(y)$.

La norme permet de définir une distance.

Propriété : *Étant donné un espace vectoriel normé $(E, ||\ ||)$ on peut définir une distance sur E par $d(x, y) =|| y - x ||$. La distance d est dite "induite par" la norme $||\ ||$.*

L'inverse est aussi possible.

Propriété : *Si une distance d sur un espace vectoriel E satisfait les propriétés*

1. $d(x, y) = d(x + a, y + a))$ *(invariance par translation)*
2. $d(\alpha x, \alpha y) =| \alpha | d(x, y))$ *(homogénéité)*

alors, on peut définir une norme sur E par $|| x ||= d(x, 0)$

Par exemple, si on considère le vecteur $\overrightarrow{x} = (x_1, x_2, \ldots, x_n)$ de $\mathbb{R}^n$, on peut alors définir

- la norme euclidienne $|| \overrightarrow{x} ||_2= \sqrt{| x_1 |^2 + \cdots + | x_n |^2}$ et elle correspond à la norme habituellement utilisée pour la distance entre deux points dans le plan ou l'espace usuels.
- la norme 1 est donnée par la somme des modules (ou valeurs absolues) des coefficients $|| \overrightarrow{x} ||_1=| x_1 | + \cdots + | x_n |$ et induit la distance de déplacement à angle droit sur un damier, dite distance de Manhattan.
- plus généralement, pour tout p supérieur ou égal à 1, la norme p est donnée par la formule suivante $|| \overrightarrow{x} ||_p= (| x_1 |^p + \cdots + | x_n |^p)^{1/p}$ elle identifie donc la norme euclidienne avec la norme 2, mais n'a surtout d'intérêt que dans sa généralisation aux espaces de fonctions,
- la norme « infinie » d'un vecteur est la limite de ses normes p lorsque p tend vers l'infini $|| \overrightarrow{x} ||_\infty= \lim_{p\to\infty} || \overrightarrow{x} ||_p= \max(| x_1 |, \ldots, | x_n |)$.

Dans la suite, le symbole $|| \cdot ||$ représentera, sauf indication contraire, la norme euclidienne.

Exemple d'utilisation de la norme : la continuité

Supposons une fonction f prenant comme argument un vecteur $\overrightarrow{u}$ d'un espace vectoriel et dont le résultat est un vecteur $\overrightarrow{v}$. On dit que f envoie $\overrightarrow{u}$ sur $\overrightarrow{v}$. Il est souvent utile d'être renseigné sur la continuité de la fonction f or celle-ci se définit précisément à l'aide de la norme. On dit d'une fonction f qu'elle est continue en $\overrightarrow{u}_0$ si

$$\forall \epsilon > 0 \quad \delta > 0 \quad \text{tel que} \quad ||\overrightarrow{u} - \overrightarrow{u}_0|| < \delta \quad ||f(\overrightarrow{u}) - f(\overrightarrow{u}_0)|| < \epsilon.$$

Exercice : Après avoir identifié les espaces vectoriels de départ et d'arrivée de la fonction $f(x_1, x_2) = \frac{x_1 x_2}{x_1^2 + x_2^2}$; étudier sa continuité en $(0, 0)$.

La norme sera utile lors des études ultérieures de la convergence d'une méthode numérique.

10.3 Produit scalaire

En géométrie vectorielle, le produit scalaire est une opération algébrique s'ajoutant aux lois s'appliquant aux vecteurs. À deux vecteurs elle associe leur produit, qui est un nombre réel (ou scalaire). Elle permet d'exploiter les notions de la géométrie euclidienne traditionnelle : longueurs, angles, orthogonalité en dimension deux et trois, mais aussi de les étendre à des espaces vectoriels réels de toute dimension, et aux espaces vectoriels complexes.

Comme il existe deux grandes manières de définir les vecteurs, soit par une approche purement algébrique, soit par une approche géométrique à l'aide des bi–points (ou couple ordonné de points), il existe de mêmes deux manières de présenter le produit scalaire : une manière algébrique, et une manière géométrique, à l'aide de bi–points.

Le produit scalaire possède de multiples applications. En physique, il est, par exemple, utilisé pour modéliser le travail d'une force. En géométrie analytique il permet de déterminer le caractère perpendiculaire de deux droites ou d'une droite et d'un plan. Dans le cas de la dimension finie quelconque, il dispose de nombreuses applications algébriques : il offre des outils pour la réduction d'endomorphismes (diagonalisation de matrice) ou encore est à la base de multiples techniques statistiques comme la méthode des moindres carrés ou l'analyse en composantes principales. En géométrie, il confère à l'espace vectoriel une structure d'**espace métrique** disposant de nombreuses propriétés comme la complétude. Le produit scalaire est aussi utilisé dans des espaces de dimension infinie, il permet alors de résoudre des équations aux dérivées partielles.

Pour un espace vectoriel de dimension n sur $\mathbb{R}$, le produit scalaire usuel (on dit plutôt "euclidien"), noté $\langle \overrightarrow{u} | \overrightarrow{v} \rangle$, associe aux vecteurs $\overrightarrow{u} = (x_1, x_2, \ldots, x_n)$ et $\overrightarrow{v} = (y_1, y_2, \ldots, y_n)$ est le nombre

$$\langle \overrightarrow{u} | \overrightarrow{v} \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

Le produit scalaire d'un vecteur par lui-même étant alors le carré de la norme euclidienne :

$$||u||^2 = \langle \overrightarrow{u} | \overrightarrow{u} \rangle = x_1^2 + x_2^2 + \cdots + x_n^2$$

Cependant, ce n'est pas l'unique façon d'associer à un vecteur un scalaire correspondant à l'idée usuelle que l'on se fait d'une norme, il y a donc lieu de généraliser

Définition : *Si H est un espace vectoriel sur $\mathbb{C}$, un produit scalaire est une application de $H \times H$ dans $\mathbb{C}$ possédant les propriétés suivantes*
- *sesquiliniéarité* $\langle \lambda \overrightarrow{u} | \overrightarrow{v} \rangle = \lambda^\dagger \langle \overrightarrow{u} | \overrightarrow{v} \rangle \qquad \langle \overrightarrow{u} | \lambda \overrightarrow{v} \rangle = \lambda \langle \overrightarrow{u} | \overrightarrow{v} \rangle$;
- *symétrie hermitienne* $\langle \overrightarrow{u} | \overrightarrow{v} \rangle = \langle \overrightarrow{v} | \overrightarrow{u} \rangle^\dagger$;
- *Le produit scalaire est défini positif* $\langle \overrightarrow{u} | \overrightarrow{u} \rangle \geq 0$

Le produit scalaire permet alors de définir une norme $|| \overrightarrow{u} || = \langle \overrightarrow{u} | \overrightarrow{u} \rangle^{1/2}$.

Définition : *Deux vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$ sont dits orthogonaux si, et seulement si,* $\langle \overrightarrow{u} | \overrightarrow{v} \rangle = 0$.

10.4 Projecteur

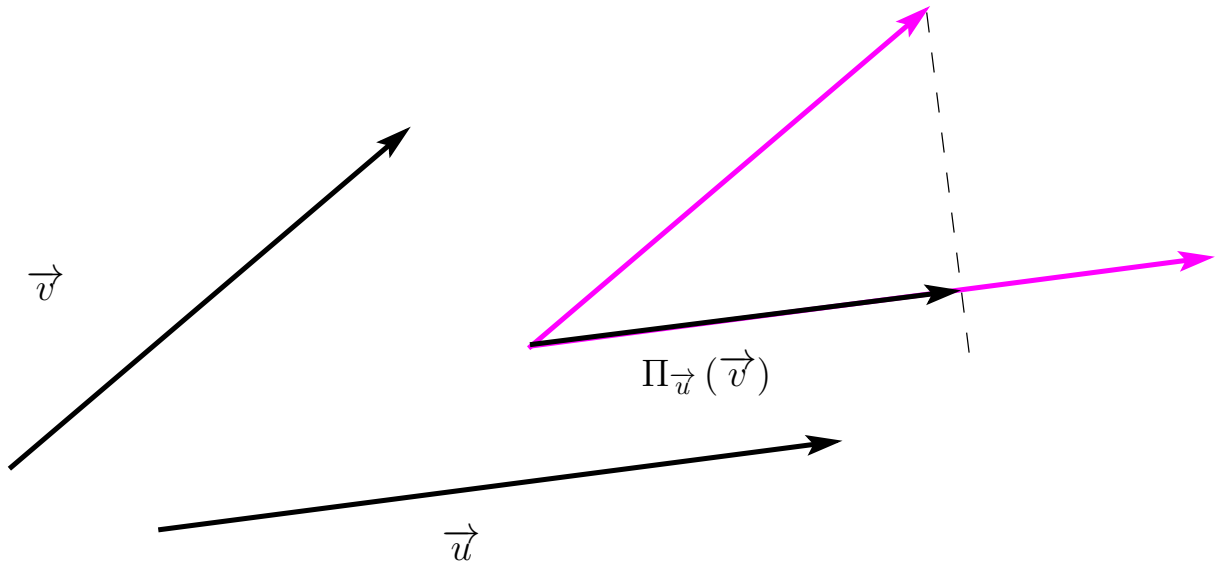
Un projecteur (ou une projection) est une application qu'on peut présenter de deux façons équivalentes
- c'est une projection linéaire associée à une décomposition de E comme somme de deux sous-espaces supplémentaires, c'est-à-dire qu'elle permet d'obtenir un des termes de la décomposition correspondante,
- c'est aussi un endomorphisme idempotent : il vérifie $p^2 = p$

Définition : *Dans un espace muni d'un produit scalaire, une projection pour laquelle les deux supplémentaires sont orthogonaux est appelée projection orthogonale.*

Considérons par exemple deux vecteurs de l'espace E à n dimensions $\overrightarrow{u}$ et $\overrightarrow{v}$ et appelons $\Pi_{\overrightarrow{u}}(\overrightarrow{v})$ la projection orthogonale de $\overrightarrow{v}$ suivant le vecteur $\overrightarrow{u}$. Le produit scalaire permet de définir cette projection

$$\Pi_{\overrightarrow{u}}(\overrightarrow{v}) = \frac{1}{\langle \overrightarrow{u} \mid \overrightarrow{u} \rangle} \langle \overrightarrow{u} | \overrightarrow{v} \rangle \overrightarrow{u}$$

La division par le carré de la norme euclidienne garantit l'idempotence de l'application. La figure suivante illustre la notion de projecteur en dimension 2.



10.5 Base d'un espace vectoriel

En mathématiques, et plus particulièrement en algèbre linéaire, une base d'un espace vectoriel est une famille de vecteurs de cet espace, telle que chaque vecteur de l'espace puisse être exprimé de manière unique comme combinaison linéaire de vecteurs de cette base. En d'autres termes, une base est une famille de vecteurs à la fois libre et génératrice d'un espace vectoriel.

10.5.1 Notion de Base

Indépendance linéaire : *Soit $\overrightarrow{e}_1, \overrightarrow{e}_2 \ldots, \overrightarrow{e}_n$ un nombre fini de vecteurs appartenant à l'espace vectoriel E. On dit que ces vecteurs sont linéairement indépendants (ou forme une famille libre) si l'unique solution du système d'équations*

$$a_1 \overrightarrow{e}_1 + a_2 \overrightarrow{e}_2 + \cdots + a_n \overrightarrow{e}_n = \overrightarrow{0}$$

est la nullité simultanée de tous les coefficients a_i ($\forall i \quad a_i = 0$). Dans le cas contraire, on dit que ces vecteurs sont linéairement dépendants.

C'est ni plus ni moins la définition la plus importante de l'algèbre linéaire !

Famille génératrice : *Soit $\overrightarrow{e}_1, \overrightarrow{e}_2 \ldots, \overrightarrow{e}_n$ un nombre fini de vecteurs appartenant à l'espace vectoriel E. On dit que ces vecteurs sont générateurs de E si, et seulement si, tout vecteur de E est une combinaison linéaire des vecteurs de cette famille.*

$$\forall \overrightarrow{u} \in E,\ \exists (a_1, a_2, \ldots, a_n) \in K^n,\ \overrightarrow{u} = \sum_{i=1}^{n} a_i \overrightarrow{e}_i$$

et bien sûr maintenant

Base d'un espace vectoriel : *Une famille de vecteurs de E est une base de E si, et seulement si, c'est une famille à la fois libre dans E et génératrice de E. De façon équivalente, une famille est une base de l'espace vectoriel E quand tout vecteur de l'espace se décompose* ***de façon unique*** *en une combinaison linéaire de vecteurs de cette base. Le nombre de vecteurs de la base est la dimension de l'espace vectoriel.*

Théorème : *Tout espace vectoriel admet une base.*

10.5.2 Base canonique

L'espace $\mathbb{R}^n$ est un espace vectoriel et chaque vecteur $\overrightarrow{u} = (x_1, x_2, \ldots, x_n)$ de $\mathbb{R}^n$ peut être écrit de manière unique comme

$$\overrightarrow{u} = x_1 \overrightarrow{e}_1 + x_2 \overrightarrow{e}_2 + \cdots + x_n \overrightarrow{e}_n \quad \text{où} \quad \overrightarrow{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \overrightarrow{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \overrightarrow{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

On appelle l'ensemble des vecteurs $\{\overrightarrow{e}_1, \overrightarrow{e}_2, \ldots, \overrightarrow{e}_n\}$ la **base canonique** de $\mathbb{R}^N$.

10.5.3 Base orthonormée

Définition : *Dans le cas d'un espace métrique, une base est dite orthonormée si, et seulement si, les vecteurs de cette base sont deux à deux orthogonaux et sont de norme égale à* 1.

Par exemple, la base canonique de $\mathbb{R}^n$ est orthonormée pour le produit scalaire usuel. À partir d'une base quelconque d'un espace euclidien, le procédé de Gram-Schmidt fournit une méthode constructive pour obtenir une base orthonormale de cet espace. Notamment, on peut affirmer :

Lemme : *Dans tout espace métrique de dimension non nulle, il existe des bases orthonormées.*

Soit $\mathcal{B} = (\overrightarrow{e}_1, \overrightarrow{e}_2 \ldots, \overrightarrow{e}_n)$ une base orthonormale d'un espace métrique E. L'unique décomposition d'un vecteur $\overrightarrow{u}$ de E dans cette base est donnée par

$$\overrightarrow{u} = \sum_{i=1}^{n} \Pi_{\overrightarrow{e}_i}(\overrightarrow{u})$$

ou encore

$$\overrightarrow{u} = \sum_{i=1}^{n} \langle \overrightarrow{e}_i \mid \overrightarrow{u} \rangle \overrightarrow{e}_i$$

Les composantes du vecteur $\overrightarrow{u}$ sont donc les projections $\langle \overrightarrow{e}_i \mid \overrightarrow{u} \rangle$. L'expression du produit scalaire de deux vecteurs de E est alors donnée par :

$$\langle \overrightarrow{u} \mid \overrightarrow{v} \rangle = \sum_{i=1}^{n} \langle \overrightarrow{e}_i \mid \overrightarrow{u} \rangle \langle \overrightarrow{e}_i \mid \overrightarrow{v} \rangle.$$

L'expression du carré de la norme d'un vecteur

$$|| \overrightarrow{u} ||^2 = \sum_{i=1}^{n} \langle \overrightarrow{e}_i \mid \overrightarrow{u} \rangle^2.$$

Ces trois propriétés sont en fait équivalentes entre elles, et équivalentes au fait que la famille $\mathcal{B}$ soit une base orthonormale de E

10.5.4 Le procédé de Gram–Schmidt

Le procédé de Gram-Schmidt est un algorithme pour construire, à partir d'une famille libre finie ou dénombrable de vecteurs, une base orthonormée du sous-espace qu'elle engendre.

Soit $\left(\overrightarrow{f}_1, \overrightarrow{f}_2, \ldots, \overrightarrow{f}_n\right)$ une famille libre de vecteurs. Le procédé de Gram–Schmidt se décompose en deux étapes. Dans un premier temps, on détermine itérativement une famille de vecteurs orthogonaux deux à deux $(\overrightarrow{e}_1, \overrightarrow{e}_2, \ldots, \overrightarrow{e}_n)$, puis on normalise (si on le souhaite) chacun des vecteurs obtenus.

La première étape de l'algorithme consiste à soustraire au vecteur $\overrightarrow{f}_{j+1}$ son projeté orthogonal sur le sous-espace engendré par $\left(\overrightarrow{f}_1, \overrightarrow{f}_2, \ldots, \overrightarrow{f}_j\right)$. On s'appuie sur la famille orthonormale déjà construite pour le calcul de ce projeté.

$$\begin{aligned}
\overrightarrow{e}_1 &= \overrightarrow{f}_1 \\
\overrightarrow{e}_2 &= \overrightarrow{f}_2 - \Pi_{\overrightarrow{e}_1}\left(\overrightarrow{f}_2\right) \\
\overrightarrow{e}_3 &= \overrightarrow{f}_3 - \Pi_{\overrightarrow{e}_1}\left(\overrightarrow{f}_3\right) - \Pi_{\overrightarrow{e}_2}\left(\overrightarrow{f}_3\right) \\
&\ldots
\end{aligned}$$

Dans un deuxième temps on peut obtenir une base orthonormée $(\overrightarrow{u}_i)$ en normalisant chacun des vecteurs $\overrightarrow{e}_i$:

$$\overrightarrow{u}_i = \frac{\overrightarrow{e}_i}{|| \overrightarrow{e}_i ||}$$

10.6 Calcul matriciel

10.6.1 Application linéaire

Les fonctions linéaires sont des exemples très simples de fonctions à plusieurs variables, mais elles sont également très importantes.

Application linéaire : *Une application linéaire est une fonction d'un espace vectoriel dans un autre satisfaisant*

- $f(\overrightarrow{u} + \overrightarrow{v}) = f(\overrightarrow{u}) + f(\overrightarrow{v})$;
- $f(\lambda \overrightarrow{u}) = \lambda f(\overrightarrow{u})$.

Si on considère souvent une application linéaire d'un espace dans lui même (comme des transformations géométriques dans l'espace matériel), l'ensemble d'arrivée de l'application linéaire est en toute généralité différent (de nature et de taille). Le projecteur défini au paragraphe 1.4 est par exemple une application linéaire d'un espace de dimension n vers un espace de dimension 1.

Définition : *Une application linéaire f est dite orthogonale si elle "conserve" le produit scalaire, i.e. si*

$$\langle f(\overrightarrow{u}) \mid f(\overrightarrow{v}) \rangle = \langle \overrightarrow{u} \mid \overrightarrow{v} \rangle$$

Une application orthogonale converse en particulier la norme.

10.6.2 Représentation matricielle d'une application linéaire

Soit g une application linéaire d'un espace vectoriel E de dimension n vers un espace vectoriel F de dimension m. Soit $\mathcal{B}_E = (\overrightarrow{e}_1, \overrightarrow{e}_2 \ldots, \overrightarrow{e}_n)$ une base de E et $\mathcal{B}_F = \left(\overrightarrow{f}_1, \overrightarrow{f}_2 \ldots, \overrightarrow{f}_n\right)$ une base de F. Soit enfin $\overrightarrow{u}$ un vecteur de E. Le vecteur $\overrightarrow{u}$ se décompose de manière unique dans la base $\mathcal{B}_E$

$$\overrightarrow{u} = x_1 \overrightarrow{e}_1 + x_2 \overrightarrow{e}_2 + \cdots + x_n \overrightarrow{e}_n = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

alors le résultat de l'application linéaire g sur le vecteur $\overrightarrow{u}$ est

$$g\left(\overrightarrow{u}\right) = g\left(x_1 \overrightarrow{e}_1 + x_2 \overrightarrow{e}_2 + \cdots + x_n \overrightarrow{e}_n\right) = x_1\, g(\overrightarrow{e}_1) + x_2\, g(\overrightarrow{e}_2) + \cdots + x_n\, g(\overrightarrow{e}_n).$$

Le résultat est donc parfaitement défini lorsque l'on connaît la décomposition de $\overrightarrow{u}$ dans la base de E et le résultat de l'application g sur chacun des vecteurs de la base $\mathcal{B}_E$ de l'ensemble de départ.

Les vecteurs $g\left(\overrightarrow{e}_i\right)$ qui appartiennent à l'espace d'arrivée F se décomposent de manière unique dans la base $\mathcal{B}_F$:

$$g(\overrightarrow{e}_i) = a_{1,i} \overrightarrow{f}_1 + a_{2,i} \overrightarrow{f}_2 + \cdots + a_{m,i} \overrightarrow{f}_m.$$

L'ensemble des coefficients $a_{i,j}$ forme un tableau à m lignes et n colonnes que l'on nomme $m \times n$–matrice

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix}.$$

La matrice A caractérise totalement l'application linéaire si on connaît les bases $\mathcal{B}_E$ et $\mathcal{B}_F$ utilisées.

Finalement le résultat de l'application linéaire vaut

$$g\left(\overrightarrow{u}\right) = \sum_{j=1}^{n} x_j \sum_{i=1}^{m} a_{i,j} \overrightarrow{f}_i = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{i,j} x_j \right) \overrightarrow{f}_i = A \cdot \overrightarrow{u}$$

Chaque application linéaire peut donc être décrite sous la forme compacte $g(\overrightarrow{u}) = A \cdot \overrightarrow{u}$. Les colonnes de la matrice sont les transformations par g des vecteurs de bases. L'écriture $g(\overrightarrow{u}) = A \cdot \overrightarrow{u}$ est incomplète, elle suppose connues du lecteur les bases d'arrivée et de départ ; sans la connaissance de ces bases, cette écriture ne veut rien dire. On dit que la matrice A est la **représentation de l'application** g. La représentation dépend de la base choisie, en conséquence, il peut exister des bases pour lesquelles la représentation de A est la plus simple.

10.6.3 Matrices remarquables

Matrice inversible

Définition : *Une $n \times n$–matrice A est inversible si il existe une $n \times n$–matrice notée A^{-1} telle que*

$$A^{-1}\, A = I_n$$

où I_n est la matrice identité. Une matrice inversible est aussi dite régulière

Propriétés des matrices inversibles Soit une $n \times n$–matrice A à coefficient dans K, les propositions suivantes sont équivalentes :

- A est inversible,
- le déterminant de A est non nul,
- 0 n'est pas valeur propre de A,
- les colonnes de A, considérées comme une famille de vecteurs, sont linéairement indépendantes,
- les colonnes de A, considérées comme une famille de vecteurs, forment une base de K^n,
- la transposée A^t de A est inversible.

Matrice orthogonale

Définition : *Une matrice représentant une application orthogonale est dite matrice orthogonale.*

Propriétés des matrices orthogonales

- Une $n \times n$–matrice A est orthogonale si, et seulement si, $A^{-1} = A^t$ où A^t est la transposée de la matrice A.
- Une matrice est orthogonale si et seulement si tous ses vecteurs colonnes sont orthogonaux entre eux et de norme 1. Ainsi une matrice orthogonale représente une base orthonormale.
- Également, une matrice est orthogonale si et seulement si sa transposée l'est, donc si et seulement si ses vecteurs lignes sont orthogonaux deux à deux et de norme 1.
- Le carré du déterminant d'une matrice orthogonale est égal à 1. Le déterminant d'une matrice orthogonale est donc égal à +1 ou -1. Si A est une matrice orthogonale et que son déterminant est +1 (respectivement -1), on dit que A est directe (respectivement indirecte).
- Le conditionnement d'une matrice orthogonale est égal à 1.
- La multiplication d'un vecteur par une matrice orthogonale préserve la norme de ce vecteur.

Les matrices orthogonales ne sont pas toutes diagonalisables.

Matrice unitaire et hermitienne

Définition : *Une $n \times n$–matrice A est unitaire si, et seulement si $A^{-1} = A^{\dagger t}$, où $A^\dagger$ est la matrice conjuguée de A obtenue en utilisant les complexe conjugué des éléments de A.*

Théorème : *Toutes les matrices unitaires sont diagonalisables*

Définition : *Une $n \times n$–matrice A est hermitienne si, et seulement si $A = A^{\dagger t}$.*

Théorème : *Toutes les matrices hermitiennes sont diagonalisables*

Théorème : *Toutes les matrices réelles symétriques sont diagonalisables*

Matrices semblables

Définition : *On dit que deux $n \times n$ matrices A et B sont semblables si il existe une $n \times n$ matrice C inversible telle que :*

$$B = C^{-1}\, A\, C$$

Théorème : *Deux matrices semblables ont les mêmes valeurs propres.*

Matrice triangulaire

Les matrices triangulaires sont des matrices carrées dont une partie triangulaire des valeurs, délimitée par la diagonale principale, est nulle

Matrices triangulaires supérieures : *Une $n \times n$–matrice A est dite triangulaire supérieure si les valeurs sous la diagonale principale sont nulles :*

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & \cdots & a_{1,n} \\ 0 & a_{2,2} & & & a_{2,n} \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & a_{n,n} \end{pmatrix}$$

Matrices triangulaires inférieures : *Une $n \times n$–matrice A est dite triangulaire inférieure si les valeurs sous la diagonale principale sont nulles :*

$$A = \begin{pmatrix} a_{1,1} & 0 & \cdots & \cdots & 0 \\ a_{2,1} & a_{2,2} & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ a_{n,1} & a_{n,2} & \cdots & \cdots & a_{n,n} \end{pmatrix}$$

Propriétés des matrices triangulaires

- Le produit de deux matrices triangulaires inférieures (respectivement supérieures) est une matrice triangulaire inférieure (respectivement supérieure).
- La transposée d'une matrice triangulaire supérieure est une triangulaire inférieure, et vice-versa.
- Une matrice triangulaire A est inversible si et seulement si tous ses termes diagonaux sont non nuls. Dans ce cas, son inverse est aussi une matrice triangulaire (supérieure si A est supérieure, inférieure sinon).
- Les valeurs propres d'une matrice triangulaire sont ses termes diagonaux.
- Le déterminant d'une matrice triangulaire est égal au produit de ses éléments diagonaux

10.6.4 Changement de base

Théorème : *Soit g une application linéaire d'un espace E de dimension n dans lui–même. Soient $\mathcal{B}_1 = (\overrightarrow{e}_1, \overrightarrow{e}_2 \ldots, \overrightarrow{e}_n)$ et $\mathcal{B}_2 \left(\overrightarrow{f}_1, \overrightarrow{f}_2 \ldots, \overrightarrow{f}_n\right)$ deux bases distinctes de E. Notons A la représentation de g dans la base $\mathcal{B}_1$ et A' la représentation de g dans la base $\mathcal{B}_2$. En notant $c_{i,j}$ les coefficients de la décomposition des vecteurs de la base $\mathcal{B}_2$ dans la base $\mathcal{B}_1$ i.e.*

$$\overrightarrow{f}_i = \sum_{i=1}^{n} c_{j,i} \overrightarrow{e}_j \qquad \forall i$$

et C la matrice constituée des coefficients $c_{i,j}$ arrangés comme

$$C = \begin{pmatrix} c_{1,1} & \ldots & c_{1,n} \\ \vdots & & \vdots \\ c_{n,1} & \ldots & c_{n,n} \end{pmatrix}.$$

La i–ème colonne de C est la décomposition du vecteur $\overrightarrow{f}_i$ dans la base $\mathcal{B}_1$. On a alors

$$A' = C^{-1}\, A\, C$$

Cas important : *Si les bases $\mathcal{B}_1$ et $\mathcal{B}_2$ sont toutes deux orthogonales alors la matrice de passage C est une matrice orthogonale et l'on a*

$$A' = C^{t}\, A\, C$$

Chapitre 11

L'algèbre linéaire au service du traitement statistique des données

11.1 La classification hiérarchique

Les principes généraux communs aux diverses techniques de classification hiérarchique sont extrêmement simples. Il est difficile de leur trouver une paternité, car ces principes relèvent plus du bon sens que d'une théorie formalisée.

Principe de la méthode

Dans le domaine des statistiques, la classification hiérarchique est une méthode d'analyse typologique qui cherche à établir une hiérarchie de classe à partir d'une population. Chaque classe rassemble les individus les plus semblables. Les stratégies pour la classification hiérarchique se répartissent en deux types :

- **Agglomératif** : C'est un "bottom up" : chaque observation est initialement dans sa propre classe. Les classes sont fusionnées par paire à mesure que l'on remonte dans la hiérarchie.
- **Divisive** : C'est un «top down» : toutes les observations sont initialement dans une seule classe que l'on fissure de façon récursive à mesure que l'on descend dans la hiérarchie.

L'algorithme ne fournit pas une partition en q agrégats d'un ensemble de n individus mais une hiérarchie de partitions, se présentant sous la forme d'arbres appelés dendrogrammes et contenant $n-1$ partitions. Chaque coupure d'un arbre fournit une partition, ayant d'autant moins de classe et des classes d'autant moins homogènes que l'on coupe plus haut.

Similarité des agrégats

Afin de déterminer quelles classes doivent être combinées (par agglomération), ou lorsqu'une classe doit être partagée (par division), une mesure de dissimilitude entre les séries d'observations est nécessaire. Dans la plupart des méthodes de classification hiérarchique, cela est réalisé par l'utilisation d'une métrique appropriée (une mesure de distance entre les paires d'observations), et un critère de liaison qui spécifie la dissemblance des ensembles en fonction des distances par paire des observations dans les ensembles.

Métrique. Le choix d'une métrique va influencer la forme des agrégats ; certains éléments peuvent être proches les uns des autres selon une distance et plus loin en fonction d'une autre. Par exemple, dans un espace à 2 dimensions, la distance entre le point $(1,0)$ et l'origine $(0,0)$ est toujours 1 selon les normes habituelles, mais la distance entre le point $(1,1)$ et l'origine $(0,0)$ peut être 2, $\sqrt{2}$ ou 1 sous la distance de Manhattan, la distance euclidienne ou la distance maximale respectivement.

Pour déterminer la distance entre deux séries de données numériques $\overrightarrow{a}$ et $\overrightarrow{b}$, on utilisera une $p-$norme

$$d(\overrightarrow{a}, \overrightarrow{b}) = || \overrightarrow{a} - \overrightarrow{b} ||_p$$

ou une mesure de colinéarité

$$d(\overrightarrow{a}, \overrightarrow{b}) = \frac{\overrightarrow{a} \cdot \overrightarrow{b}}{|| \overrightarrow{a} ||_2 || \overrightarrow{b} ||_2}$$

.

Pour le texte ou les autres données non numériques on peut utiliser

- la distance de Hamming entre deux chaînes de longueur égale qui est le nombre de positions où les symboles correspondants sont différents. Autrement dit, elle mesure le nombre minimum de substitutions nécessaires pour transformer une chaîne dans l'autre, ou le nombre d'erreurs qui a transformé une chaîne dans l'autre.
- La distance de Levenshtein mesure la similarité entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre.

Lorsque les données ne sont pas homogènes (textes et nombres) il faut combiner astucieusement les différentes distances.

Critère de liaison. Une fois constitué un groupe d'individus, il faut déterminer une mesure de distance entre un individu d'une classe et une autre classe, puis la distance entre deux classes définissant ainsi les critères de liaisons entre classes. Cela revient à déterminer une stratégie de regroupement des individus. La distance entre classes pourra en général se calculer directement à partir des distances des différents éléments impliqués dans le regroupement.

Soit deux classes $A = (a_1, a_2, \ldots, a_p)$ et $B = (b_1, b_2, \ldots, b_q)$ et une métrique d entre individus, les distances (ou liaison) entre A et B couramment utilisées sont

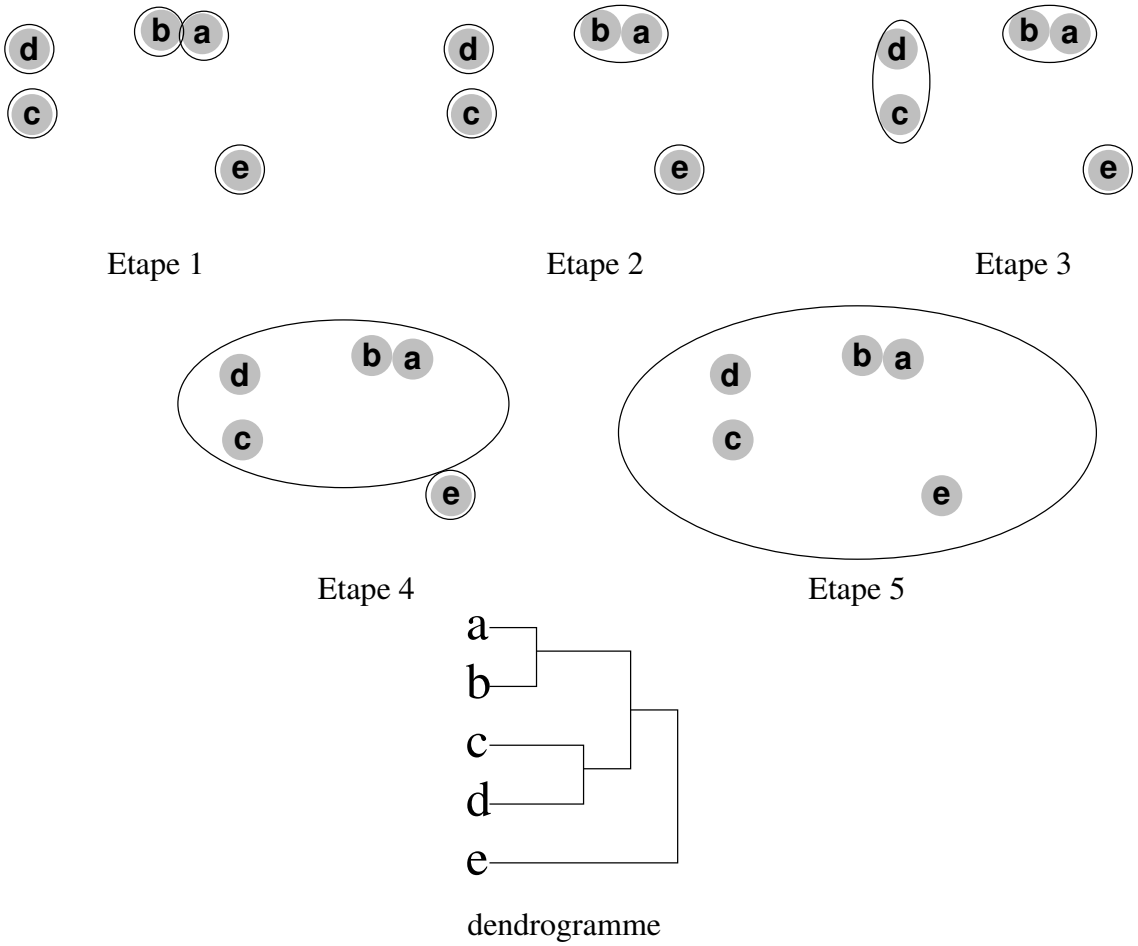
- Plus grande distance : $D(A, B) = \max\{d(a_i, b_j),\ i \in [1, p],\ j \in [1, q]\}$;
- Plus petite distance : $D(A, B) = \min\{d(a_i, b_j),\ i \in [1, p],\ j \in [1, q]\}$;
- Distance moyenne : $D(A, B) = \frac{1}{p\,q} \sum_{i=1}^{p} \sum_{j=1}^{q} d(a_i, b_j)$.

La méthode

Partant d'une population de N individu et ayant défini la métrique et le critère de lien, on peut lancer le clustering. On détaille ici le procédé par agrégation

- **Étape 0 :** Calculer la distance entre les $N(N-1)$ paires d'individus distinctes.
- **Étape 1 :** Ranger les N individus dans N agrégats.
- **Étape 2 :** Fusionner les deux agrégats les plus proches, il reste donc $N-1$ agrégats.
- **Étape 3 :** Déterminer les liaisons entre les $(N-1)(N-2)$ agrégats à l'aide du critère de liaison choisi et fusionner les deux agrégats les plus proches, il reste donc $N-2$ agrégats.
- **Étape 4 :** Déterminer les liaisons entre les $(N-2)(N-3)$ agrégats à l'aide du critère de liaison choisi et fusionner les deux agrégats les plus proches, il reste donc $N-3$ agrégats.
- ...
- **Étape N-1 :** Il ne reste que deux agrégats, la dernière étape consiste donc à mettre tous les individus sont dans 1 agrégat.

Le procédé est illustré en prenant comme objet à classer cinq points du plan.



Pour de plus amples détails sur les méthodes d'exploration statistique des données, voir par exemple *"Statistique exploratoire multidimensionnelle", Lebart L., Morineau A., Piron M.* Sciences Sup (Dunod)

11.2 L'analyse en composantes principales

L'analyse en Composantes Principales (ACP) est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites "corrélées" en statistique) en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées "composantes principales", ou axes. Elle permet au praticien de réduire l'information en un nombre de composantes plus limité que le nombre initial de variables.

Il s'agit d'une approche à la fois géométrique (représentation des variables dans un nouvel espace géométrique selon des directions d'inertie maximale) et statistique (recherche d'axes indépendants expliquant au mieux la variabilité - la variance - des données). Lorsqu'on veut alors compresser un ensemble de N variables aléatoires, les n premiers axes de l'ACP sont un meilleur choix, du point de vue de l'inertie ou la variance expliquée

Les champs d'application sont aujourd'hui multiples, allant de la biologie à la recherche économique et sociale, et plus récemment le traitement d'images. L'ACP est majoritairement utilisée pour :

— décrire et visualiser des données ;
— les décorréler ; la nouvelle base est constituée d'axes qui ne sont pas corrélés entre eux ;
— les débruiter, en considérant que les axes que l'on décide d'oublier sont des axes bruités.

La puissance de l'ACP est qu'elle sait aussi prendre en compte des données de nature hétérogène : par exemple un tableau des différents pays du monde avec le PNB par habitant, le taux d'alphabétisation, le taux d'équipement en téléphones portables, le prix moyen du hamburger, etc. Elle permet d'avoir une intuition rapide des effets conjoints entre ces variables.

Échantillon

On applique usuellement une ACP sur un ensemble de N variables aléatoires $\left(\overrightarrow{X}_1, \ldots, \overrightarrow{X}_N\right)$ connues à partir d'un échantillon de K réalisations conjointes de ces variables. Cet échantillon de ces N variables aléatoires peut être structuré dans une matrice M à K lignes et N colonnes.

$$M = \begin{pmatrix} X_{1,1} & \ldots & X_{1,N} \\ \vdots & \ldots & \vdots \\ X_{K,1} & \ldots & X_{K,N} \end{pmatrix} \tag{11.1}$$

Chaque variable aléatoire $\overrightarrow{X}_n = (X_{1,n}, \ldots, X_{K,n})$ a une moyenne $\bar{X}_n$ et un écart type σ_{X_n}.

Transformation de l'échantillon

Le vecteur $\overrightarrow{g} = \left(\bar{X}_{1,n}, \ldots, \bar{X}_{K,n}\right)$ est le centre de gravité du nuage de points. La matrice M est généralement centrée sur le centre de gravité, chaque vecteur colonne de la matrice à ainsi une moyenne nulle.

$$\bar{M} = \begin{pmatrix} X_{1,1} - \bar{X}_1 & \ldots & X_{1,N} - \bar{X}_N \\ \vdots & \ldots & \vdots \\ X_{K,1} - \bar{X}_1 & \ldots & X_{K,N} - \bar{X}_N \end{pmatrix} = M - \overrightarrow{g}\mathbb{I}. \tag{11.2}$$

La matrice M peut être aussi réduite, chaque vecteur colonne de la matrice à ainsi une variance de 1.

$$\tilde{M} = \begin{pmatrix} \frac{X_{1,1} - \bar{X}_1}{\sigma(X_1)} & \ldots & \frac{X_{1,N} - \bar{X}_N}{\sigma(X_N)} \\ \vdots & \ldots & \vdots \\ \frac{X_{K,1} - \bar{X}_1}{\sigma(X_1)} & \ldots & \frac{X_{K,N} - \bar{X}_N}{\sigma(X_N)} \end{pmatrix}. \tag{11.3}$$

Le choix de réduire ou non le nuage de points (i.e. les K réalisations de la variable aléatoire $\left(\overrightarrow{X}_1, \ldots, \overrightarrow{X}_N\right)$ est un choix de modèle :

- si on ne réduit pas le nuage : une variable à forte variance va « tirer » tout l'effet de l'ACP à elle ;
- si on réduit le nuage : une variable qui n'est qu'un bruit va se retrouver avec une variance apparente égale à une variable informative.

Calcul de covariances et de corrélations

Une fois la matrice M transformée en $\bar{M}$ ou $\tilde{M}$, il suffit de la multiplier par sa transposée pour obtenir :

- la matrice de variance-covariance des $\left(\overrightarrow{X}_1, \ldots, \overrightarrow{X}_N\right)$ si M n'est pas réduite ;
- la matrice de corrélation des $\left(\overrightarrow{X}_1, \ldots, \overrightarrow{X}_N\right)$ si M est réduite.

Ces deux matrices sont carrées (de taille N), symétriques, et réelles. Elles sont donc diagonalisables dans une base orthonormée et ont des valeurs propres réelles.

Principe de l'ACP

Le principe de l'ACP est de trouver un axe paramétré par un vecteur unitaire $\overrightarrow{u}$, issu d'une combinaison linéaire des $\overrightarrow{X}_n$, tel que la variance (l'étalement) du nuage autour de cet axe soit maximale.

Pour bien comprendre, imaginons que la variance sur $\overrightarrow{u}$ soit égale à la variance du nuage ; on aurait alors trouvé une combinaison des $\overrightarrow{X}_n$ qui contient toute la diversité du nuage original (en tout cas toute la part de sa diversité captée par la variance).

Projections

Finalement, on cherche le vecteur unitaire $\overrightarrow{u}$ tel que la projection du nuage sur $\overrightarrow{u}$ ait une variance maximale. La projection de l'échantillon des $\overrightarrow{X}_n$ centré (et éventuellement réduit) sur $\overrightarrow{u}$ s'écrit

$$\overrightarrow{\Pi}_{\overrightarrow{u}}\left(\tilde{M}\right) = \tilde{M} \cdot \overrightarrow{u}.$$

Cette projection appliquée à une matrice diffère dans sa définition de celle appliquée aux vecteurs. Le résultat est ici un vecteur dont la i–ème composante correspond au produit scalaire du i–ème vecteur colonne de la matrice $\tilde{M}$ et du vecteur unitaire $\overrightarrow{u}$ recherché.

La variance de la projection est le carré de la norme du vecteur de projection

$$|| \overrightarrow{\Pi}_{\overrightarrow{u}}\left(\tilde{M}\right) ||^2 = \langle \overrightarrow{\Pi}_{\overrightarrow{u}}\left(\tilde{M}\right) \mid \overrightarrow{\Pi}_{\overrightarrow{u}}\left(\tilde{M}\right)\rangle = \overrightarrow{u}^T \tilde{M}^T \cdot M \overrightarrow{u} = \overrightarrow{u}^T C \overrightarrow{u}$$

où C est la matrice de corrélation (ou variance–covariance si on utilise $\bar{M}$)

Comme il a été vu plus haut que C est diagonalisable dans une base orthonormée, notons Q la matrice orthogonale de changement de base associée et $\Lambda = Diag\,(\lambda_1, \ldots, \lambda_N)$ la matrice diagonale formée de ses valeurs propres λ_i rangées par ordre décroissant ($\lambda_1 > \lambda_2 > \cdots > \lambda_N$). On a alors

$$C = Q^T \cdot \Delta \cdot Q$$

et donc

$$|| \overrightarrow{\Pi}_{\overrightarrow{u}}\left(\tilde{M}\right) ||^2 = \overrightarrow{u}^T Q^T \cdot \Delta \cdot Q \overrightarrow{u} = (Q\overrightarrow{u})^T \cdot \Delta \cdot (Q\overrightarrow{u})\,.$$

Le vecteur $\overrightarrow{v} = Q\overrightarrow{u}$ est de même norme que $\overrightarrow{u}$ puisque Q est orthogonale. L'objectif est alors de déterminer le vecteur unitaire $\overrightarrow{v}$ qui maximise $\overrightarrow{v}^T \cdot \Delta \cdot \overrightarrow{v}$. En posant $\overrightarrow{v} = (v_1, v_2, \ldots, v_n)$ cela revient à résoudre

$$\min\left(\lambda_1\, v_1^2 + \lambda_2\, v_2^2 + \cdots + \lambda_n\, v_n^2\right) \quad \text{avec} \quad v_1^2 + v_2^2 + \cdots + v_n^2 = 1,$$

dont la solution est évidente $v_1 = 1$ et $v_{i>1} = 0$. En conséquence $\overrightarrow{u}$ est le vecteur propre de C associé à la valeur propre λ_1. La valeur propre λ_1 est la variance empirique sur le premier axe de l'ACP, elle exprime le pourcentage d'inertie d'écrit par cet axe.

On continue la recherche du deuxième axe de projection $\overrightarrow{w}$ sur le même principe en imposant qu'il soit orthogonal à $\overrightarrow{u}$ on trouve le second vecteur propre et ainsi de suite. Finalement, la question de l'ACP se ramène à un problème de diagonalisation de la matrice de corrélation. On obtient en outre que la variance expliquée par le k-ème vecteur propre vaut λ_k.

Application : La régression linéaire

La régression linéaire est un cas particulier de l'analyse en composantes principales à deux dimensions. On suppose une collection de mesures $\overrightarrow{x} = (x_1, x_2, \ldots, x_n)$ et $\overrightarrow{y} = (y_1, y_2, \ldots, y_n)$. Montrer que l'ACP permet de déterminer la droite de régression $y = ax + b$.

Troisième partie

Analyse numérique élémentaires

Chapitre 12

Introduction aux problèmes numériques

Sommaire

L'analyse numérique est une discipline des mathématiques. Elle s'intéresse tant aux fondements théoriques qu'à la mise en pratique des méthodes permettant de résoudre, par des calculs purement numériques, des problèmes d'analyse mathématique.

L'analyse numérique traite de nombreux problèmes de sciences physiques, biologiques, technologiques ou des problèmes issus de modèles économiques et sociaux. Elle intervient dans le développement de codes de calcul ainsi que dans les problèmes de simulations ou d'expérimentations mathématiques. Elle entretient des liens étroits avec l'informatique. Si sa partie théorique relève plus des mathématiques, sa mise en pratique aboutit généralement à l'implémentation d'algorithme sur ordinateur, quelque que soit le langage de programmation utilisé.

En analyse numérique, un certain nombre de calculs sont faits de manière répétitive pour :
— résoudre un système linéaire,
— résoudre un système d'équations différentielles,
— inverser une matrice,
— calculer le déterminant, valeurs propres, vecteurs propres
— interpoler/extrapoler une fonction,
— décomposer une fonction sur une base de fonctions (ex : TF)
— ajuster ("fitter") des points de mesure avec des fonctions connues
— ...

Des algorithmes performants existent souvent, mais ils ont chacun leurs spécificités. En fonction de vos besoins, il faudra choisir l'algorithme le mieux adapté. La plupart de ces outils « de base » sont souvent déjà programmés dans des librairies de calculs mathématiques (la plupart des codes sont disponibles sur *www.netlib.org*). L'implémentation des codes numérique (le codage) n'est donc pas une difficulté majeure. Par contre, la connaissance des principes mathématiques et des performances de chaque algorithme est un investissement indispensable à tous "numéricien". L'utilisation de boîte noire est donc

à proscrire dans le cadre d'une analyse numérique sérieuse. En fonction du problème, en choisissant l'algorithme le plus adapté, on peut gagner du temps de calcul, ou mieux : éviter de fonder un résultat scientifique sur une erreur numérique.

12.1 Erreurs de calcul

Eh oui, l'ordinateur aussi fait des erreurs de calcul !

12.1.1 Sources d'erreur

Tout calcul numérique est par essence entaché d'erreur, seul le calcul analytique permet une description exacte du résultat. L'usage à bon escient du calcul numérique exige la parfaite connaissance des sources d'erreur et des techniques de contrôle de cette erreur. Donner un résultat numérique sans indiction de la précision est aussi inutilisable qu'une mesure sans incertitude. La précision du résultat dépend du contrôle de l'erreur numérique.

Pour évaluer la précision d'un résultat, il faut connaître les erreurs qui ont été commises. Donnons trois exemples.

1. **Les erreurs de troncature** sont liés à la précision de l'algorithme utilisé : la solution numérique obtenue diffère toujours de la solution mathématique exacte. Ces erreurs peuvent et doivent être contrôlées par l'algorithme, lui–même. Par exemple, si une fonction est approchée par son développement de Taylor, l'erreur de troncature sera obtenue par l'évaluation du reste du développement. Son contrôle sera obtenu par la majoration de ce reste.
2. **Les erreurs de méthode** se produisent lorsqu'une expression est mal équilibrée et mélange des valeurs dont la différence est importante. C'est un problème de calibrage numérique qui est sensible aux erreurs d'arrondi. Dans la plupart des cas, l'algorithme doit être modifié. Par exemple, le calcul de l'inverse d'une matrice de déterminant proche de 0 , mais non exactement 0, est très sensibles aux erreurs. Dans ce cas il faut préférer une inversion itérative, plutôt qu'une inversion exacte.
3. **Les erreurs de discrétisation** sont imposées par le calculateur. La représentation d'un nombre dans la mémoire de l'ordinateur (au sens large) étant finie, tout nombre réel n'est connu qu'avec une précision donnée de p chiffres significatifs. Dans les nombres réels cela entraîne des erreurs de d'arrondi qui vont se propager et polluer la suite des calculs.

Les différentes sources d'erreur interfèrent avec les erreurs d'arrondi lors d'un calcul. Par exemple, la résolution d'un problème très sensible aux variations sur les données donne lieu a des calculs avec de grandes erreurs d'arrondi.

12.1.2 Mesures de l'erreur

Tout calcul devrait s'accompagner d'une estimation des erreurs d'approximation commises. Pour mesurer celles-ci, nous définissons les erreurs absolues et relatives, ainsi que la notion de chiffres significatifs.

> **Erreurs absolue et relative :** *Soit x un réel et $\tilde{x}$ une approximation de x. L'erreur sur $\tilde{x}$ est $|\tilde{x} - x|$. Si $x \neq 0$, l'erreur relative est $\left|\frac{\tilde{x}-x}{x}\right|$.*
>
> **Nombre de chiffres significatifs :** *Le nombre de chiffres significatifs de x est le nombre de chiffres à partir du premier chiffre non nul.*

Exemple : Le nombre de chiffres significatifs de $x = 0,0034560$ est 5, L'erreur relative sur $\tilde{x} = 0,00346$ vaut environ $1,6\ 10^{-3}$ et le nombre de chiffres significatifs exact est ici 2.

Dans le cas de vecteurs, on définit les erreurs sur les normes ou les erreurs composante par composante.

12.1.3 Arithmétique flottante

Avant d'analyser l'impact des erreurs d'arrondi et des différentes sources d'erreur sur des calculs, nous allons détailler les caractéristiques arithmétiques d'un ordinateur. L'arithmétique flottante est définie en détails dans plusieurs ouvrages [1]. Un logiciel de calcul scientifique utilise plusieurs types de variables, qui correspondent a un codage spécifique. Les types arithmétiques les plus usuels sont le type *entier*, le type *réel* et le type *complexe*. Un complexe est formé de deux réels, la partie réelle et la partie imaginaire. Un réel est codé par un nombre flottant. Pour simplifié ne décrit ici que le système flottant en base 10 mais les résultats reste vrai pour n'importe qu'elle base : 2, 16, ...

Système flottant en base 10 : *La représentation d'un nombre réel à la forme suivante en virgule flottante*

$$\pm m_1, m_2 m_3 \dots m_p E^{\pm e_1 e_2}$$

cette représentation comporte p chiffres (entiers) significatifs pour la mantisse et 2 chiffres pour la puissance de 10. *L'ensemble des nombres flottants normalisés est noté* $\mathbb{F}$

Système flottant en base b **:** *Un système flottant est défini par une base b, un exposant minimal* e_{min}*, un exposant maximal* e_{max}*, un nombre de chiffres p. L'ensemble des nombres flottants normalisés est noté* $\mathbb{F}$*. Un nombre flottant non nul normalisé est*

$$x = (-1)^s \; m \; b^e,$$

où $s \in 0, 1$ *est le bit de signe, l'exposant e est un entier tel que* $e_{min} \leq e \leq e_{max}$*, la mantisse m vérifie* $1 \leq m \leq b$ *est s'écrit*

$$\begin{cases} m = a_0 + a_1\, b^{-1} + a_2\, b^{-2} + \cdots + a_p\, b^{-p} \\ avec\, 0 \leq a_i \leq b - 1 \; et \; a_0 \neq 0 \end{cases}$$

Par convention, le nombre 0 *à un exposant et une mantisse nuls.*

L'ensemble des nombres flottant est fini et ne décrit que quelques réels.

Précision : *Le plus petit nombre qui, ajouté à* $1,0$*, produit un nombre différent de* $1,0$ *est appelé la précision machine* ϵ *et dépend du nombre de bit de la mantisse.*

Tout réel x est encadré par deux flottants consécutifs x^+ et x^- tels que

$$x^- = \max\{y \in \mathbb{F},\; y < x\} \quad et \quad x^+ = \min\{y \in \mathbb{F},\; y > x\}.$$

Il n'existe pas de flottant entre x^- et x^+ et $0 \leq |x^+ - x^-| \leq |x|\epsilon$. Si $x \in \mathbb{F}$ alors $x^+ = x^- = x$.

Arrondis : *On note* $[x]$ *la représentation du réel x dans le système flottant choisi. On a soit* $[x] = x^+$ *soit* $[x] = x^-$*.*

Les propriétés des quatre opérations sur les nombres réelles $\{+, -, \times, /\}$ ne sont pas toutes vérifiées avec les nombres flottants

- Les opérations flottantes sont commutatives,
- elles ne sont pas associatives,
- elles ne sont pas distributives,
- 0 est l'élément neutre de l'addition flottante, 1 est l'élément neutre de la multiplication flottante,
- l'opposé de $x \in \mathbb{F}$ existe et vaut $-x$
- par contre $x \in \mathbb{F}$ n'a pas toujours d'inverse dans $\mathbb{F}$

Exemple : Soit $x = 1 + 10^{-3}$, $y = 1 + 10^{-8}$ et $z = 10^{-8}$, les nombres $u = \frac{(y+z))-y}{z}$ et $v = \frac{z+(y-y)}{z}$ valent bien évidement 1. Dans une arithmétique à 4 chiffres significatifs on aura [u]=+0,000E+00 et [v]=+1,000E+00. En analyse numérique l'ordre des opérations est très important, particulièrement lorsque l'on additionne des nombres d'ordres de grandeur très différents.

1. citons par exemple *Qualité des Calculs sur Ordinateurs. Vers des arithmétiques plus fiables ?* Masson, 1997.

Exceptions Le résultat exact d'une opération peut être extérieur au système flottant (un nombre trop grand, une division par zéro). Pour fermer le système arithmétique flottant, l'ensemble $\mathbb{F}$ est doté de nombres spéciaux, notés $+\infty$, $-\infty$, NaN (Plus l'infini, Moins l'infini, Not a Number). Le nombre 0 a un signe et les opération sont définies avec ces nombres spéciaux. Une opération ayant pour résultat un des nombre spéciaux déclenche une exception.

12.1.4 Norme IEEE-754

La norme IEEE-754, publiée en 1985, spécifie un système flottant et l'arithmétique flottante associée. La plupart des constructeurs respectent aujourd'hui cette norme, ce qui facilite grandement le transfert de logiciels entre machines. Cette norme permet aussi d'établir des preuves sur le comportement des erreurs d'arrondi dans un algorithme.

Type	Base b	Mantisse p	Exposant e_{min}	Exposant e_{max}
Simple	2	24	-126	$+127$
Double	2	53	-1022	$+1023$

TABLE 12.1 – Formats simple et double précision de la norme IEEE-754

Type	précision ϵ	Overflow x_{max}	Underflow u
Simple	10^{-7}	10^{+38}	10^{-38}
Double	10^{-16}	10^{+308}	10^{-308}

TABLE 12.2 – Valeurs caractéristiques approchées des formats simple et double précision de la norme IEEE-754

La norme IEEE-754 spécifie deux formats, appelés simple et double précision, qui sont résumés dans les tables (12.1) et (12.2). La norme définit quatre arrondis :

— L'arrondi vers moins l'infini de x vaut x^-,
— l'arrondi vers plus l'infini de x vaut x^+,
— l'arrondi vers zéro ou par troncature de x vaut x^+ si $x < 0$ et x^- si $x > 0$,
— l'arrondi au plus près de x vaut x^- si $x - x^- < x^+ - x$, il vaut x^+ sinon. Dans le cas d'égalité une règle de parité fixe le choix.

Exemple : Les deux nombres réels $x = \sqrt{2} \simeq 1.4142135623730951$ et $y = -\sqrt{2}$ sont encadrés par $x^- = 1.4142135$ et $x^+ = 1.4142137$ pour x ; $y^- = -1.4142137$ et $y^+ = -1.4142135$ pour y. Les représentations sont $[x] = x^-$ et $[y] = y^+$ de telle sorte que $[x] + [y] = 0$.

12.1.5 Phénomènes d'absorption et de cancellation

Un code scientifique performant vérifie la pertinence numérique de chaque opération ! Notons que la précision machine ϵ n'est pas le plus petit réel pouvant être représenté sur la machine qui dépend du nombre de bit de l'exposant.

L'absorption se produit lors de l'addition de deux quantités avec des ordres de grandeur très différents.

> **Absorption :** *L'addition de deux nombres réels x et y sera numériquement inexistante si*
>
> $$|y| \leq |x|\epsilon/2.$$
>
> *L'addition aura pour résultat :* $[x + y] = [x]$.

Exemple : La série harmonique

$$S_N = \sum_{n=1}^{N} \frac{1}{n}$$

est bien connue pour divergée : $\lim_{N\to\infty} S_N = \infty$. Or un calcul numérique utilisant un codage des réels sur 4 octets (simple précision) fait apparaître une convergence artefactuelle (voir le tableau suivant). Lorsque de rapport $\frac{1}{NS_N}$ est inférieure à la précision machine ($1,2 \times 10^{-7}$ sur la machine utilisée), l'addition d'un terme supplémentaire n'a plus d'effet.

N	$1/N$	S_N
400000	0.0000025	13.481427
800000	0.00000125	14.166623
1200000	8.3333333E-7	14.548093
1600000	6.25E-7	14.929563
2000000	5.E-7	15.311032
2400000	4.1666667E-7	15.403683
2800000	3.5714285E-7	15.403683

Le phénomène de **cancellation** se produit lors de la soustraction de deux nombres voisins. Il faut alors renormaliser le résultat car les chiffres de gauche s'éliminent (d'où le nom cancellation). Les chiffres de droite deviennent les premiers chiffres significatifs. Si les opérandes sont exacts, la cancellation est bénigne car la soustraction est en général exacte. Mais si les opérandes sont entachés d'erreur, l'ordre de grandeur du résultat exact est celui des incertitudes, donc le résultat flottant risque de n'avoir aucun sens ; la cancellation est dite catastrophique. Il faut noter que la cancellation est un révélateur des erreurs précédentes.

Cancellation : *Soit x et y deux nombres réels. il y a élimination (cancellation) de x et y si*

$$0 < |x - y| \leq (|x| + |y|)\,\epsilon/2.$$

De manière générale, il y a cancellation catastrophique entre plusieurs nombres si la somme de ces nombres est beaucoup plus petite que la somme de leurs valeurs absolues, autrement dit si

$$0 < \left|\sum x_i\right| \leq \left(\sum |x_i|\right)\epsilon/2.$$

Exemple : Prenons l'exemple du calcul numérique de la dérivée d'une fonction analytique connue f par la méthode des différences finies. L'expression mathématique exacte sous forme symétrique $f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x-h)}{2h}$ est remplacée par une approximation dépendant d'un paramètre h $f'_h(x) = \frac{f(x+h)-f(x-h)}{2h}$. La définition mathématique indique que l'approximation est d'autant meilleure que la valeur de h est petite. Illustrons les résultats obtenus en utilisant un codage des réels sur 4 et 8 octets en prenant pour f la fonction sinus et en évaluant la dérivée en $x = 1$. L'avantage étant que la solution analytique nous est connue : $f'(1) = \cos(1)$ on peut donc calculer simplement l'erreur relative en fonction du paramètre de différentiation h :

$$\epsilon_h = \left|\frac{f'_h(1) - f'(1)}{f'(1)}\right|.$$

La représentation graphique de l'erreur relative (figure 12.2) met en évidence une valeur optimale de h qui dépend du codage utilisé. La perte de précision pour les faibles valeurs de h provient de l'évaluation numérique de la différence $f(x+h) - f(x-h)$: un phénomène dit de "cancelation" survient lorsque les deux quantités $f(x \pm h)$ sont très proches.

12.1.6 Propagation de l'erreur

Chaque opération arithmétique introduit une erreur d'ordre ϵ. Si les erreurs sont aléatoires et uniforme (positives et négatives), le résultat de N opérations produira une erreur totale d'environ $\sqrt{N}\epsilon$. Si, par contre, les erreurs s'accumulent dans une direction l'erreur totale sera $N\epsilon$

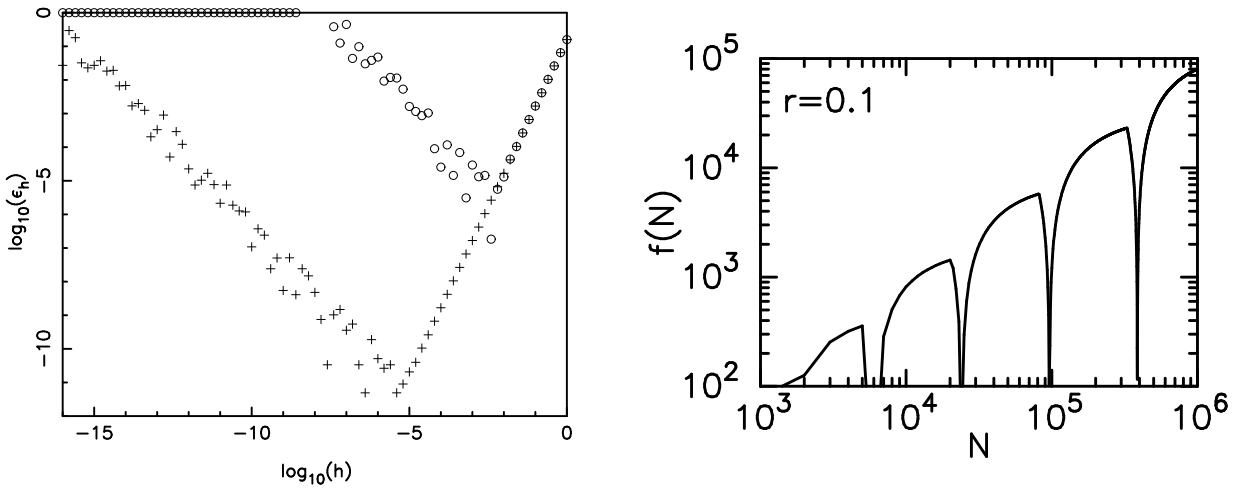


FIGURE 12.1 – À droite : Évolution de $f(N)$ pour le calcul des termes d'une suite arithmétique avec $r = 0.1$. À gauche : Évolution de l'erreur relative ϵ_h en fonction du paramètre de différentiation h. Les cercles correspondent à un codage des réels sur 4 octets et les croix sur 8.

Exemple : Supposons le calcul des termes d'une suite $u_{n+1} = u_n + r$ avec $u_0 = 0$ et r réel ; le calcul mathématique donne $u_n = nr$. On peut définir la fonction de propagation de l'erreur $f(N) = \frac{u_N - Nr}{Nr\,\epsilon}$. Comme le montre la partie gauche de la figure 1.1, l'évolution de $f(N)$ est loin d'être triviale.

Exercice : Considérons le calcul des termes d'une suite arithmetico–géométrique définie par la relation $x_{n+1} = ax_n + b$. Montrer que l'erreur évolue exponentiellement. Calculer l'amplification de l'erreur à l'étape $n = 3000$ si $a = 1.01$, $a = 0.99$.

12.2 Suites numériques et calcul itératif

De nombreux algorithmes numériques sont itératifs et se modélisent simplement à l'aide de la notion de suite mathématique.[2] Prenons par exemple le procédé d'extraction d'une racine carrée par la méthode de Héron d'Alexandrie : "*Pour extraire la racine carrée de A, choisir une expression arbitraire a et prendre la moyenne entre a et* $\frac{A}{a}$ *et recommencer aussi loin que l'on veut le processus précédent*" qui est un des plus vieux algorithmes numériques connus. En notation moderne, cela définit la suite de nombres (u_n) telle que $u_0 = a$ et $\forall n \in \mathbb{N},\ u_{n+1} = \frac{1}{2}\left(u_n + \frac{A}{u_n}\right)$ il est trivial de montrer que, si la suite converge, elle converge vers $\sqrt{A}$. Dans ce calcul itératif, les erreurs numériques commissent à chaque évaluation de u_n se propagent dans les évaluations suivantes du fait du caractère itératif du procédé, ce qui à pour effet éventuellement d'amplifier l'erreur globale et de diminuer la précision du calcul. La propriété de *stabilité* garantit que les erreurs ne s'amplifient pas au cours du déroulement de l'algorithme.

Les méthodes numériques utilisées pour résoudre un problème approché conduisent à des résultats toujours entachés d'erreur. Ces erreurs doivent être suffisamment petites pour que solution numérique converge vers la solution réelle. Dans ce cas l'algorithme est dit *convergent*. Si un raisonnement mathématique permet de montrer qu'une méthode diverge, elle ne pourra en aucun être utilisée sur un calculateur. En revanche si la méthode converge en théorie, il se peut qu'en pratique elle diverge.

Les suites numériques forment le prototype pour l'étude de la convergence. En effet, certaines suites ont leurs éléments qui "tendent" vers un nombre réel bien défini lorsque l'indice s'accroît. Ce comportement de certaines suites est nommé convergence.

> **Convergence :** *Une suite* (u_n) *converge vers* u^* *si quel que soit* $\varepsilon > 0$ *il existe un entier* n_0 *tel que* $\mid u_n - u^* \mid < \varepsilon$ *pour tout* $n \geq n_0$

Pour prouver la convergence d'une suite sans connaissance de la limite, on peut utiliser le critère de convergence de Cauchy

2. En mathématiques, une suite est une famille d'éléments indexée par les entiers naturels. Une suite finie est une famille indexée par les entiers strictement positifs inférieurs ou égaux à un certain entier, ce dernier étant appelé « longueur » de la suite.

Critère de convergence de Cauchy : *Une suite est converge si, et seulement si, quel que soit $\varepsilon > 0$ il existe un entier n_0 tel que $\mid u_n - u_m \mid < \varepsilon$ pour tous $m, n \geq n_0$*

On aura l'occasion de rencontrer des suites numériques dans de nombreux algorithmes. Profitions de l'occasion pour rappeler les grands classiques des suites numériques

- **Suite arithmétique :** la suite définie par $U_0 = a$ et $u_{n+1} = u_n + r$ a pour terme général $u_n = a + nr$ (non convergente)
- **Suite géométrique :** la suite définie par $U_0 = a$ et $u_{n+1} = qu_n$ a pour terme général $u_n = a\, q^n$ (convergente si $|q| < 1$)
- **arithmético–géométriques :** la suite définie par $U_0 = a$ et $u_{n+1} = q\ u_n + r$ (avec $q \neq 1$) a pour terme général $u_n = \frac{r}{1-q} + q^n \left(a - \frac{r}{1-q}\right))$

Vitesse de convergence d'une suite En analyse numérique, on peut classer les suites convergentes en fonction de leur vitesse de convergence vers leur point limite. C'est une manière d'apprécier l'efficacité des algorithmes qui les génèrent.

Les suites considérées ici sont convergentes sans être stationnaires ; plus précisément, tous leurs éléments sont supposés différents du point limite. Si une suite est stationnaire, tous ses éléments sont égaux à partir d'un certain rang et il est alors normal de s'intéresser au nombre d'éléments différents du point limite. C'est ce que l'on fait lorsqu'on étudie la complexité des algorithmes trouvant ce qu'ils cherchent en un nombre fini d'étapes.

La notion vitesse de convergence d'une suite $(u_{n>1})$ d'un espace normé vers sa limite u^* est fondée sur la comparaison de la norme de l'erreur $\mid u_n - u^* \mid$ de deux éléments successifs. L'erreur est toujours supposée non nulle : $u_n \neq u^*$, pour tout indice n. Cette hypothèse est raisonnable lorsque la suite est générée par un algorithme bien conçu, car dès que $u_n = u^*$, la suite devient stationnaire après u_n (tous les itérés suivants sont égaux à u^*) et il n'y a plus de sens à parler de vitesse de convergence. On s'intéresse donc au quotient

$$\frac{\mid u_{n+1} - u^* \mid}{\mid u_n - u^* \mid^{\alpha}}$$

où α est un nombre réel strictement positif. L'intérêt de ce quotient est que l'on peut souvent l'estimer en faisant un développement de Taylor autour u^* des fonctions du problème que l'on cherche à résoudre et dont u^* est solution.

Brièvement, on dit que l'ordre de convergence est $\alpha > 1$ si le quotient ci-dessus est borné. On dit aussi que la convergence est :

- **linéaire** (ordre 1) si il existe un réel $\tau \in [0, 1[$ et un indice $n_0 > 1$, tels que pour tout $n \geq n_0$, on a $\mid u_{n+1} - u^* \mid \leq \tau \mid u_n - u^* \mid$
- **super–linéaire** si $\frac{|u_{n+1} - u^*|}{|u_n - u^*|^{\alpha}} \to 0$
- **quadratique** (ordre 2) si il existe un réel $C > 0$ et un indice $n_0 > 1$, tels que pour tout $n \geq n_0$, on a $\mid u_{n+1} - u^* \mid \leq C \mid u_n - u^* \mid^2$
- **cubique** (ordre 3) si il existe un réel $C > 0$ et un indice $n_0 > 1$, tels que pour tout $n \geq n_0$, on a $\mid u_{n+1} - u^* \mid \leq C \mid u_n - u^* \mid^3$

Nombre de chiffres significatifs Numériquement, plus la convergence est rapide, plus le nombre de chiffres significatifs corrects de x_k (ceux identiques à ceux de u^*) augmente vite. Donnons une définition plus précise de cette notion. On suppose que, $u^* \neq 0$ car on ne peut définir ce que sont les chiffres significatifs de zéro. Si $\frac{|u_n - u^*|}{|u^*|}$ vaut 10^{-4}, on dira que u_n a 4 chiffres significatifs corrects. Ceci conduit à la définition suivante.

Nombre de chiffres significatifs corrects : *Le nombre de chiffres significatifs corrects de x_k par rapport à $u^* \neq 0$ est le nombre réel défini par*

$$\sigma_n = -\log_{10} \frac{\mid u_n - u^* \mid}{\mid u^* \mid}$$

12.3 Les outils du calcul numérique

Après avoir posé les bases théoriques du calcul numérique, il est temps d'envisager ses aspects matériels.

Un point de comparaison : Du premier ordinateur à ceux de l'an 2000, les progrès ont été tels que, dans le cas d'une automobile, celle-ci aurait les caractéristiques suivantes :

— Son prix serait inférieur à 1€ ;
— 10 000 km/h pour sa vitesse ;
— 1 ml de consommation aux 100 km ; et,
— 100 g en poids.

Des supercalculateurs aux clusters HPC. Dans les années 70-80 Fujitsu, Nec, Control Data puis Cray proposent des supercalculateurs de plus en plus puissants. On fabrique ces engins comme des Formules 1, chaque élément étant le summum de ce que permettent la recherche et la technologie. Les coûts sont très élevés et la puissance double tous les 18 mois, comme le prévoit la loi de Moore. Les premiers supercalculateurs étaient de simples ordinateurs mono-processeurs (mais possédant parfois jusqu'à dix processeurs périphériques pour les entrées-sorties). Dans les années 1970, la plupart des supercalculateurs ont adopté un processeur vectoriel, qui effectue le décodage d'une instruction une seule fois pour l'appliquer à toute une série d'opérandes. C'est seulement vers la fin des années 1980 que la technique des systèmes massivement parallèles a été adoptée, avec l'utilisation dans un même superordinateur de milliers de processeurs.

Aujourd'hui, le supercalculateur est basé sur l'emploi de composants standards d'origine Intel, AMD ou PowerPC (les COTS) dont la puissance brute est remarquable. On les assemble entre eux pour obtenir une puissance encore supérieure. Par exemple, un seul microprocesseur Itanium d'Intel délivre 6,4 gigaflops, soit les performances de 50 à 60 Cray-1 de 1976 (100 mégaflops en puissance de crête). Il est fabriqué par dizaine de millions d'unités à un coût comparativement très faible.

Les clés du succès de l'architecture cluster HPC

— La disponibilité de sous-ensembles standards (microprocesseurs, cartes mères, disques et cartes d'interface de réseau) produits en masse, fiables et de faible coût.
— L'existence des logiciels Open source, comme Linux, les compilateurs GNU et les outils de programmation parallèles (MPI, PVM).
— L'expérience accumulée par les chercheurs dans les algorithmes parallèles.
— Le développement des réseaux à hautes performances à base de composants standard.
— L'augmentation des besoins de calcul dans tous les domaines

Le concept de clusters HPC n'a fait que se développer (dans la liste des 500 plus grands supercalculateurs, 296 sont à base de clusters HPC). Le besoin en moyens de calcul pour répondre à des problématiques telles que la modélisation et la simulation de systèmes complexes, et l'évolution rapide des moyens informatiques (augmentation de la vitesse des processeurs et des mémoires ainsi que de la vitesse des réseaux) ont donné naissance aux clusters de PC (grappes de PC).

Le rapport prix / performance d'une grappe de PC est de 3 à 10 fois inférieur à celui des supercalculateurs traditionnels.

Les clusters sont composés d'un ensemble de serveurs interconnectés entre eux par un réseau rapide (cf. Gigabit Ethernet, Dolphin, Infiniband, Myrinet, ou PathScale). Ils permettent de répondre aux problématiques de Haute Performance et de Haute Disponibilité.

Les processeurs étant multi coeurs, on parle plus de nombres de coeurs que de de nombre de processeurs. Pour une idée de tarif, un cluster de 128 coeurs se négocie actuellement autour de 10k€ (2011), c'est moins d'un tiers du salaire annuel brut d'un ingénieur débutant pour une machine pouvant servir à de nombreux utilisateurs. Ce genre d'équipement est donc accessible par toute entreprise et n'est plus réservé aux centres de calcul.

Un niveau supérieur d'intégration des moyens de calcul a récemment été franchi. Les différents centres de calcul, comme le CRI de l'université de Lille, étant reliés entre eux par un réseau rapide, ils offrent de mutualiser leurs ressources pour créer un gigantesque HPC délocalisé sur l'Europe entière.

L'utilisateur soumettant son calcul (on parle de job) à exécution ne sait pas où exactement celui-ci sera exécuté, mieux encore, une partie du calcul peut être exécuté à Londres, l'autre à Paris ...

Avec quoi calculer Voici un exemple des moyens de calcul dont je dispose :
- sur mon bureau : papier-crayon, calculatrice, un PC octo–coeurs ;
- dans mon laboratoire : un cluster de 256 processeurs ;
- dans mon université : un supercalculateur Bluegene/L d'IBM 2048 coeurs ;
- plus globalement, les grilles nationale et européenne de calcul.

Le calcul parallèle L'évolution architecturale des machines de calcul vers la parallélisation massive semble aujourd'hui irréversible. Pour tirer parti de la puissance de calcul théorique, il est indispensable de concevoir d'emblée le code numérique comme un code parallèle, c–à–d. en tenant compte de la communication entre les différents processeurs exécutant les différentes parties du code.

Il s'avère dans la pratique que la multiplication du nombre d'unités de calcul ne divise pas spontanément le temps d'exécution des programmes. De nombreuses autres considérations entrent en jeu comme la perte de temps pour l'attente de données calculées par une autre tâche.

De façon idéale, l'accélération due à la parallélisation devrait être linéaire, en doublant le nombre d'unités de calcul, on devrait réduire de moitié le temps d'exécution, et ainsi de suite. Malheureusement très peu de programmes peuvent prétendre à de telles performances. Dans les années 1960, Gene Amdahl formula une loi empirique éponyme restée célèbre, elle aussi fut développée par d'autres auteurs. Dans sa version originale, la loi d'Amdahl s'articule sur une simple règle de trois. Elle indique le gain de temps que va apporter un système multiprocesseur en fonction :
- du nombre de processeurs N
- de la proportion d'activité parallélisable s

le tout en négligeant à ce stade le surcroît d'activité lié à la gestion du parallélisme lui-même. La loi a la forme :

$$R = \frac{1}{(1-s) + \frac{s}{N}}$$

Avec N tendant vers l'infini, on obtient : $R = \frac{1}{(1-s)}$.

Ce que montre la loi d'Amdahl, c'est que la fraction du temps d'exécution qui peut tirer profit de l'amélioration limite le gain de performance global, quelle que soit la valeur de l'amélioration de la composante.

En ce qui concerne l'analyse numérique, l'algorithme séquentiel optimal n'est pas forcément le plus facilement parallélisable. On tirera souvent profit d'un changement d'algorithme. Pour résoudre un système linéaire par exemple on remplacera la Méthode de Gauss-Seidel par la Méthode de Jacobi qui converge plus lentement, mais qui est totalement parallélisable.

12.4 Problèmes

12.4.1 Quelques applications directes

Exercice : Considérons l'équation du second degré $10^{-8}x^2 - 0.8x + 10^{-8} = 0$. Cette équation admet deux racines $r_1 \simeq 0,8 \times 10^8$ et $r_2 \simeq 1,25 \times 10^{-8}$. Le calcul de r_1 est souvent entaché d'erreur, essayer d'en trouver l'origine. Pour obtenir une meilleure valeur, on peut utiliser l'expression du produit des racines $r_1 r_2 = 1$.

Exercice : La terre est continuellement bombardée par des poussières cosmiques. La masse totale de poussières reçue par an est estimée à 5×10^7 kg/a.

1. En supposant que la densité des poussières vaut $\rho = 2,5 \times 10^3$ kg/m^3 donner l'accroissement de volume annuel de la Terre.
2. En déduire la variation annuelle du rayon de la Terre.
3. Faîtes l'application numérique. Conclusion.

4. Utiliser un développement limité pour calculer la variation du rayon de la Terre.
5. De quelle quantité le rayon de la Terre a–t–il varié au cours des Âges.

Exercice : Soit le calcul des intégrales $I_n = \int_0^1 \frac{x^n}{x+10}$ avec $n \in \mathbb{N}^*$; montrer que $0 \leq I_n \leq 1$ ainsi que $I_{n+1} = \frac{1}{n} - 10\, I_n$. Calculer I_0 et les 30 premiers éléments de la suite. Conclure.

12.4.2 Méthode d'accélération de la convergence

Voici sous forme de problème deux méthodes classiques pour accélérer la convergence d'un algorithme.

Le procédé Δ^2 d'Aitken (1926) *Voici un algorithme très utilisé dans les codes numériques.*

Étant donnée une suite $S_0, S_1, S_2, \ldots$ qui converge lentement vers la valeur de S. Le but est de trouver une autre suite avec la même limite, mais qui converge plus rapidement.

Souvent, on peut observer que la suite satisfait

$$S_{n+1} - S \approx \rho(S_n - S)$$

ou de façon équivalente $S_n \approx S + C\ \rho^n$.

L'idée est de remplacer $\approx$ par $=$ et de déterminer ρ, C et S de trois formules consécutives.

1. En notant
$$\Delta S_n = S_{n+1} - S_n$$
(différence finie) calculer les expressions de ΔS_n, ΔS_{n+1} et $\Delta^2 S_n = \Delta(\Delta S_n) = \Delta S_{n+1} - \Delta S_n$ dans l'hypothèse où $S_n = S + C\ \rho^n$.
2. En déduire S.
Si $S_n = S + C\ \rho^n$ n'est pas satisfait exactement, la valeur de S déterminer va dépendre de n. On obtient ainsi une autre suite S'_n définie par (procédé Δ^2 d'Aitken)
$$S'_n = S_{n+1} - \frac{\Delta S_n \cdot \Delta S_{n+1}}{\Delta^2 S_n},$$
qui en générale converge plus vite.
3. Montrer que la suite S'_n converge vers S.
4. Utiliser le procédé Δ^2 d'Aitken pour le calcul de la suite
$$S_n = \sum_{i=1}^{n} \frac{(-1)^{i+1}}{i}$$
qui converge vers $\ln(2)$.
5. Considérons la suite x_n donnée par $x_{n+1} = x_n + 1 - x_n^2/2,\ x_0 = 0$.
— Quelle est sa limite ?
— Appliquer l'algorithme Δ^2 d'Aitken pour accélérer la convergence.
— En utilisant $x_0, x_1, \ldots, x_8$ comparer l'erreur des suites obtenues avec ou sans Δ^2 d'Aitken.

L'extrapolation de Richardson En analyse numérique, le procédé d'extrapolation de Richardson est une technique d'accélération de la convergence proche du Δ^2 d'Aitken. Ce procédé est notamment utilisé pour définir une méthode numérique d'intégration : la méthode de Romberg.

On suppose que la quantité inconnue A peut être approchée par une fonction $A(h)$ avec une convergence d'ordre n en h, c–à–d.

$$A = A(h) + a_n\, h^n + O(h^m),\ a_n \neq 0,\ m > n,$$

où a_n est un coefficient inconnu. Le principe d'extrapolation consiste à supprimer le terme en h^n par combinaison linéaire de deux valeurs de $A(h)$, calculées avec des h différents : par exemple $A(h)$ et $A(h/2)$.

1. Montrer que la quantité
$$R(h) = \frac{2^n A(h/2) - A(h)}{2^n - 1},$$
qui est l'extrapolée de Richardson, approche A à l'ordre $m > n$ en h.
2. Déterminer $R(h)$ à partir de $A(h)$ et $A(h/3)$.
3. Utiliser l'extrapolation pour améliorer l'algorithme dévaluation de la dérivée d'une fonction par la méthode des différences finies.

L'intérêt de la méthode est qu'il sera fréquemment plus aisé d'obtenir une précision donnée en calculant $R(h)$ que $A(h')$ avec un h' beaucoup plus petit (risque d'erreur d'arrondi, grande quantité de calcul).

12.4.3 Problème de synthèse : Autour de π (Novembre 2012)

De nombreuses formules, de physique, d'ingénierie et bien sûr de mathématiques, impliquent π, qui est une des constantes les plus importantes des mathématiques.

Le nombre π est irrationnel, c'est-à-dire qu'on ne peut pas l'exprimer comme un rapport de deux nombres entiers ; ceci entraîne que son écriture décimale n'est ni finie, ni périodique. C'est même un nombre transcendant, ce qui signifie qu'il n'existe pas de polynôme non nul à coefficients entiers dont π soit une racine.

La détermination d'une valeur approchée suffisamment précise de π, et la compréhension de sa nature sont des enjeux qui ont traversé l'histoire des mathématiques ; la fascination exercée par ce nombre l'a même fait entrer dans la culture populaire.

Cette partie vise l'étude de quelques méthodes numériques d'approximation de π. Une grande famille de méthode utilise la relation
$$\tan \frac{\pi}{4} = 1$$
pour déterminer une approximation de π par la fonction réciproque
$$\pi = 4 \ \arctan(1).$$

Calcul en série

Formules de Leibnitz La formule de Leibnitz utilise le développement en série entière de la fonction arctan évalué en 1 pour trouver une approximation de π. Pour établir ce développement, on utilise la relation
$$\frac{d}{dx} \arctan(x) = \frac{1}{1+x^2} \quad \text{soit} \quad \arctan(x) = \int_0^x \frac{1}{1+u^2}\, du.$$

1. Utiliser la formule de Taylor Young pour retrouver le développement en série entière de la fonction $\frac{1}{1+u}$:
$$\frac{1}{1+u} = 1 - u + u^2 - u^3 + \cdots = \sum_{n=0}^{\infty} (-1)^n \, u^n.$$
2. Expliquer comment on peut en déduire le développement en série de $\frac{1}{1+u^2}$.
$$\frac{1}{1+u^2} = 1 - u^2 + u^4 - u^6 + \cdots = \sum_{n=0}^{\infty} (-1)^n \, u^{2n}.$$
3. Montrer qu'en intégrant la relation précédente entre $u = 0$ et $u = x$ on retrouve simplement le développement en série de la fonction arctan :
$$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} \cdots = \sum_{n=0}^{\infty} (-1)^n \, \frac{x^{2n+1}}{2n+1}.$$
4. En déduire la formule de Leibnitz en évaluant la série en $x = 1$.

5. La table 12.3 présente le calcul numérique des N premiers termes de la série ainsi que le nombre de chiffres significatifs σ_N. Commentez les résultats numériques en terme de convergence et de vitesse de convergence. La vitesse de convergence est-elle linéaire ? Justifier votre réponse.
6. Combien de termes faut-il conserver pour obtenir une approximation de π à une précision relative de 10^{-3} ? Cette méthode vous semble-t-elle efficace ?
7. Sachant que le rayon de convergence de la série entière définissant la fonction arctan vaut 1, comment peut-on expliquer les résultats précédents ?

N	S_N	ε_N	σ_N
200	3.1366421888702996	4.95046471949356359E-003	2.8025039029245380
400	3.1391050952489445	2.48755834084857241E-003	3.1013765976235121
600	3.1399315251675919	1.66112842220123014E-003	3.2767466635435345
800	3.1403457712814098	1.24688230838332359E-003	3.4013244097784217
1000	3.1405946498462800	9.98003743513109498E-004	3.4980177023646410
1200	3.1407607069783365	8.31946611456579888E-004	3.5770544155064838
1400	3.1408793869187388	7.13266671054313406E-004	3.6438979415591541
1600	3.1409684339252517	6.24219664541403318E-004	3.7018124267446662
1800	3.1410377146757980	5.54938913995162153E-004	3.7529046927694805
2000	3.1410931531214445	4.99500468348568205E-004	3.7986139729227166

TABLE 12.3 – **Approximation de π par la formule de Leibnitz.** S_N est la valeur des N premiers termes de la série de Leibnitz, σ_N est le nombre de chiffres significatifs correspondant.

Formules de Machin La formule de Machin utilise aussi le développement en série entière de la fonction arctan pour approché π mais elle utilise des évaluations de arctan en des points plus proches de zéro que la formule de Leibnitz. Les formules de Machin utilisent l'identité trigonométrique

$$\arctan a + \arctan b = \arctan \frac{a+b}{1-ab} \quad \text{si} \quad ab < 1.$$

La plus simple des formules de Machin est la formule d'Euler ici envisagée.

1. Vérifier l'identité
$$\arctan \frac{1}{2} + \arctan \frac{1}{3} = \frac{\pi}{4}.$$
2. Utiliser le développement en série de la fonction arctan pour déterminer une série dont la limite vaut π (formule d'Euler).
3. La table 12.4 présente le calcul numérique utilisant les N premiers termes des développement de arctan ainsi que le nombre de chiffres significatifs. Commentez les résultats numériques en terme de convergence et de vitesse de convergence. Comparer l'efficacité relative des formules d'Euler et de Leibnitz.
4. La vitesse de convergence de la formule d'Euler est-elle plutôt linéaire ou quadratique ? Justifier votre réponse.

Intégration numérique

Une forme alternative d'approximation de π consiste à évaluer numériquement l'intégrale

$$\int_0^1 \frac{4}{1+x^2}\, dx = \pi.$$

1. Donner une définition d'une formule de quadrature en précisant les notions de poids, nœuds, étage et ordre.

N	S_N	ε_N	σ_N
2	3.1408505617610554	7.42091828737745374E-004	3.6266922231475869
4	3.1415615878775909	3.10657122022384158E-005	5.0048685581392833
6	3.1415911843609066	1.46922888655254269E-006	6.3300604142602985
8	3.1415925796063511	7.39834420393492564E-008	7.6280153399304709
10	3.1415926497167881	3.87300502779908129E-009	8.9091018119218433
12	3.1415926533815393	2.08253858602347464E-010	10.178556815776496
14	3.1415926535783725	1.14206422097140603E-011	11.439459346692093
16	3.1415926535891572	6.35935748505289666E-013	12.693736633585338
18	3.1415926535897571	3.59712259978550719E-014	13.941194632678526
20	3.1415926535897909	2.22044604925031308E-015	15.150709647221156

TABLE 12.4 – **Approximation de π par la formule de Machin (Euler).** S_N est la valeur de la formule d'Euler tronquée aux N premiers termes des séries entière des fonction arctan, σ_N est le nombre de chiffres significatifs correspondant.

2. Expliciter les particularités des formules de quadrature de Newton–Cotes et de Gauss.
3. Expliquer en vous aidant d'une illustration graphique le principe de la formule du trapèze.
4. Utiliser les méthodes simples de quadrature suivantes pour obtenir une approximation de π
   — Trapèze
   — Simpson
   — Gauss d'ordre 6.
5. Qu'appelle-t–on une formule de quadrature composite ?
6. La figure 12.2 représente l'erreur d'évaluation de l'intégrale $\int_0^1 \frac{4}{1+x^2}\,dx$ par deux méthodes composites en fonction du nombre fe d'évaluations numériques de la fonction $f(x) = \frac{4}{1+x^2}$. Les calculs numériques sont effectués avec un codage des réels sur 8 octets, ce qui correspond à une précision machine $\epsilon = 10^{-16}$.
   — Expliquer l'origine du plateau horizontal indiqué par la zone hachurée pour les faibles erreurs.
   — Déterminer une estimation de l'ordre de chacune des méthodes (sans tenir compte de la zone hachurée). Quelle courbe représente la meilleur méthode.
   — Comparer quantitativement les performances de la meilleur méthode d'intégration numérique présentée sur la figure 12.2 et celle de la méthode de Machin dont les résultats sont présenté dans la table 12.4

Racine d'équation

On peut enfin envisager d'approcher π comme la racine de l'équation $\sin x = 0$ comprise dans l'intervalle $[3, 4]$.

1. Expliquer, en vous aidant d'une représentation graphique, le principe de la méthode de Newton pour la recherche d'une solution de l'équation $f(x) = 0$ qui conduit au schéma itératif suivant
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$
2. Calculer les itérés successif jusqu'à x_4 de la méthode de Newton en partant de $x_0 = 3$. *Pour faciliter les calculs, remarquer d'abord que f/f' correspond à une fonction élémentaire.*
3. Vos résultats numériques illustrent–ils ce que vous savez de la vitesse de convergence de la méthode de Newton ? (Soyez précis dans votre réponse).
4. Qu'advient-il si on initie la méthode avec $x_0 = 1$? Donner les résultats des premiers itérés et une explication graphique du problème.
5. Expliquer le principe d'une autre méthode de recherche de racine garantissant la convergence dans un intervalle défini par l'utilisateur.

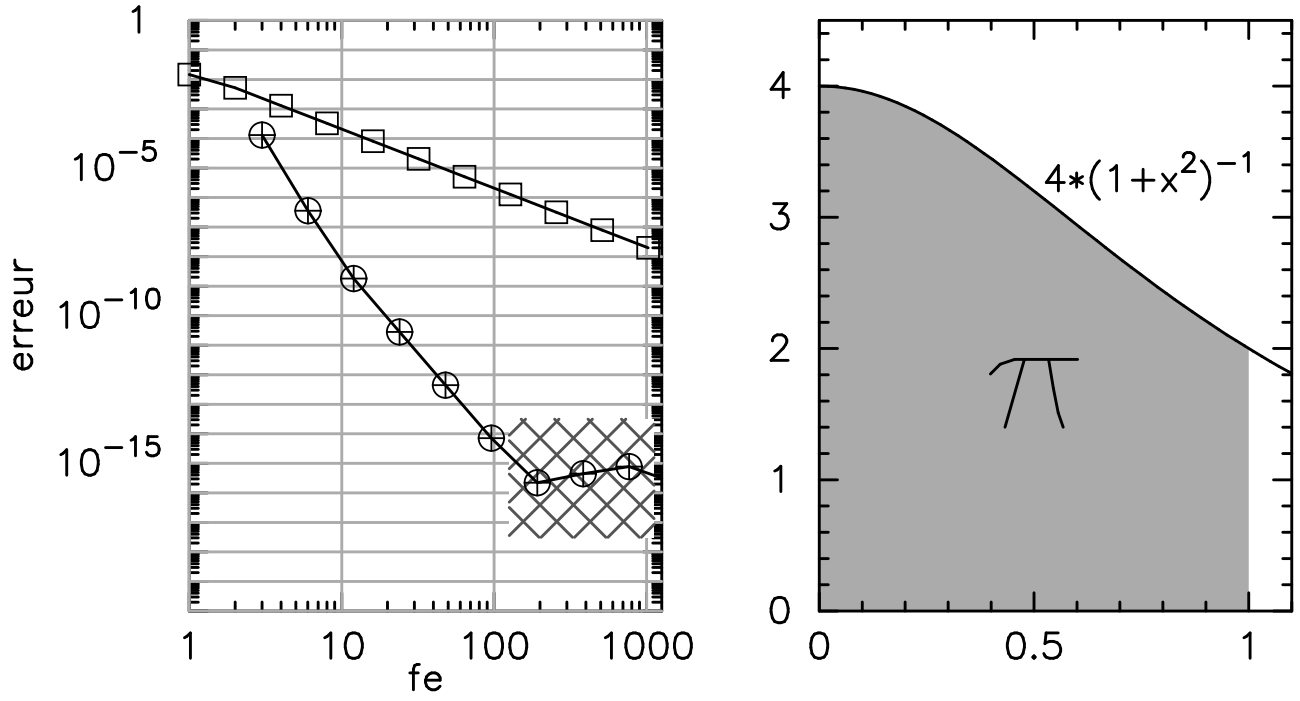


FIGURE 12.2 – **Performance de deux calculs numériques de $\int_0^1 \frac{4}{1+x^2}\,dx$ par des formules de quadratures composites.** sur la figure de droite, le trait plein représente la fonction $\frac{4}{1+x^2}$ et la zone grisée l'intégrale $\int_0^1 \frac{4}{1+x^2}\,dx$.

Chapitre 13

Résolution d'équations

Sommaire

La résolution d'équation à une variable est la recherche des valeurs de la variable réelle x telles que

$$f(x) = 0$$

où f est une fonction donnée. En mathématique on appelle un **zéro de** f une telle solution. Si f est une fonction linéaire ou quadratique, on utilisera bien évidemment les méthodes de résolution analytique (dans la mesure du possible, il ne faut pas être idiot). Si on connaît une solution d'un problème voisin, on utilisera une méthode de perturbation (par exemple pour $x^3 + 10^{-5}x - 1 = 0$). Dans les autres cas, on est contraint d'utiliser une méthode numérique pour trouver une approximation d'une solution.

Les méthodes numériques permettent dans le meilleur des cas de déterminer une solution en fonction des conditions initiales données. Aucune méthode "simple" ne vous donnera l'ensemble des solutions.

13.1 Méthode de la bissection

La méthode de dichotomie ou méthode de la bissection est un algorithme de recherche d'un zéro d'une fonction qui consiste à répéter des partages d'un intervalle en deux parties égales puis à sélectionner le sous-intervalle dans lequel existe un zéro de la fonction. La méthode requiert

— la connaissance préalable d'un intervalle $[a, b]$ encadrant le zéro voulu.

— LA continuité de f sur $[a, b]$.

Ces contraintes ne sont pas toujours évidente à satisfaire.

Algorithme La méthode de dichotomie commence par deux points a_0 et b_0 tels que $f(a_0)$ et $f(b_0)$ soient de signes opposés avec f continue sur $[a_0, b_0]$, ce qui implique d'après le théorème des valeurs intermédiaires que la fonction f possède au moins un zéro dans l'intervalle $[a_0, b_0]$. La méthode consiste à produire une suite décroissante d'intervalles $[a_k, b_k]$ qui contiennent tous un zéro de f.

À l'étape k, le nombre

$$c_k = \frac{a_k + b_k}{2)}$$

est calculé, c_k est le milieu de l'intervalle $[a_k, b_k]$.

— Si $f(a_k) \cdot f(c_k) > 0$: $f(a_k)$ et $f(c_k)$ sont de mêmes signes, alors $a_{k+1} = c_k$ et $b_{k+1} = b_k$
— sinon $a_{k+1} = a_k$ et $b_{k+1} = c_k$.
Ce procédé est répété jusqu'à ce que le zéro soit suffisamment approché. L'algorithme de dichotomie est en soi récursif.

Convergence L'erreur absolue de la méthode de dichotomie est au plus

$$\frac{b-a}{2^{n+1}}$$

après n étapes. En d'autres termes, l'erreur est diminuée de moitié à chaque étape, ainsi la méthode converge linéairement, ce qui est très lent par comparaison avec la méthode de Newton.

L'avantage par rapport à cette dernière est son domaine d'application plus vaste : il suffit que $f(a)$ et $f(b)$ soient de signes opposés et qu'on puisse déterminer le signe de $f(c)$ à chaque itération. De plus, si on se donne la tolérance relative ϵ, on connaît en théorie le nombre maximum d'itérations nécessaires pour satisfaire cette tolérance :

$$2^{n+1} = 1/\epsilon$$

C'est un cas assez peu habituel en calcul numérique pour être noté.

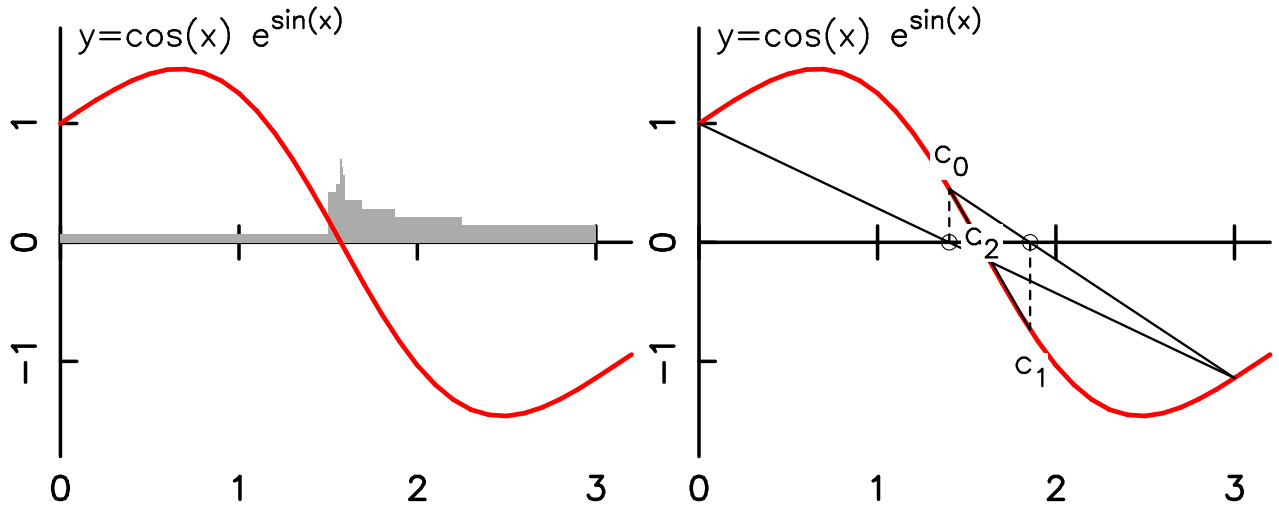


FIGURE 13.1 – Apllication de la méthode de la dichotomie (à gauche) et de la fausse position (à dorite) à la recherche d'un zéro de $f(x) = \cos(x)\exp(\sin(x))$ dans l'intervalle $[0, 3]$. Les intervalles successifs de la dichotomie sont représenté par les rectangle gris. Les sécantes des la suite des points d'intersection c_k sont représentées en pour les quatre première itérations.

13.2 Méthode de la fausse position

Comme la méthode de dichotomie, la méthode de la fausse position commence par deux points a_0 et b_0 tels que $f(a_0)$ et $f(b_0)$ soient de signes opposés, ce qui implique d'après le théorème des valeurs intermédiaires que la fonction f possède au moins un zéro dans l'intervalle $[a_0, b_0]$. La méthode consiste à produire une suite décroissante d'intervalles $[a_k, b_k]$ qui contiennent tous un zéro de f.

À l'étape k, le nombre

$$c_k = b_k - \frac{b_k - a_k}{f(a_k) - f(b_k)} f(b_k)$$

est calculé. Comme expliqué ci-dessous c_k est l'abscisse de l'intersection de la droite passant par $(a_k, f(a_k))$ et $(b_k, f(b_k))$ avec l'axe des abscisses, que nous appellerons pour simplifier zéro de la sécante.

— Si $f(a_k) \cdot f(c_k) > 0$: $f(a_k)$ et $f(c_k)$ sont de mêmes signes, alors $a_{k+1} = c_k$ et $b_{k+1} = b_k$
— sinon $a_{k+1} = a_k$ et $b_{k+1} = c_k$.

Ce procédé est répété jusqu'à ce que le zéro soit suffisamment approché.

La formule ci-dessus est également employée dans la méthode de la sécante, mais la méthode de la sécante retient systématiquement les deux derniers points calculés, alors que la méthode de la fausse position retient deux points qui encadrent certainement un zéro. D'autre part, la seule différence entre la méthode de la fausse position et la méthode de dichotomie est l'utilisation la relation $c_k = (a_k + b_k)/2$.

Recherche du zéro de la sécante Étant donnés a et b, nous construisons la droite passant par les points $(a, f(a))$ et $(b, f(b))$, comme dans la figure ci-contre. Remarquons que cette droite est une sécante ou une corde du graphe de la fonction f. En utilisant la pente et un point, l'équation de la droite peut s'écrire

$$y - f(b) = \frac{f(b) - f(a)}{b - a}(x - b).$$

Nous déterminons maintenant c, l'abscisse du point d'intersection de cette droite avec l'axe des abscisses (zéro de la sécante) donnée par

$$f(b) + \frac{f(b) - f(a)}{b - a}(c - b) = 0.$$

La résolution de l'équation précédente donne c_k.

Vitesse de convergence Si les valeurs initiales a_0 et b_0 sont prises telles que $f(a_0)$ et $f(b_0)$ soient de signes opposés, alors la méthode de fausse position convergera vers un zéro de f. La vitesse de convergence sera typiquement superlinéaire, ainsi plus rapide que la méthode de dichotomie, mais plus lente que la méthode de la sécante.

13.3 Méthode du point fixe

La méthode du point fixe permet la recherche par itération d'une Racine de l'équation $x = f(x)$ en cherchant la limite de la suite $u_{n+1} = f(u_n)$. Avant de détailler la méthode, il faut faire un détour par la notion de différentiabilité d'une fonction .

> **Différentiabilité au sens de Fréchet :** *Soit f une fonction de $\mathbb{R}$ dans $\mathbb{R}$, on dit que f est différentiable au sens de Fréchet au voisinage de x si*
>
> $$f(x+h) = f(x) + h\ f'(x) + \Phi(x,h) \quad avec \quad \lim_{h\to 0} \frac{\Phi(x,h)}{h} = 0$$
>
> *f' est dit différentielle de f en x.*

Avec des mots simples, une fonction différentiable en x est une fonction que l'on peut approcher par une droite dont le coefficient directeur est la dérivée de la fonction au point considéré.

Le résultat suivant statue sur la convergence de la méthode de l'itération

> **Convergence de la méthode de l'itération :** *Si la suite $u_{n+1} = f(u_n)$ admet un point fixe u^* (i.e. $f(u^*) = u^*$) avec f différentiable en u^* et $\mid f'(u^*) \mid < 1$ alors la suite (u_n) converge vers u^* et la valeur de $\mid f'(u^*)$ détermine la vitesse de convergence.*

Étudions maintenant la vitesse de convergence de la suite en utilisant la différentiabilité de f

$$u_{n+1} = f(u_n) = f\left(u^* + \underbrace{u_n - u^*}_{h}\right) = \underbrace{f(u^*)}_{u^*} + h\ f'(u^*) + \Phi(u^*, h)$$

soit

$$u_{n+1} - u^* = (u_n - u^*)\ f'(u^*) + \Phi(u^*, u_n - u^*)$$

et donc

$$\frac{u_{n+1}-u^*}{u_n-u^*} = f'(u^*) + \frac{\Phi(u^*, u_n-u^*)}{u_n-u^*} \to f'(u^*)$$

On retrouve la conclusion du théorème précèdent : la suite converge si $\mid f'(u^*) \mid < 1$. L'ordre de la suite est 1 et le nombre de chiffres significatifs suit une progression arithmétique de raison $\mid f'(u^*) \mid$:

$$\sigma_{n+1} = \sigma_{n+1} - \log_{10} \mid f'(u^*) \mid .$$

Les résultats précédents sous-entendent bien sûr $f'(u^*) \neq 0$. Ils se généralisent dans le cas $f'(u^*) = 0$ et l'on montre que la méthode est ordre ≥ 2.

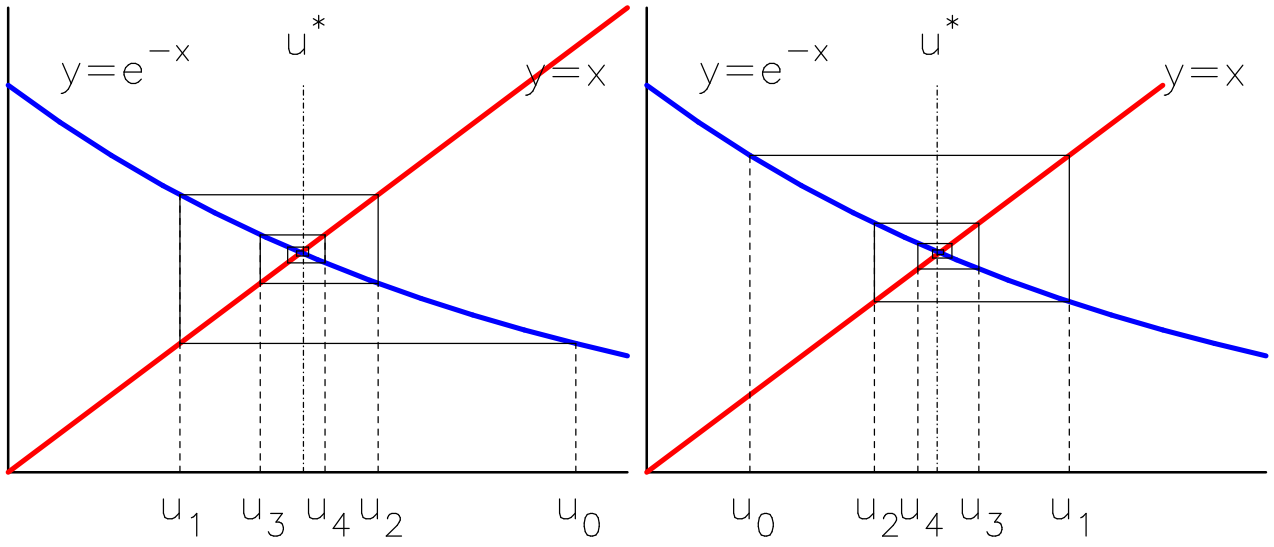


FIGURE 13.2 – Itération de la méthode du point fixe sur un exemple simple.

Exercice : Étudier la convergence de la suite $u_{n+1} = \exp(-u_n)$. Calculer l'ordre et déterminer la progression de l'erreur. Calculer la racine de $x = \exp(-x)$.

13.4 Méthode de Newton

La méthode de Newton ou méthode de Newton-Raphson est, dans son application la plus simple, un algorithme efficace pour trouver numériquement une approximation précise d'un zéro (ou racine) d'une fonction réelle d'une variable réelle. Cette méthode doit son nom aux mathématiciens anglais Isaac Newton (1643-1727) et Joseph Raphson (peut-être 1648-1715), qui furent les premiers à la décrire pour la recherche des zéros d'une équation polynomiale. On n'oubliera pas Thomas Simpson (1710-1761) qui élargit considérablement le domaine d'application de l'algorithme en montrant, grâce à la notion de dérivée, comment on pouvait l'utiliser pour calculer un zéro d'une équation non linéaire, pouvant ne pas être un polynôme, et d'un système formé de telles équations.

Sous sa forme moderne, l'algorithme peut être présenté brièvement comme suit : à chaque itération, la fonction dont on cherche un zéro est linéarisée en l'itéré (ou point) courant et l'itéré suivant est pris égal au zéro de la fonction linéarisée. Cette description sommaire indique qu'au moins deux conditions sont requises pour la bonne marche de l'algorithme : la fonction doit être différentiable aux points visités (pour pouvoir y linéariser la fonction) et les dérivées ne doivent pas s'y annuler (pour que la fonction linéarisée ait un zéro) ; s'ajoute à ces conditions la contrainte forte de devoir prendre le premier itéré assez proche d'un zéro régulier de la fonction (i.e., en lequel la dérivée de la fonction ne s'annule pas), pour que la convergence du processus soit assurée.

L'intérêt principal de l'algorithme de Newton est sa convergence quadratique locale. En termes imagés, mais peu précis, cela signifie que le nombre de chiffres significatifs corrects des itérés double à chaque itération, asymptotiquement. Comme le nombre de chiffres significatifs représentables par un ordinateur est limité (environ 15 chiffres décimaux sur un ordinateur avec un processeur 32-bits), on

peut simplifier grossièrement les propriétés de convergence de l'algorithme de Newton en disant que, soit il converge en moins de 10 itérations, soit il diverge. En effet, si l'itéré initial n'est pas pris suffisamment proche d'un zéro, la suite des itérés générée par l'algorithme a un comportement erratique, dont la convergence éventuelle ne peut être que le fruit du hasard (un des itérés est par chance proche d'un zéro).

L'algorithme On va donc chercher à construire une bonne approximation d'un zéro de la fonction d'une variable réelle $f(x)$ en se basant sur son développement de Taylor au premier ordre. Pour cela, partant d'un point x_0 que l'on choisit de préférence proche du zéro à trouver (en faisant des estimations grossières par exemple), on approche la fonction au premier ordre, autrement dit, on la considère à peu près égale à sa tangente en ce point :

$$f(x) \simeq f(x_0) + (x - x_0) \cdot f'(x_0)$$

Partant de là, pour trouver un zéro de cette fonction d'approximation, il suffit de calculer l'intersection de la droite tangente avec l'axe des abscisses, c'est-à-dire résoudre l'équation affine

$$0 \simeq f(x_0) + (x - x_0) \cdot f'(x_0)$$

On obtient alors un point x_1 qui en général a de bonnes chances d'être plus proche du vrai zéro de f que le point x_0 précédent. Par cette opération, on peut donc espérer améliorer l'approximation par itérations successives (voir figure 2.3) : on approche à nouveau la fonction par sa tangente en x_1 pour obtenir un nouveau point x_2 etc.

Cette méthode requiert que la fonction possède une tangente en chacun des points de la suite que l'on construit par itération, par exemple il suffit que f soit dérivable.

Formellement, on part d'un point x_0 appartenant à l'ensemble de définition de la fonction et on construit par récurrence la suite

$$u_{n+1} = u_n - \frac{f(u_n)}{f'(u_n)}.$$

Il se peut que la récurrence doive se terminer, si à l'étape n, u_n n'appartient pas au domaine de définition ou si la dérivée $f'(u_n)$ est nulle ; dans ces cas, la méthode échoue.

Bien que la méthode soit très efficace, certains aspects pratiques doivent être pris en compte. Avant tout, la méthode de Newton nécessite que la dérivée soit effectivement calculée. Dans les cas où la dérivée est seulement estimée en prenant la pente entre deux points de la fonction, la méthode prend le nom de méthode de la sécante, moins efficace (d'ordre 1,618 qui est le nombre d'or) et inférieure à d'autres algorithmes. Par ailleurs, si la valeur de départ est trop éloignée du vrai zéro, la méthode de Newton peut entrer en boucle infinie sans produire d'approximation améliorée. À cause de cela, toute mise en oeuvre de la méthode de Newton doit inclure un code de contrôle du nombre d'itérations.

Convergence de la méthode de Newton Pour étudier la convergence de la méthode de Newton, on utilise encore une fois le développement de Taylor de $g(x) = x - \frac{f(x)}{f'(x)}$ au voisinage de u^* telle que $g(u^*) = u^*$

$$g(u^* + h) = g(u^*) + h\, g'(u^*) + \frac{1}{2}h^2\, g''(u^*) + \Phi(u^*, h).$$

Le schéma itératif de la méthode de Newton devient

$$u_{n+1} = g(u_n) = g(u^* + u_n - u^*) = g(u^*) + (u_n - u^*)\, g'(u^*) + \frac{1}{2}(u_n - u^*)^2\, g''(u^*) + \Phi(u^*, (u_n - u^*)).$$

En remarquant que $g(u^*) = u^*$, $g'(u^*) = 0$ et $g''(u^*) = \frac{f''(u^*)}{f'(u^*)}$ on obtient l'évolution de l'erreur

$$u_{n+1} - u^* = \frac{1}{2}\frac{f''(u^*)}{f'(u^*)}\,(u_n - u^*)^2 + \Phi(u^*, (u_n - u^*)).$$

La méthode de Newton est donc au moins d'ordre 2 et le nombre de chiffres significatif suit une progression arithmetico–géométrique de raison $|\, \frac{1}{2}\frac{f''(u^*)}{f'(u^*)}\, |$:

$$\sigma_{n+1} = 2\sigma_n - \log_{10} |\, \frac{1}{2}\frac{f''(u^*)}{f'(u^*)}\, | \,.$$

Exemples de non-convergence

— La tangente à la courbe peut couper l'axe des abscisses hors du domaine de définition de la fonction.

— Si l'on utilise l'algorithme de Newton pour trouver l'unique zéro $x^* = 0$ de la fonction $x \in \mathbb{R} \to \sqrt{|x|}$ en prenant un itéré initial $x_0 \neq 0$, on constate que, pour tout $k \in \mathbb{N}$, $x_{k+1} = -x_k$; la suite générée ne converge donc pas, même localement (c'est-à-dire même si x_0 est pris proche du zéro $x^* = 0$). Le problème provient ici, en particulier, de la non-différentiabilité de la fonction en l'unique zéro $x^* = 0$.

Critère d'arrêt Des critères d'arrêt possibles, déterminés relativement à une grandeur numériquement négligeable, sont :

$$\|f(x_k)\| < \varepsilon_1 \qquad \text{ou} \qquad \|x_{k+1} - x_k\| < \varepsilon_2$$

où $\varepsilon_1, \varepsilon_2 \in \mathbb{R}^+$ representent des erreurs d'approximations caractérisant la qualité de la solution numérique.

Dans tous les cas, il se peut que le critère d'arrêt soit vérifié en des points ne correspondant pas à des solutions de l'équation à résoudre.

Généralisation aux systèmes d'équations à plusieurs variables On peut aussi utiliser la méthode de Newton pour résoudre un système d' n équations (non linéaires) à n inconnues $x = (x_1, \ldots, x_n)$, ce qui revient à trouver un zéro d'une fonction F de $\mathbb{R}^n$ dans $\mathbb{R}^n$, qui devra être différentiable. Dans la formulation donnée ci-dessus, il faut multiplier par l'inverse de la matrice jacobienne $F'(x_k)$ au lieu de diviser par $f'(x_k)$. Évidemment, pour économiser du temps de calcul, on ne calculera pas l'inverse de la jacobienne, mais on résoudra le système d'équations linéaires suivant

$$F'(x_k)(x_{k+1} - x_k) = -F(x_k)$$

en l'inconnue $x_{k+1} - x_k$. Encore une fois, cette méthode ne fonctionne que pour une valeur initiale x_0 suffisamment proche d'un zéro de F.

Exercice : Montrer que l'algorithme de Babylone est une implémentation de la méthode de Newton à la fonction $f(x) = x^2 - A$

Exercice : Généraliser la méthode au calcul de racine n-ème

Exercice : Calculer la solution de $x = \cos(x)$ en utilisant la méthode de Newton.

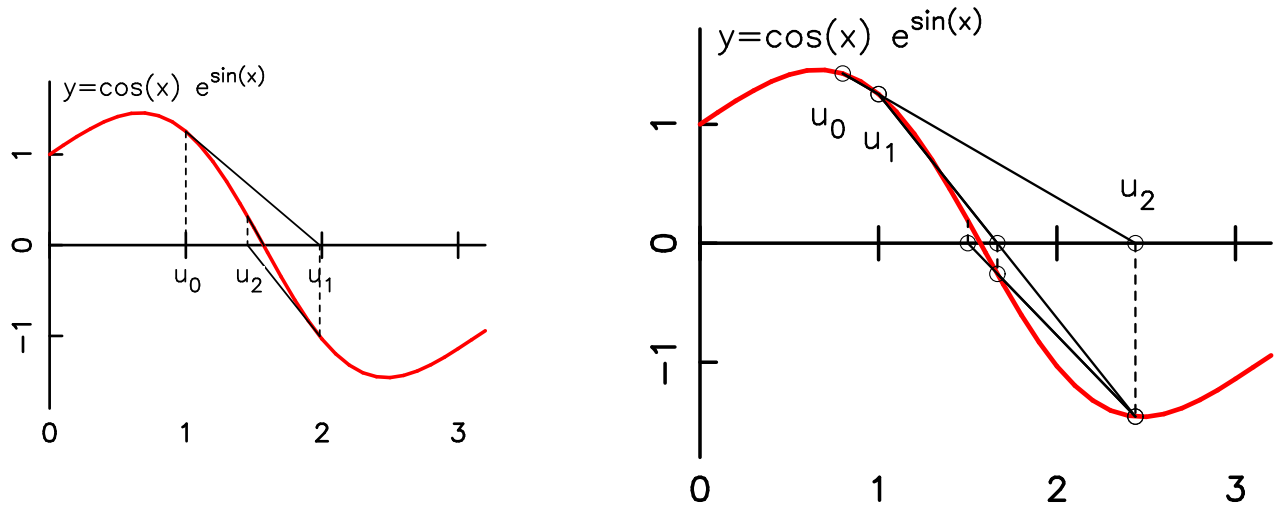


FIGURE 13.3 – Itération de la méthode de Newton (à gauche) et de la sécante (à droite) sur un exemple simple. La méthode de Newton est initialisée avec $u_0 = 1$, la méthode de la sécante avec $u_0 = 0,8$ et $u_1 = 1$

13.5 Méthode de la sécante

La méthode de la sécante est une adaptation de la méthode de Newton où le calcul de la dérivée est approché en utilisant les résultats des étapes n et $n-1$ du calcul :

$$f'(u_n) = \frac{f(u_n) - f(u_{n-1})}{u_n - u_{n-1}}.$$

Cette méthode est avantageuse lorsque le calcul de la dérivée est soit délicat soit coûteux. L'initialisation nécessite deux points u_0 et u_1, proches, si possible, de la solution recherchée. Il n'est pas nécessaire, contrairement à la méthode de la fausse position, que u_0 et u_1 encadre une racine de $f(x)$. Le schéma de la méthode de la sécante est alors le suivant

$$u_{n+1} = u_n - f(u_n) \ \frac{u_n - u_{n-1}}{f(u_n) - f(u_{n-1})}.$$

La méthode de la sécante converge toujours dans un voisinage proche de la solution et son ordre est $\frac{1+\sqrt{5}}{2} \simeq 1,6$.

On peut comparer les méthodes de Newton et de la sécante en supposant identique le coût numérique de l'évaluation de f et de f'. Dans ce cas, une itération de la méthode de Newton demande deux évaluations (f et f') pour doubler le nombre de chiffre significatif. Pour le même coût numérique, la méthode de la sécante fait deux itérations (deux évaluations de f) ce qui multiplie le nombre de chiffres significatifs par environ $2,5$ $(1,6^2)$.

Cela dit, la méthode de la sécante est sujette un un problème de stabilité lors de l'évaluation de la dérivée : à mesure que u_n tend vers u^* la différence $u_{n+1} - u_n$ devient petite et peut conduire à de grosses pertes (problème de la différence de nombres très proche). C'est pourquoi on implémente toujours l'algorithme sous la forme

$$\begin{aligned} \tau_n &= \frac{f(u_n) - f(u_{n-1})}{u_n - u_{n-1}} \\ u_{n+1} &= u_n - \frac{f(u_n)}{\tau_n}. \end{aligned}$$

Essayer de comprendre l'avantage de cette implémentation.

En pratique, on utilise la méthode de la sécante quand $\mid u_{n+1} - u_n \mid$ est grand et dès que

$$\mid u_{n+1} - u_n \mid \leq \sqrt{\varepsilon} \cdot \sup(u_n, u_{n+1}),$$

avec ε la précision relative de la machine, on garde le même τ_n pour la suite des calculs.

13.6 Interpolation quadratique inverse

L'idée est d'utiliser une interpolation quadratique afin d'approcher la fonction inverse de f. Cet algorithme est rarement utilisé seul, mais prend sa place dans la populaire Méthode de Brent.

La méthode L'algorithme de l'interpolation quadratique inverse est donné par la relation de récurrence :

$$x_{n+1} = \frac{f_{n-1}f_n}{(f_{n-2} - f_{n-1})(f_{n-2} - f_n)}x_{n-2} + \frac{f_{n-2}f_n}{(f_{n-1} - f_{n-2})(f_{n-1} - f_n)}x_{n-1} + \frac{f_{n-2}f_{n-1}}{(f_n - f_{n-2})(f_n - f_{n-1})}x_n,$$

où $f_k = f(x_k)$. Cette méthode nécessite donc trois valeurs initiales x_0, x_1 et x_2 .

Explication de la méthode On utilise les trois itérations précédentes x_{n-2}, x_{n-1} et x_n, avec leur image, f_{n-2}, f_{n-1} et f_n. En appliquant une interpolation lagrangienne quadratique, on trouve l'approximation suivante de l'inverse de f

$$f^{-1}(y) = \frac{(y-f_{n-1})(y-f_n)}{(f_{n-2}-f_{n-1})(f_{n-2}-f_n)}x_{n-2} + \frac{(y-f_{n-2})(y-f_n)}{(f_{n-1}-f_{n-2})(f_{n-1}-f_n)}x_{n-1} + \frac{(y-f_{n-2})(y-f_{n-1})}{(f_n-f_{n-2})(f_n-f_{n-1})}x_n.$$

On recherche une racine de f, donc on remplace $y = f(x) = 0$ dans la relation précédente et on obtient la relation souhaitée.

Comportement Le comportement asymptotique est très bon : en général, les itérations x_n convergent rapidement vers la racine, une fois qu'elles en sont proches. Toutefois, les performances sont souvent assez pauvres si on part trop loin de la racine. Par exemple, si par malchance deux des images f_{n-2}, f_{n-1} et f_n coïncident, l'algorithme échoue complètement.

13.7 Méthode de Brent

La méthode de Brent combine la méthode de dichotomie, la méthode de la sécante et l'interpolation quadratique inverse. À chaque itération, elle décide laquelle de ces trois méthodes est susceptible d'approcher au mieux le zéro, et effectue une itération en utilisant cette méthode. L'idée principale est d'utiliser la méthode de la sécante ou d'interpolation quadratique inverse parce qu'elles convergent vite, et de revenir à la robuste méthode de dichotomie si besoin est. Cela donne une méthode alliant robustesse et rapidité, qui se trouve être très populaire et très appréciée. L'idée d'allier ces méthodes différentes est due à Theodorus Dekker (1969) et à Richard Brent (1973).

La méthode de Dekker. L'idée de combiner les méthodes de dichotomie et de la sécante remonte à Dekker. À l'image de la méthode de dichotomie, la méthode de Dekker est initialisée par deux points, a_0 et b_0, tels que $f(a_0)$ et $f(b_0)$ aient des signes opposés. Si f est continue sur $[a_0, b_0]$, le théorème des valeurs intermédiaires indique l'existence d'une solution entre a_0 et b_0.

À chaque itération, trois points entrent en jeu :

— b_k est l'itération courante, c.-à-d. l'approximation courante de la racine de f ;

— a_k est le "contrepoint," c.-à-d. un point tel que $f(a_k)$ et $f(b_k)$ aient des signes opposés. Ainsi, l'intervalle $[a_k, b_k]$ contient à coup sûr la solution. De plus, $|f(b_k)|$ doit être plus petit (en magnitude) que $|f(a_k)|$, de telle sorte que b_k soit une meilleure approximation de la racine que a_k ;

— b_{k-1} est l'itération précédente (pour la première itération, on pose $b_{-1} = a_0$).

Deux candidats à la prochaine itération sont calculés : le premier est obtenu par la méthode de la sécante

$$s = b_k - \frac{b_k - b_{k-1}}{f(b_k) - f(b_{k-1})} f(b_k),$$

et le second par la méthode de dichotomie

$$m = \frac{a_k + b_k}{2}.$$

Si le résultat de la méthode de la sécante, s, tombe entre b_k et m, alors il peut devenir le prochain itéré ($b_{k+1} = s$), et dans le cas contraire, le point milieu entre en jeu ($b_{k+1} = m$).

Alors, la valeur du nouveau contrepoint est choisie de telle sorte que $f(a_{k+1})$ et $f(b_{k+1})$ aient des signes opposés. Si $f(a_k)$ et $f(b_{k+1})$ sont de signe opposé, alors le contrepoint ne change pas : $a_{k+1} = a_k$. Sinon, $f(b_{k+1})$ et $f(b_k)$ sont de signe opposé et le nouveau contrepoint devient $a_{k+1} = b_k$.

Finalement, si $|f(a_{k+1})| < |f(b_{k+1})|$, alors a_{k+1} est probablement une meilleure approximation de la solution que b_{k+1}, et par suite, les valeurs de a_{k+1} et b_{k+1} sont échangées.

En arrivant à ce point, la méthode vient de réaliser une itération. La méthode est répétée jusqu'à convergence.

La méthode de Brent. La méthode de Dekker est efficace si f se comporte raisonnablement bien. Toutefois, dans certaines circonstances, chaque itération emploie la méthode de la sécante, mais la suite des b_k converge très lentement (en particulier, $|b_k - b_{k-1}|$ peut devenir arbitrairement petit). Dans une telle configuration, la méthode de Dekker nécessite alors plus d'itérations que la méthode de dichotomie.

Brent propose une petite modification pour éviter ce problème : un test supplémentaire doit être vérifié avant que le résultat de la méthode de la sécante soit accepté comme la prochaine itération. En particulier, si l'étape précédente utilisait la méthode de dichotomie, l'inégalité

$$|s - b_k| < \tfrac{1}{2}|b_k - b_{k-1}|$$

doit être vraie, sinon le point milieu m devient le prochain itéré. Si l'étape précédente utilisait l'interpolation, alors l'inégalité

$$|s - b_k| < \tfrac{1}{2}|b_{k-1} - b_{k-2}|$$

est utilisée à la place.

Cette modification permet de remplacer la méthode de la sécante par la méthode de dichotomie lorsque la première progresse trop lentement. Brent a prouvé que cette méthode requiert au plus N^2 itérations, où N représente le nombre d'itérations pour la méthode de dichotomie. Si la fonction f se comporte bien, la méthode de Brent utilise au choix l'interpolation quadratique inverse ou l'interpolation linéaire, auquel cas la vitesse de convergence est superlinéaire.

La méthode de Brent se base sur l'interpolation quadratique inverse plutôt que l'interpolation linéaire (qui apparaît dans la méthode de la sécante) si $f(b_k)$, $f(a_k)$ et $f(b_{k-1})$ sont distincts. Ceci améliore significativement l'efficacité. Par conséquent, la condition pour accepter la valeur d'interpolation s est modifiée : s doit tomber entre $(3a_k + b_k)/4$ et b_k.

La librairie Netlib contient une implémentation en Fortran. La fonction MATLAB nommée fzero ou la fonction solve de PARI/GP sont des exemples d'implémentation de la méthode de Brent.

Chapitre 14

Intégration numérique

Sommaire

Rappelons brièvement la définition la plus simple de l'intégrale : l'intégrale de Riemann. Si f est une fonction réelle positive continue prenant ses valeurs dans un segment $I = [0, a]$, alors l'intégrale de f sur I, notée

$$\int_{x\in I} f(x)\,\mathrm{d}x,$$

est l'aire d'une surface délimitée par la représentation graphique de f et par les trois droites d'équation $x = 0$, $x = a$, $y = 0$. On note S_f cette surface :

$$S_f = \{(x, y) \in \mathbb{R}_+^2 \mid x \in I \text{ et } 0 \leq y \leq f(x)\}$$

On donne un signe positif à l'aire des surfaces comme S_f situées au-dessus de l'axe des abscisses. Pour pouvoir traiter aussi les fonctions négatives, on donne un signe négatif aux portions situées sous cet axe.

Ainsi, pour définir l'intégrale d'une fonction continue dans le cas général (positive ou négative), il suffit de définir f^+ et f^- comme suit :

$$f^+(x) = \begin{cases} f(x) & \text{si } f(x) > 0 \\ 0, & \text{sinon} \end{cases}$$

$$f^-(x) = \begin{cases} -f(x) & \text{si } f(x) < 0 \\ 0, & \text{sinon} \end{cases}$$

puis de définir l'intégrale de f à partir de $f^+ +$ et f^-, fonctions continues et positives :

$$\int_{x\in I} f\,\mathrm{d}x = \int_{x\in I} f^+\,\mathrm{d}x - \int_{x\in I} f^-\,\mathrm{d}x$$

L'**intégration numérique** est un thème pionner de l'analyse numérique puisqu'il remonte à Newton et les programmes qui ont tourné sur les premiers ordinateurs furent en grande partie les calculs d'intégrales : ces problèmes sont parmi les plus faciles à programmer.

Problème : *Étant donnée une fonction continue sur un intervalle borné f : $[a,b] \to \mathbb{R}$, on cherche une évaluation de l'intégrale*

$$\int_a^b f(x)dx$$

14.1 Formules de quadrature et leurs ordres

On appelle formule de quadrature une expression linéaire dont l'évaluation fournit une valeur approchée de l'intégrale sur l'intervalle $[0,1]$. En effet, une transformation affine permet de transposer la formule sur un morceau particulier :

$$\int_a^b f(x)dx = (b-a)\int_0^1 f\left(a+t\,(b-a)\right)dt.$$

En notant $g(t) = f\left(a+t\,(b-a)\right)$, l'intégrale voulue est équivalente à

$$\int_a^b f(x)\;dx = (b-a)\int_0^1 g(t)\,dt.$$

La formule de quadrature fait intervenir des valeurs pondérées de la fonction (et parfois également celles de sa dérivée) en certains nœuds : les coefficients de pondération et les nœuds dépendent de la méthode employée. Ces formules de quadrature sont en effet obtenues à l'aide de la substitution de la fonction par une approximation, c'est-à-dire par une fonction proche dont l'intégrale peut être déterminée algébriquement.

Définition : *Une formule de quadrature à s étages est donnée par*

$$\int_0^1 g(t)dt \approx \sum_{i=1}^{s} b_i g(c_i).$$

Les c_i sont les noeuds de la formule de quadrature et les b_i en sont les poids. On dit que l'ordre de la formule de quadrature est p si elle est exacte pour tous les polynômes de degrés inférieurs à p

Théorème : *La formule de quadrature a un ordre p si et seulement si*

$$\sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q} \quad pour \quad q = 1,2,\ldots,p$$

14.2 formules de Newton-Cotes

Les formules de Newton-Cotes utilisent l'interpolation des fonctions à intégrer par des polynômes dont la primitive est connue. Elles ont la particularité d'utiliser un maillage équidistant de l'intervalle $[a,b]$. On note $I = \int_a^b f(x)\,dx$, l'intégrale exacte et $I(f)$, la forme approché de l'intégrale. la valeur de $E(f) = I - I(f)$ définit la mesure de l'erreur de troncature de la méthode.

14.2.1 Formules simples

Formules du rectangle et du point milieu C'est la méthode la plus simple qui consiste à interpoler la fonction f à intégrer par une fonction constante (polynôme de degré 0). Si ξ est le point d'interpolation, la formule est la suivante :

$$I(f) = (b-a)f(\xi),$$

Le choix de ξ influence l'erreur de troncature :

— Si $\xi = a$, ou $\xi = b$, l'erreur est donnée par

$$E(f) = \frac{(b-a)^2}{2} f'(\eta), \quad \eta \in [a,b].$$

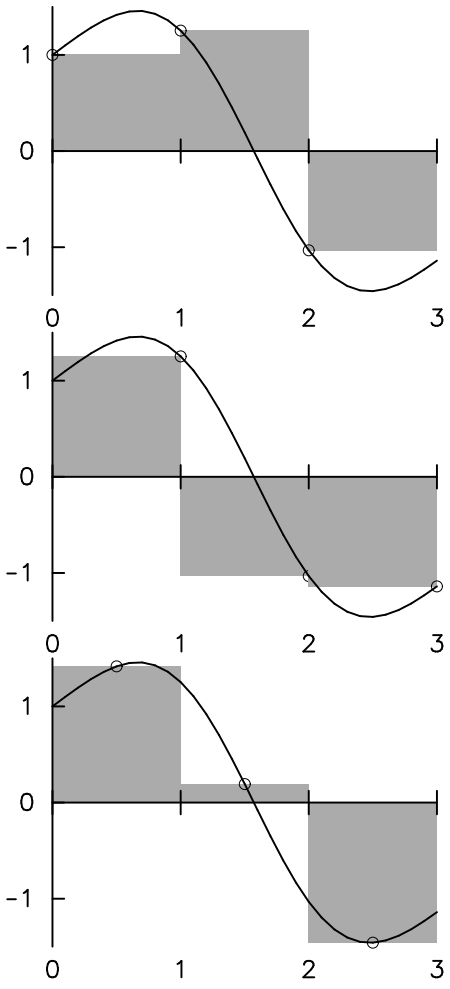
C'est la **méthode du rectangle** qui est d'ordre 0.

— Si $\xi = (a+b)/2$, l'erreur est donnée par

$$E(f) = \frac{(b-a)^3}{24} f''(\eta), \quad \eta \in [a,b].$$

Il s'agit de la méthode du **point milieu** qui est d'ordre 1.

Ainsi, le choix du point milieu améliore l'ordre de la méthode : celle du rectangle est exacte (c'est-à-dire $E(f) = 0$) pour les fonctions constantes alors que celle du point milieu est exacte pour les polynômes de degré 1. Ceci s'explique par le fait que l'écart d'intégration de la méthode du point milieu donne lieu à deux erreurs d'évaluation, de valeurs absolues environ égales et de signes opposés.



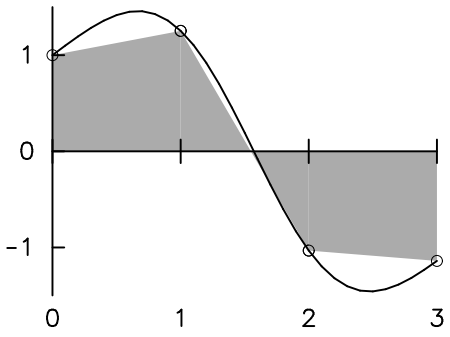
Formule du trapèze En interpolant f par un polynôme de degré 1, les deux points d'interpolation $(a,\ f(a))$ et $(b,\ f(b))$ suffisent à tracer un segment dont l'intégrale correspond à l'aire d'un trapèze, justifiant le nom de méthode des trapèzes qui est d'ordre 1 :

$$I(f) = (b-a)\,\frac{f(a)+f(b)}{2}$$

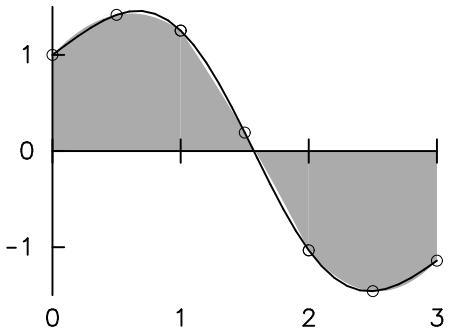
L'erreur est donnée par

$$E(f) = -\frac{(b-a)^3}{12} f''(\eta), \quad \eta \in [a,b]$$

Conformément aux expressions de l'erreur, la méthode des trapèzes est souvent moins performante que celle du point milieu.



Formule de Simpson En interpolant f par un polynôme de degré 2 (3 degrés de liberté), 3 points (ou conditions) sont nécessaires pour le caractériser : les valeurs aux extrémités a, b, et celle choisie en leur milieu $x_{1/2} = (a+b)/2$. La méthode de Simpson est basée sur un polynôme de degré 2 (intégrale d'une parabole), tout en restant exacte pour des polynômes de degré 3 ; elle est donc d'ordre 3 :



$$I(f) = \frac{(b-a)}{6}[f(a) + 4f(x_{1/2}) + f(b)]$$

L'erreur globale est donnée par

$$E(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\eta), \quad \eta \in [a,b]$$

L'analyse de l'erreur de troncature est obtenue en utilisant la formule de Lagrange que l'on rappelle ici

> **Formule de Taylor–Lagrange :** *Si la fonction f est à valeurs réelles et qu'elle est dérivable sur jusqu'à l'ordre $n+1$, alors il existe un nombre ξ strictement compris entre a et x tel que :*
>
> $$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1}.$$

Remarque : Comme la méthode du point milieu qui caractérise un polynôme de degré 0 et qui reste exacte pour tout polynôme de degré 1, la méthode de Simpson caractérise un polynôme de degré 2 et reste exacte pour tout polynôme de degré 3. Il s'agit d'une sorte d'anomalie où se produisent des compensations bénéfiques à l'ordre de la méthode.

Généralisation En généralisant cette idée (passer un polynôme de degrés $s-1$ par les s points équidistants), on obtient les *formules de Newton–Cotes (1676 et 1711)*. Pour $s \leq 7$ les coefficients de ces formules sont donnés dans le tableau 3.1. On peut montrer que les poids "explosent" au-delà de $s = 10$. Si on veut augmenter la précision, il vaut mieux raffiner les subdivisions $\{x_j\}$ qu'augmenter le degrés s.

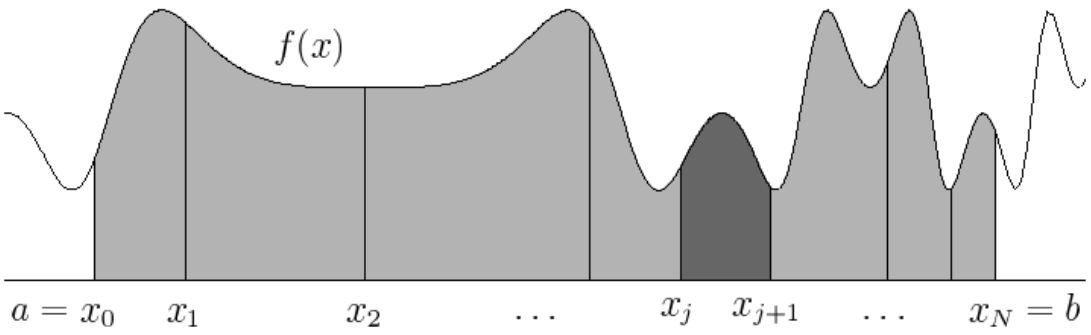
s	ordre	poids b_i							nom
2	2	$\frac{1}{2}$	$\frac{1}{2}$						trapèze
3	4	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$					Simpson
4	4	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$				Newton
5	6	$\frac{7}{90}$	$\frac{32}{90}$	$\frac{12}{90}$	$\frac{32}{90}$	$\frac{7}{90}$			Boole
6	6	$\frac{19}{288}$	$\frac{75}{288}$	$\frac{50}{288}$	$\frac{50}{288}$	$\frac{75}{288}$	$\frac{19}{288}$		
7	8	$\frac{41}{840}$	$\frac{216}{840}$	$\frac{27}{840}$	$\frac{272}{840}$	$\frac{27}{840}$	$\frac{216}{840}$	$\frac{41}{840}$	Weddle

TABLE 14.1 – Formules de quadrature de Newton–Cotes. Les nœuds équidistants sont données par la relation $c_i = \frac{i-1}{s-1}$

14.2.2 Formules composites

La plupart des algorithmes numériques procèdent comme suit : on subdivise $[a, b,]$ en plusieurs sous–intervalles $(a = x_0 < x_1 < x_2 < \cdots < x_N = b)$ et on utilise l'additivité de l'intégrale

$$\int_a^b f(x)dx = \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} f(x)dx$$



De cette manière, on est ramené au calcul de plusieurs intégrales pour lesquelles la longueur de l'intervalle d'intégration est relativement petite. Prenons une de ces intégrales et notons la longueur de l'intervalle par $h_j = x_{j+1} - x_j$. un changement de variable donne alors

$$\int_{x_j}^{x_{j+1}} f(x)dx = h_j \int_0^1 f(x_j + th_j)dt$$

Notons enfin $g(t) = f(x_j + th_j)$. Il reste alors à calculer une bonne approximation de $\int_0^1 g(t)dt$.

Si l'intervalle $[a, b]$ est simplement décomposé en n sous-intervalles de longueurs égales sur lesquels la même formule simple est appliquée, alors on obtient les formules composites suivantes :

— Méthode du point milieu d'ordre 1 :

$$I(f) = \frac{(b-a)}{n} \sum_{k=0}^{n-1} f(a + (k + 1/2)h)$$

— Méthode des trapèzes d'ordre 1 :

$$I(f) = \frac{(b-a)}{n} \left(\frac{f(a) + f(b)}{2} + \sum_{k=1}^{n-1} f(a + kh) \right)$$

— Méthode de Simpson d'ordre 3 :

$$I(f) = \frac{h}{6} \left(f(a) + f(b) + 2 \sum_{k=1}^{n-1} f(x_k) + 4 \sum_{k=0}^{n-1} f(m_k) \right)$$

avec k l'indice des n sous-intervalles, $h = (b-a)/n$ la longueur de chacun d'eux, $x_k = a + kh$ la borne inférieure, ceci pour k entre 0 et $n-1$.

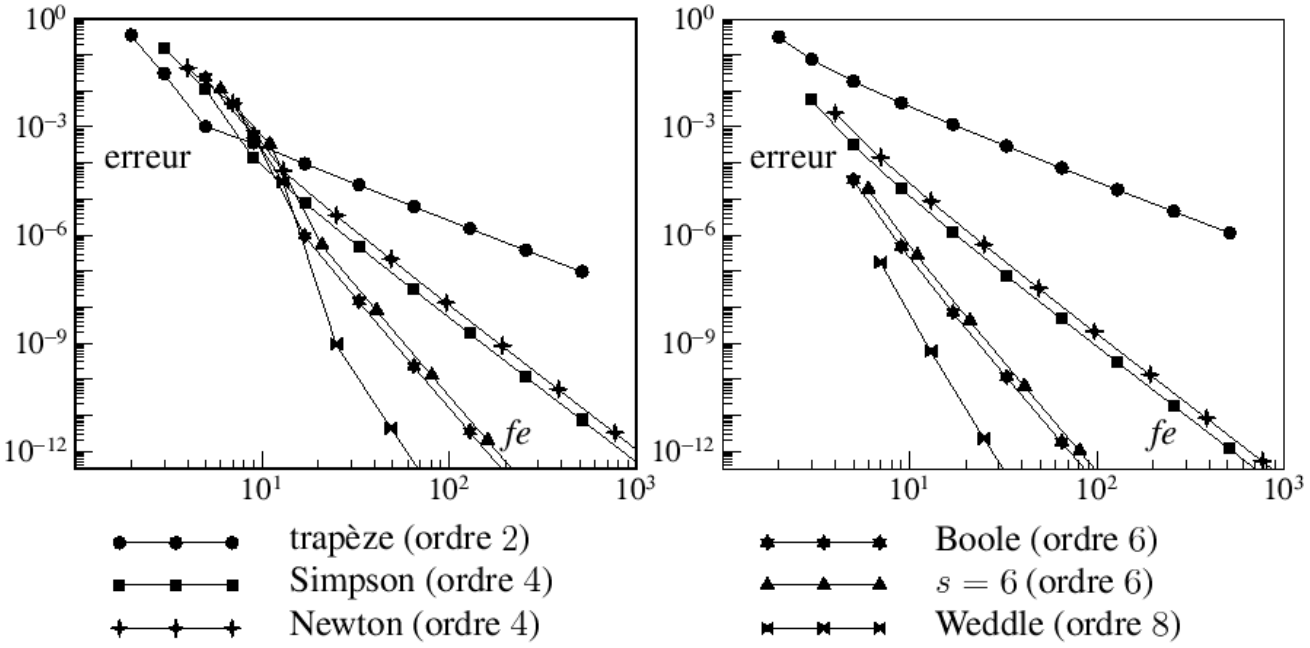
14.2.3 Étude de l'erreur

Afin étudier l'erreur commise en approchant l'intégrale par une formule de quadrature, commençons par une expérience numérique : Prenons une fonction $f(x)$, définie sur [a,b], divisons l'intervalle en plusieurs sous–intervalles équidistants ($h = (b-1)/N$) et appliquons une formule de quadrature du paragraphe précédent. Ensuite, étudions l'erreur (en échelle logarithmique)

$$err = \int_a^b f(x)dx - \sum_{j=0}^{N-1} h \sum_{i=1}^{s} b_i f(x_j + c_i h)$$

en fonction du nombre d'évaluations de la fonction $f(x)$: $fe = N(s-1)+1$. Le nombre fe représente une mesure du travail qui doit être proportionnel au temps de calcul sur un ordinateur. La figure suivante présente les résultats pour $N = 1, 2, 4, 8, 16, 32, \ldots$ obtenus par les formules de Newton-Cotes pour les deux intégrales :

$$\int_0^3 \cos(x)\exp(\sin(x))\,dx \qquad \text{et} \qquad \int_0^2 \cos(x)\,dx.$$



En étudiants les résultats on constate que
— le nombre de chiffres exacts, donné par $-\log(err)$, dépend linéairement de $\log(fe)$;
— la pente de chaque droite est $-p$ (où p est l'ordre de la formule) ;
— pour un travail équivalent, les formules avec un ordre élevé ont une meilleure précision.

Explication des résultats précédents Étudions d'abord l'erreur faite sur un sous intervalle de longueur h

$$E(f,x_0,h) = \int_{x_0}^{x_0+h} f(x)dx - h\sum_{i=1}^{s} b_i f(x_0+c_i h)$$

$$E(f,x_0,h) = h\left(\int_0^1 f(x_0+th)dx - \sum_{i=1}^{s} b_i f(x_0+c_i h)\right)$$

On peut évaluer l'erreur si f est suffisamment différentiable, on peut alors remplacer $f(x_0+th)$ et $f(x_0+c_ih)$ par les séries de Taylor développées autour de x_0, on obtient ainsi

$$E(f,x_0,h) = \sum_{q\geq 0} \frac{h^{q+1}}{q!}\left(\int_0^1 t^q dt - \sum_{i=1}^{s} b_i c_i^q\right) f^{(q)}(x_0)$$

Si la méthode est d'ordre p, les p premiers termes de la somme s'annulent est l'on a

$$E(f,x_0,h) = \frac{h^{p+1}}{p!}\left(\int_0^1 t^p dt - \sum_{i=1}^{s} b_i c_i^p\right) f^{(p)}(x_0) + \Theta(h^{p+2})$$

La constante $C = \frac{1}{p!}\left(\int_0^1 t^p dt - \sum_{i=1}^{s} b_i c_i^p\right)$ s'appelle la constante d'erreur. Supposons que h soit petit de manière à ce que le terme $\Theta(h^{p+2})$ soit négligeable par rapport au terme $h^{p+1}Cf^{(p)}(x_0)$, alors on obtient

$$err = \sum_{j=0}^{N-1} E(f,x_j,h) \approx Ch^p \sum_{j=0}^{N-1} h\, f^{(p)}(x_j) \approx Ch^p \int_a^b f^{(p)}(x)dx = Ch^p\left(f^{(p-1)}(b) - f^{(p-1)}(a)\right).$$

Cette formule permet de mieux comprendre les résultats numériques. Comme $err \approx C_1 h^p$ et $fe \approx C_2/h$, on a

$$log(err) \approx \log(C_1) + p\log(h) \approx Const - p\log(fe).$$

Ceci montre la dépendance linéaire entre $log(err)$ et $log(fe)$ ainsi que la valeur de la pente.

14.3 Méthode de Romberg

La méthode d'intégration de Romberg est une méthode récursive de calcul numérique d'intégrale, basée sur l'application du procédé d'extrapolation de Richardson à la méthode des trapèzes. Cette technique d'accélération permet d'améliorer l'ordre de convergence de la méthode des trapèzes, en appliquant cette dernière à des divisions dyadiques successives de l'intervalle d'étude et en en formant une combinaison judicieuse.

Principe On souhaite calculer l'intégrale $I = \int_a^b f(x)\,dx$ d'une fonction f supposée régulière sur $[a, b]$ en subdivisant $[a, b]$ en n sous-intervalles identiques (n pair, $n = 2p$ par exemple), du type $[a + kh, a + (k + 1)h]$ pour $k = 0, 1, \ldots, n - 1$ et $h = (b - a)/n$. La méthode des trapèzes, notée $T(n)$, est définie à l'aide de cette grille régulière :

$$T(n) = \frac{h}{2}\left[\sum_{k=0}^{n} \omega_k f(a + kh)\right]$$

où les pondérations ω_k sont égales à 1 pour les points extrêmes, et à 2 pour les autres. Puisque la méthode est exacte pour un polynôme de degré 1 (méthode d'ordre 1), l'erreur commise, notée $E = I - T(n)$, vérifie

$$E(h) = c_1 h^2 + c_2 h^4 + c_3 h^6 \cdots .$$

Cette relation exprime le fait que la méthode des trapèzes présente une erreur proportionnelle à h^2, ce qui permet d'utiliser l'extrapolation de Richardson.

Algorithme On formalise la technique précédente de réduction de l'erreur en deux étapes :

1. Initialisations : Soit $R(0, 0) = \frac{1}{2}(b-a)(f(a)+f(b))$, et $R(n, 0) = T(2^n)$ le résultat de la méthode des trapèzes basée sur $h_n = \frac{b-a}{2^n}$.
2. Récurrence : On forme le tableau suivant

$$R(n, m) = \frac{4^m R(n, m - 1) - R(n - 1, m - 1)}{4^m - 1}$$

On montre par récurrence que l'approximation à l'étape n, soit $R(n, n)$, fournit une approximation de I qui est en $O(h_n^{2n+2})$, ceci à condition que la fonction soit de classe $\mathcal{C}^{2n+2}$. Le calcul de $R(n, n)$ est exact pour les polynômes de degré $2n + 1$: elle d'ordre $2n + 1$.

L'algorithme peut être représenté sous la forme d'un tableau. La première diagonale $R(n, n)$ fournit les approximations successives ; la première colonne $R(n, 0)$ correspond (par définition) à la méthode du trapèze, la seconde $R(1, n)$ reflète la méthode de Simpson, la troisième celle de Boole qui est la formule de Newton-Cotes à 5 points appliquée à une décomposition en $2n - 2 + 1$ sous-intervalles. Par contre, les colonnes suivantes correspondent à des formules de quadrature différentes de celles de Newton-Cotes d'ordre supérieur, évitant ainsi les problèmes d'instabilité.

Pour obtenir une approximation de I avec une précision $\epsilon > 0$ donnée, le critère d'arrêt est satisfait lorsque $|R(n, n) - R(n - 1, n - 1)| < \epsilon$. Ceci implique généralement $|R(n, n) - I| < \epsilon$.

Exemple : Soit à calculer $I = \int_{-1}^{2} \frac{2}{1+4x^2}\, dx$ par ma méthode de Romberg, on obtient le tableau suivant

2^n									
1	0.7764706								
2	1.8882353	1.9623529							
4	2.3510142	2.3818661	2.3885251						
8	2.4235526	2.4283885	2.4291270	2.4292862					
16	2.4307736	2.4312550	2.4313005	2.4313090	2.4313110				
32	2.4324170	2.4325265	2.4325467	2.4325516	2.4325528	2.4325531			
64	2.4328289	2.4328564	2.4328616	2.4328629	2.4328632	2.4328633	2.4328633		
128	2.4329320	2.4329389	2.4329402	2.4329405	2.4329406	2.4329406	2.4329406	2.4329406	
256	2.4329578	2.4329595	2.4329598	2.4329599	2.4329599	2.4329599	2.4329599	2.4329599	2.4329599

En utilisant le résultat exact de l'intégrale, $I = \arctan(4) + \arctan(2)$ on détermine le nombre de chiffres significatifs corrects

2^n									
1	0.17								
2	0.65	0.71							
4	1.47	1.68	1.74						
8	2.41	2.73	2.80	2.82					
16	3.05	3.15	3.16	3.17	3.17				
32	3.65	3.74	3.76	3.77	3.77	3.77			
64	4.25	4.34	4.37	4.37	4.37	4.37	4.37		
128	4.85	4.95	4.97	4.97	4.97	4.97	4.97	4.97	
256	5.45	5.55	5.57	5.58	5.58	5.58	5.58	5.58	5.58

14.4 Formules d'un ordre supérieur

Jusqu'à présent, on a utilisé une subdivision régulière des sous-intervalles pour obtenir les formules de Newton–Cotes. Cela a conduit à des formules à s étages au mieux d'ordre $s+1$. Si la répartition équidistante des noeuds c_i a le mérite de la simplicité, il n'y pas de raison qu'elle soit la plus efficace. La question est donc : *Y a–t–il un choix de c_i, permettant d'avoir un ordre supérieur ?*

Théorème : *Soit $(b_i, c_i)_{i=1}^{s}$ une formule de quadrature d'ordre $p \geq s$ et soit*

$$M(t) = (t - c_1) \cdot (t - c_2) \cdot \cdots \cdot (t - c_s).$$

Alors, l'ordre est supérieur ou égale à $s + m$ si et seulement si $\int_0^1 M(t)g(t)dt = 0$ pour tout polynôme $g(t)$ de degrés $\leq m - 1$.

Ce théorème que l'on ne démontre pas ici permet en pratique de déterminer les noeuds c_i, on verra ensuite comment en déduire les poids b_i.

Exemple : Essayons de construire une formule de quadrature à $s = 3$ étages pour que l'ordre soit $p = 6$. Dans la terminologie du théorème précédant cela implique $m = 3$ donc le polynôme générique $g(t)$ est au maximum de degrés 2. Posons donc $g(t) = \alpha t^2 + \beta t + \gamma$. En remplaçant, on obtient une méthode d'ordre 6 si et seulement si $\int_0^1 M(t)g(t)dt = 0$ soit

$$\alpha \int_0^1 M(t)t^2 dt + \beta \int_0^1 M(t)t dt + \gamma \int_0^1 M(t)dt = 0$$

Cette égalité devant être vraie pour des polynômes de degrés 1 ($\alpha = 0$ et $\beta = 0$) et de degrés 2 ($\alpha = 0$), on en déduit

$$\int_0^1 M(t)dt = 0; \quad \int_0^1 M(t)t dt = 0; \quad \int_0^1 M(t)t^2 dt = 0$$

En développant

$$M(t) = (t - c_1)(t - c_2)(t - c_3) = t^3 - \sigma_1 t^2 + \sigma_2 t - \sigma_3$$

où $\sigma_1 = c_1 + c_2 + c_3$, $\sigma_2 = c_1c_2 + c_2c_3 + c_1c_3$ et $\sigma_3 = c_1c_2c_3$, on obtient les trois égalités

$$\begin{aligned} \frac{1}{3}\sigma_1 - \frac{1}{2}\sigma_1 + \sigma_3 &= \frac{1}{4} && (14.1)\\ \frac{1}{4}\sigma_1 - \frac{1}{3}\sigma_1 + \frac{1}{2}\sigma_3 &= \frac{1}{5} && (14.2)\\ \frac{1}{5}\sigma_1 - \frac{1}{4}\sigma_1 + \frac{1}{3}\sigma_3 &= \frac{1}{6} && (14.3) \end{aligned}$$

La résolution de ce système donne $\sigma_1 = 3/2$, $\sigma_2 = 3/5$ et $\sigma_3 = 1/20$. Il reste à déterminer les noeuds, c'est-à-dire déterminer les racines de $M(t)$

$$M(t) = t^3 - \sigma_1 t^2 + \sigma_2 t - \sigma_3 = (t - \frac{1}{2})(t - \frac{5-\sqrt{15}}{10})(t - \frac{5+\sqrt{15}}{10})$$

Par chance, le polynôme $M(t)$ ne possède que des racines réelles. Elles sont toutes dans l'intervalle [0,1], ce qui convient.

Il reste donc à déterminer les poids b_i connaissant les noeuds c_i, pour cela on utilise la relation

$$\sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q} \quad \text{pour} \quad q = 1, 2, \ldots, p$$

dont les trois premières égalités ($q = 1, 2, 3$) donne

$$\begin{aligned} b_1 + b_2 + b_3 &= 1 && (14.4)\\ c_1b_1 + c_2b_2 + c_3b_3 &= 1/2 && (14.5)\\ c_1^2b_1 + c_2^2b_2 + c_3^2b_3 &= 1/3 && (14.6)\\ & && (14.7) \end{aligned}$$

d'où on tire les poids $b_1 = 8/18$ et $b_2 = b_3 = 5/18$. On peut vérifier à l'aide des trois égalités suivantes que la méthode est bien d'ordre 6. On a donc trouvé une formule de quadrature d'ordre $p = 6$ avec seulement $s = 3$ étages :

$$\int_0^1 g(t)dt \approx \frac{5}{18}\, g\left(\frac{5+\sqrt{15}}{10}\right) + \frac{8}{18}\, g\left(\frac{1}{2}\right) + \frac{5}{18}\, g\left(\frac{5-\sqrt{15}}{10}\right)$$

La question naturelle est peut–on continuer ainsi ou y a–t–il une limitation dans l'ordre que l'on peut obtenir. La réponse est bien sûr la seconde alternative

> **Théorème :** *Si p est l'ordre d'une formule de quadrature à s étages, alors $p \leq 2s$.*

Démonstration : Supposons, par l'absurde, que l'ordre satisfasse $p \geq 2s + 1$. Alors, l'intégrale $\int_0^1 M(t)g(t)dt = 0$ est nulle pour tout polynôme $g(t)$ de degrés $\leq s$. Ceci contredit le fait que $\int_0^1 M(t)M(t)dt = \in_0^1 (t - c_1)^2 \cdot (t - c_2)^2 \cdot \cdots \cdot (t - c_s)^2 dt > 0$.

14.5 Formules de quadrature de Gauss

Pour construire une formule de quadrature d'ordre $2s$ avec $s = 4, 5, \ldots$, on peut en principe faire le même calcul que dans l'exemple précédent. Toutefois, l'approche peut être généralisée en utilisant la base des polynômes de Legendre, ce qui fournit des formules simples pour $M(t)$.

Pour construire une formule de quadrature ayant un ordre $p = 2s$, on pose

$$M(t) = C \cdot P_s(2t - 1);$$

avec P_s le polynôme de Legendre de degré s. On obtient ainsi

$$\int_0^1 P_s(2t-1)g(2t-1)\, dt = \frac{1}{2}\int_{-1}^1 P_s(\tau)g(\tau)\, d\tau = 0 \quad \text{si} \quad \deg(g) \leq s - 1.$$

On en déduit directement le théorème suivant

Théorème : *Pour chaque entier positif s, il existe une formule de quadrature unique à s étages d'ordre $p = 2s$. Elle est donnée par :*

— $c_1, c_2, \ldots, c_s$ *les racines de* $P_s(2t-1)$;

— $b_1, b_2, \ldots, b_s$ *données par* $b_i = \frac{1-c_i^2}{s^2(P_{s-1}(c_i))^2}$.

On en vient donc directement aux résultats qui sont connus sous le nom de *formule de quadrature de Gauss* ou *Gauss–Legendre*.

$$s=1: \quad \int_0^1 g(t)\,dt \approx g\Big(\frac{1}{2}\Big) \qquad \text{(formule du point milieu)}$$

$$s=2: \quad \int_0^1 g(t)\,dt \approx \frac{1}{2}g\Big(\frac{1}{2}-\frac{\sqrt{3}}{6}\Big)+\frac{1}{2}g\Big(\frac{1}{2}+\frac{\sqrt{3}}{6}\Big)$$

$$s=3: \quad \int_0^1 g(t)\,dt \approx \frac{5}{18}g\Big(\frac{1}{2}-\frac{\sqrt{15}}{10}\Big)+\frac{8}{18}g\Big(\frac{1}{2}\Big)+\frac{5}{18}g\Big(\frac{1}{2}+\frac{\sqrt{15}}{10}\Big)$$

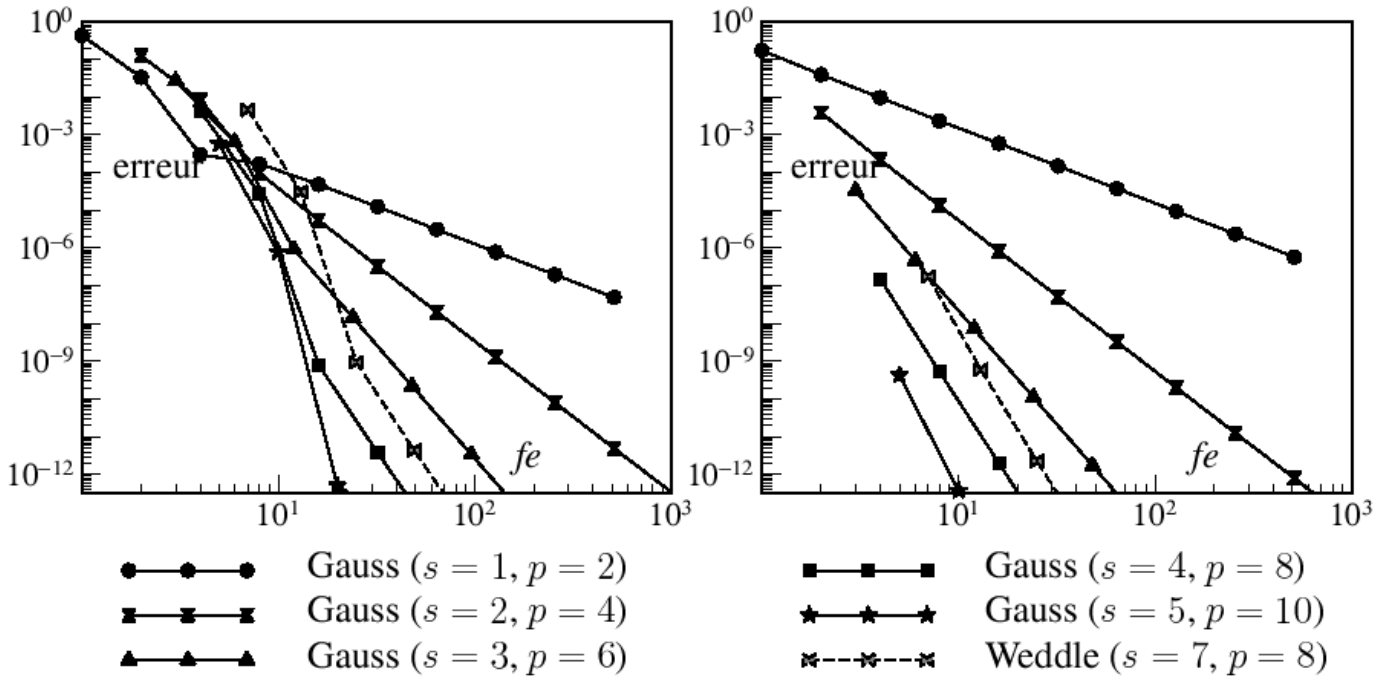
$$s=4: \quad \int_0^1 g(t)\,dt \approx \mu g\Big(\frac{1}{2}-\delta\Big)+\mu' g\Big(\frac{1}{2}-\delta'\Big)+\mu' g\Big(\frac{1}{2}+\delta'\Big)+\mu g\Big(\frac{1}{2}+\delta\Big)$$

$$s=5: \quad \int_0^1 g(t)\,dt \approx \nu g\Big(\frac{1}{2}-\epsilon\Big)+\nu' g\Big(\frac{1}{2}-\epsilon'\Big)+\frac{64}{225}g\Big(\frac{1}{2}\Big)+\nu' g\Big(\frac{1}{2}+\epsilon'\Big)+\nu g\Big(\frac{1}{2}+\epsilon\Big)$$

$$\delta=\frac{1}{2}\sqrt{\frac{15+2\sqrt{30}}{35}}\,,\quad \delta'=\frac{1}{2}\sqrt{\frac{15-2\sqrt{30}}{35}}\,,\quad \mu=\frac{1}{4}-\frac{\sqrt{30}}{72}\,,\quad \mu'=\frac{1}{4}+\frac{\sqrt{30}}{72}\,,$$

$$\epsilon=\frac{1}{2}\sqrt{\frac{35+2\sqrt{70}}{63}}\,,\quad \epsilon'=\frac{1}{2}\sqrt{\frac{35-2\sqrt{70}}{63}}\,,\quad \nu=\frac{322-13\sqrt{70}}{1800}\,,\quad \nu'=\frac{322+13\sqrt{70}}{1800}\,.$$

Expérience numérique. Avec ces formules optimales, on peut refaire les calculs numériques des deux intégrales déjà traitées par les formules de Newton-Cotes. Les résultats correspondants peuvent être admirés ; ils montrent une claire amélioration.



14.6 Un programme adaptatif

On a étudié jusqu'à présent les méthodes de base d'intégration numérique. Il reste à envisager comment les employer au mieux pour construire un algorithme généraliste d'intégration numérique qui

puisse se comporter comme une boîte noire. Ce que l'utilisateur attend d'un tel code peut être résumé comme suit : connaissant la fonction et l'intervalle d'intégration, quelle est la valeur de l'intégrale avec une précision relative de TOL. Il lui importe peu de savoir quelle sera la méthode utilisée. Si l'on fixe la formule de quadrature, il faut encore trouver une division Δ de l'intervalle $[a, b]$. Si l'on nomme par I_Δ l'approximation numérique la condition de l'utilisateur peut–être formulée comme

$$\left| I_\Delta - \int_a^b f(x)dx \right| \leq TOL \cdot \int_a^b |f(x)|dx.$$

Les deux problèmes sont

— choix de la partition Δ,

— estimation de l'erreur $I_\Delta - \int_a^b f(x)dx$

Choix de la partition Supposons le deuxième point résolu, une méthode simple pour la partition est de procéder par dichotomie : on divise en deux l'intervalle ou l'erreur est maximale tant que la condition imposée par l'utilisateur n'est pas vérifier.

Estimation de l'erreur Malheureusement les formules de majoration de l'erreur obtenues ne sont pas ici d'une grande utilité, car on ne connaît que très rarement la dérivée d'ordre p de la fonction $f(x)$. L'idée est d'appliquer une deuxième formule de quadrature d'ordre différent et d'utiliser la différence de deux approximations numériques comme estimation de l'erreur du moins bon résultat. Pour que le travail supplémentaire soit négligeable, on choisit des formules avec les mêmes coefficients c_i.

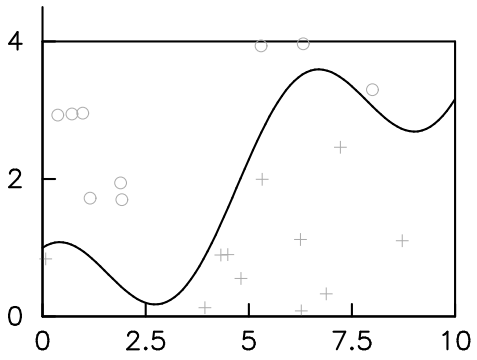
Un exemple de partition est donné dans la première figure de ce chapitre

14.7 Méthode de Monte–Carlo

Le terme méthode de Monte-Carlo, ou méthode Monte-Carlo, désigne toute méthode visant à calculer une valeur numérique en utilisant des procédés aléatoires, c'est-à-dire des techniques probabilistes. Le nom de ces méthodes, qui fait allusion aux jeux de hasard pratiqués à Monte-Carlo, a été inventé en 1947 par Nicholas Metropolis, et publié pour la première fois en 1949 dans un article co-écrit avec Stanislas Ulam.

Les méthodes de Monte-Carlo sont particulièrement utilisées pour calculer des intégrales en dimensions plus grandes que 1.

Principe en dimension 1 Supposons le calcul de l'intégrale de la fonction f représenter ci-contre entre 0 et 10. Si on tire au hasard un grand nombre de points dans la boîte $x \in [0, 10]$, $y \in [0, 4]$ alors le nombre de points sous la courbe (les croix ci-contre) est proportionnel à l'aire sous la courbe. Si on tire N points dont n sous la courbe dans une boîte d'aire A on obtient une approximation de l'intégrale $I \simeq \frac{n}{N} \times A$.



Généralisation La méthode de Monte–Carlo converge lentement, mais elle est pertinente en grande dimension lorsque le domaine d'intégration est difficile à déterminer.

Si on souhaite calculer

$$I = \int_A f(\overrightarrow{x})\mathrm{d}\overrightarrow{x},$$

sur un domaine A compliqué avec $\overrightarrow{x}$ multidimensionnel, On se donne un domaine plus simple D de volume V qui englobe A. L'inégrale I est réécrite sous la forme

$$I = \int_A f(\overrightarrow{x})\mathrm{d}\overrightarrow{x} = \frac{1}{V}\int_D f(\overrightarrow{x}) \cdot g(\overrightarrow{x})V\mathrm{d}\overrightarrow{x},$$

où

$$g(\overrightarrow{x}) = \begin{cases} 1 & \text{si } \overrightarrow{x} \in A \\ 0, & \text{sinon} \end{cases}.$$

Dès lors, I correspond à la valeur moyenne de $h(\overrightarrow{x}) = f(\overrightarrow{x}) \cdot g(\overrightarrow{x})V$ sur le domaine D. Si les points $\overrightarrow{x}_i$ sont uniformément distribués dans D alors

$$I \simeq \frac{1}{N}\sum_{i=0}^{N} h(\overrightarrow{x}_i) = \frac{V}{N}\sum_{i=0}^{N} f(\overrightarrow{x}_i) \cdot g(\overrightarrow{x}_i).$$

De manière générale, l'erreur décroît en $\frac{1}{\sqrt{N}}$: pour avoir une erreur 2 fois plus faible, il faut 4 fois plus de points !

14.8 Exercices d'applications

14.8.1 Applications directes

1 - Calculer par les méthodes de Newton l'intégrale $\int_0^1 \frac{1}{1+x^2}dx$ et évaluer l'erreur commise.

2 - Pour l'intégrale $\int_{-4}^{4} e^{-x^2}$ comparer les méthodes d'intégration des trapèzes, de Simpson et de Gauss.

3 - Comment peut–on, à partir de ce qui a été vu en cours, calculer l'intégrale $\int_1^\infty \frac{1}{x^2}dx$?

14.8.2 Intégrale généralisée

Déterminer une approximation numérique l'intégrale

$$\int_0^1 \frac{1}{\sqrt{x}}dx.$$

1 - Peut–on utiliser toutes les formules de Newton–cotes ?

2 - Construire la suite S_N des approximations successives en utilisant la formule du points milieu et la partition de $[0, 1]$ en 2^N sous–intervalles équidistants.

3 - Comparer avec l'approximation utilisant une formule de Gauss avec $s = 3$, $s = 5$ ainsi qu'avec la valeur exacte. Les erreurs sont–elles conforment à ce que l'on pouvait estimer ?

14.8.3 Intégrale double

Construire une méthode pour évaluer l'intégrale double $\int_0^2 \int_1^4 xy^2 dxdy$ et évaluer l'erreur commise.

14.8.4 Évaluation Novembre 2010

On souhaite calculer une approximation de l'intégrale suivante

$$I = \int_0^a f(x)dx.$$

Méthode des trapèzes

On veut déterminer une approximation de I en subdivisant l'intervalle $[0, 2]$ en N sous intervalles égaux et en appliquant dans chaque sous intervalle la méthode du trapèze.

1 - Monter qu'en posant $h = \frac{a}{N}$, l'écriture formelle de la méthode donne

$$I \approx h\left(\frac{f(0) + f(a)}{2} + \sum_{k=1}^{N-1} f(kh)\right)$$

2 - Présenter en tableau une approximation de I pour $N = 1, 2, 4, 8$, si $f(x) = xe^{-x}$ et $a = 2$.

Algorithme de Romberg

On note $T_0(h)$ l'approximation obtenue précédemment $T_0(h) = h\left(\frac{f(0)+f(a)}{2} + \sum_{k=1}^{N-1} f(kh)\right)$. La méthode du trapèze étant une méthode d'ordre 2, l'erreur commise $E(h) = T_0(h) - I$ est de la forme

$$E(h) = c_1 h^2 + c_2 h^4 + c_3 h^6 + \dots$$

où les coefficients c_i dépendent ni de h, ni de N. On a donc

$$T_0(h) = I + c_1 h^2 + c_2 h^4 \dots .$$

3 - Montrer que $T_1(h) = \frac{1}{3}\left(4T_0(\frac{h}{2}) - T_0(h)\right)$ est une approximation de I. Quelle est l'ordre du premier terme de l'erreur ? $T_1(h)$ est-elle une meilleure approximation de I que $T_0(h)$?

On peut itérer ce principe en posant

$$T_n\left(\frac{h}{2^k}\right) = \frac{4^n T_{n-1}\left(\frac{h}{2^{k+1}}\right) - T_{n-1}\left(\frac{h}{2^k}\right)}{4^n - 1}$$

ce qui permet d'établir de meilleurs approximations de I.

4 - Dans la première partie, vous avez calculé $T_0(a)$, $T_0(\frac{a}{2})$, $T_0(\frac{a}{4})$, $T_0(\frac{a}{8})$ pour $f(x) = xe^{-x}$ et $a = 2$. Utiliser l'algorithme précédent pour calculer $T_1(a)$, $T_1(\frac{a}{2})$, $T_1(\frac{a}{4})$, $T_2(a)$, $T_2(\frac{a}{2})$, $T_3(a)$

Chapitre 15

Intégration numérique des équations différentielles

Sommaire

La résolution d'équations différentielles ordinaires est un problème très courant en physique, chimie, écologie des populations .... Dans la grande majorité des cas, il faut employer des méthodes numériques, car la résolution analytique est impossible. Posons le problème

Problème : *Trouver une approximation de* $y(T)$ *avec* $T > t_0$ *si*

$$\frac{dy}{dt} = f(t, y), \quad \text{et} \quad y(t_0) = y_0$$

Ce chapitre présente exclusivement les méthodes de **Runge–Kutta** qui sont les plus employées, d'autres méthodes existent et ont leurs avantages propres, mais la curiosité de les découvrir est laissée au lecteur. Les autres méthodes les plus connues sont les méthodes d'Adams et leurs variantes, les méthodes de Prédiction-Correction, les méthodes de Lobatto, celle utilisant l'extrapolation de Richardson ....

15.1 Méthodes de Runge–Kutta

L'idée générale des méthodes de Runge–Kutta est de diviser l'intervalle $[t_0, T]$ en sous-intervalles d'extrémités $t_0 < t_1 < \cdots < t_N = T$, on dénote $h_n = t_{n+1} - t_n$ et on calcule l'approximation $y_n = y(t_n)$ par une formule du type

$$y_{n+1} = y_n + h_n \Phi(t_n, y_n, h_n).$$

Une telle formule s'appelle "méthode à un pas", car le calcul de y_{n+1} utilise uniquement les valeurs h_n, t_n, y_n et non $h_{n-1}, t_{n-1}, y_{n-1} \ldots$.

Méthode d'Euler (1768).

La méthode la plus simple est

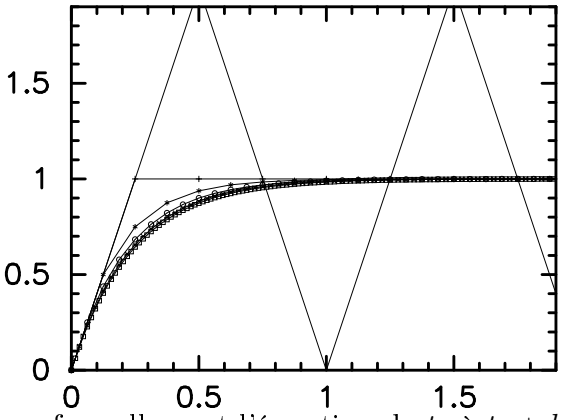
$$y_{n+1} = y_n + h_n f(t_n, y_n).$$

Elle est obtenue en remplaçant la solution $y(t)$ par sa tangente au point (t_n, y_n).

Prenons un exemple simple, recherchons une approximation de $y(t > 0)$ solution de

$$\frac{dy}{dt} = -4 \cdot (y - 1), \quad \text{avec} \quad y(0) = 0$$

en utilisant la méthode d'Euler avec $h_n = h = \frac{1}{2^n}$. On peut observer la convergence de la méthode sur la figure de droite.



Pour la dérivation d'autres méthodes numériques, intégrons formellement l'équation de t_0 à $t_0 + h$

$$y(t_0 + h) = y_0 + \int_{t_0}^{t_0+h} f\left(y(\tau, y(\tau)\right) d\tau.$$

La résolution de l'équation différentielle se ramène donc à un problème similaire à une quadrature. Si l'on remplace l'intégrale par $hf(t_0, y_0)$, on obtient la méthode d'Euler. L'idée évidente est d'approcher cette intégrale par une formule de quadrature ayant un ordre plus élevé.

Méthode de Runge (1895).

On prend la formule du point milieu

$$y(t_0 + h) \approx y_0 + hf\left(t_0 + \frac{h}{2}, y(t_0 + \frac{h}{2})\right)$$

et on remplace la valeur inconnue $y(t_0 + \frac{h}{2})$ par une estimation *via* la méthode d'Euler. Ceci donne

$$y_1 = y_0 + hf\left(t_0 + \frac{h}{2}, y_0 + \frac{h}{2} f(t_0, y_0)\right)$$

Méthode de Heun (1900).

On prend une formule de quadrature d'ordre 3

$$y(t_0 + h) \approx y_0 + \frac{h}{4}\left[f\left(t_0, y(t_0)\right) + 3f\left(t_0 + \frac{2h}{3}, y(t_0 + \frac{2h}{3})\right)\right]$$

et on remplace la valeur inconnue $y(t_0 + 2h/3)$ par l'approximation de la méthode de Runge. Ceci donne

$$y_1 = y_0 + \frac{h}{4}\left[f(t_0, y_0) + 3f\left(t_0 + \frac{2h}{3}, y_0 + \frac{2h}{3} f\left(t_0 + \frac{h}{3}, y_0 + \frac{h}{3} f(t_0, y_0)\right)\right)\right]$$

En généralisant cette idée à une formule de quadrature plus générale et en introduisant la notation $k_i = f(\dots)$ pour les expressions $f(t, y)$ qui apparaissent, on est conduit à la définition suivante

Définition : *Une méthode de Runge–Kutta à s étages est donnée par*

$$\begin{aligned}
k_1 &= f(t_0, y_0) \\
k_2 &= f(t_0 + c_2 h, y_0 + h a_{21} k_1) \\
k_3 &= f\left(t_0 + c_3 h, y_0 + h(a_{31} k_1 + a_{32} k_2)\right) \\
&\vdots \\
k_s &= f\left(t_0 + c_s h, y_0 + h(a_{s1} k_1 + \cdots + a_{s,s-1} k_{s-1})\right) \\
y_1 &= y_0 + h(b_1 k_1 + \cdots + b_s k_s)
\end{aligned}$$

où c_j, a_{ij}, b_j sont des coefficients. On la représente à l'aide du schéma $\begin{array}{c|c} c_i & a_{i,j} \\ \hline & b_i \end{array}$

Exercice : Donner les coefficients des méthodes d'Euler de Runge et de Heun.

On peut maintenant étendre la notion de l'ordre défini pour les formules de quadrature aux méthodes de Runge–Kutta.

Définition : *On dit que la méthode de Runge–Kutta a l'ordre p si, pour chaque problème $y' = f(t,y)$, $y(t_0 = y_0$ (avec f suffisamment différentiable), l'erreur après un pas satisfait*

$$y_1 - y(t_0 + h) = \Theta(h^{p+1}) \quad pour \quad h \to 0.$$

La différence $y_1 - y(t_0 + h)$ s'appelle l'erreur locale de la méthode.

Exercice : Montrer que la méthode d'Euler et d'ordre $p = 1$ et la méthode de Runge d'ordre $p = 2$.

Les méthodes d'ordre 4 les plus classique sont les suivantes

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	2/6	2/6	1/6

"La" Méthode de Runge–Kutta

0				
1/3	1/3			
1/3	-1/3	1		
1	1	-1	1	
	1/8	3/8	3/8	1/8

La méthode dite du 3/8

Celle de gauche est basée sur la formule de Simpson, l'autre sur la formule de Newton.

Séance de TP

15.2 Convergence des méthodes de Runge–Kutta

En TP, on a pu (on pourra) constaté que pour un calcul avec des pas constants, l'erreur globale se comporte comme $\log(err) \approx C_0 - p \cdot \log(fe)$, ce qui est équivalent à $err \approx C_1(fe)^{-p} \approx C_2 h^p$. Ceci montre que la solution numérique converge vers la solution exacte si $h \to 0$. Dans ce paragraphe, on va démontrer ce résultat.

On applique une méthode à un pas

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

à une équation différentielle $y' = f(t,y)$, $y(t_0 = y_0$, et on cherche à estimer l'*erreur globale* $y(t_n) - y_n$.

Théorème : *Soit $y(t)$ la solution de $y' = f(t,y)$, $y(t_0 = y_0$ sur $[t_0, T]$.*
Supposons que

1. *l'erreur locale satisfasse pour $t \in [t_0, T]$ et $h < h_max$*

$$||y(t+h) - y(t) - h\Phi(t, y(t), h)|| \leq C \cdot h^{p+1}$$

2. *la fonction $\Phi(t,y,h)$ satisfasse une condition de Lipshitz*

$$||\Phi(t,y,h) - \Phi(t,z,h)|| \leq \Lambda \cdot ||y - z||$$

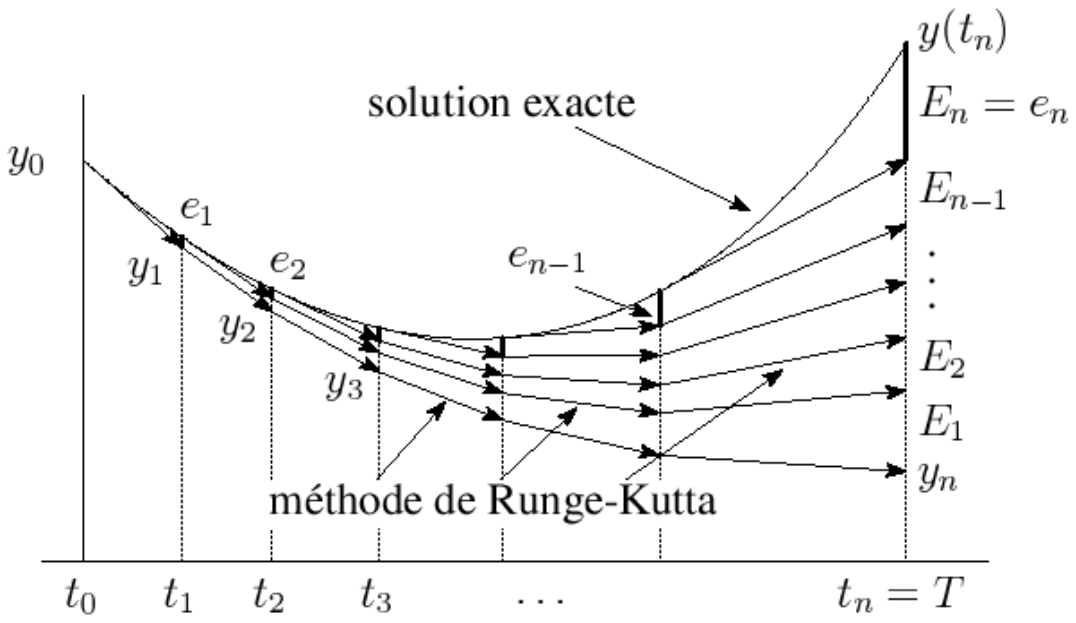
pour $h \leq h_max$ et (t,y), (t,z) dans un voisinage de la solution.

Alors, l'erreur globale admet pour $t_n \leq T$ l'estimation

$$||y(t_n) - y_n|| \leq h^p \cdot \frac{C}{\lambda} \cdot \left(e^{\lambda(t_n - t_0)} - 1\right)$$

où $h = \max_i h_i$, sous la condition de h soit suffisamment petit.

L'idée de la démonstration est d'étudier l'influence de l'erreur locale, commise au i^{me} pas, sur l'approximation y_n. Ensuite, on additionne les erreurs accumulées.



Exercice : Vérifier les conditions du théorème pour la méthode d'Euler.

On peut montrer que les hypothèses du théorème se vérifient pour une méthode d'ordre p. Ce théorème montre que pour h *suffisamment petit*, la solution numérique converge vers la solution de l'équation différentielle. On pourrait y voir la panacée, mais malheureusement il se trouve parfois (et même souvent) que le h suffisamment petit est si petit qu'il conduise à une explosion du temps de calcul. On verra qu'il existe alors d'autres méthodes.

15.3 Équations différentielles raides (stiff)

On a pu voir en TP que la résolution numérique d'une équation différentielle simple par une méthode de Runge–Kutta pouvait réclamer des pas très petits pour obtenir une approximation acceptable. Pour mieux comprendre ce phénomène, considérons le problème plus simple

$$\epsilon y' = -y + \cos(t), \quad 0 < \epsilon \ll 1$$

qui possède les mêmes caractéristiques.

Exercice : Montrer que la solution générale de l'équation est de la forme

$$y(t) = C \cdot e^{-t/\varepsilon} + cost + \varepsilon \sin t + \Theta(\varepsilon^2).$$

Méthodes d'Euler explicite et implicite

Une manière simple de retrouver la méthode d4Euler et de remplacer dans l'équation

$$\frac{dy}{dt} = f(y,t)$$

la dérivée par une différence finie

$$\frac{dy}{dt} \approx \frac{y_{n+1} - y_n}{h}.$$

On a alors deux choix pour évaluer la solution y dans le membre de droite

$$y = y_n \quad \text{ou} \quad y = y_{n+1}$$

Le premier choix conduit à la méthode d'Euler dite explicite qui a déjà été utilisée

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

Le second à la méthode dite implicite dans la mesure ou elle réclame une évaluation de f en un point indéterminé

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

La méthode d'Euler implicite n'est pas directement exploitable sauf dans le cas ou f est linéaire en y comme dans l'exemple proposé.

Appliquons les deux schémas d'Euler avec un pas constant h tels que $t_n = nh$

Solution numérique 1 (Euler explicite).

Le schéma conduit à

$$y_{n+1} = \left(1 - \frac{h}{\varepsilon}\right) y_n + \frac{h}{\varepsilon} \cos t_n.$$

C'est une équation aux différences finies qui est linéaire et inhomogène. La solution est obtenue comme pour une équation différentielle. On cherche d'abord une solution particulière de la forme $y_n = A \cos t_n + B \sin t_n$. En utilisant les relations trigonométriques et en identifiant les deux membres on obtient deux équations pour A et B dont la solution est de la forme $A = 1 + \Theta(h\varepsilon)$ et $B = \varepsilon + \Theta(h^2\varepsilon)$. En ajoutant la solution générale $y_n = \left(1 - \frac{h}{\varepsilon}\right)^n$, on obtient

$$y_n = C \cdot \left(1 - \frac{h}{\varepsilon}\right)^n + \cos t_n + \varepsilon \sin t_n + \Theta(h\varepsilon).$$

On voit que la solution numérique y_n converge vers la solution exacte si $|1 - h/\varepsilon| < 1$, c–à–d. si $h < 2\varepsilon$. Si ε très petit (par ex. 10^{-6}) une telle restriction est inacceptable. Remarquer que la restriction sur h est indépendante de la précision demandée, elle est uniquement conditionnée par la propagation des erreurs dans la méthode.

Solution numérique 2 (Euler implicite).

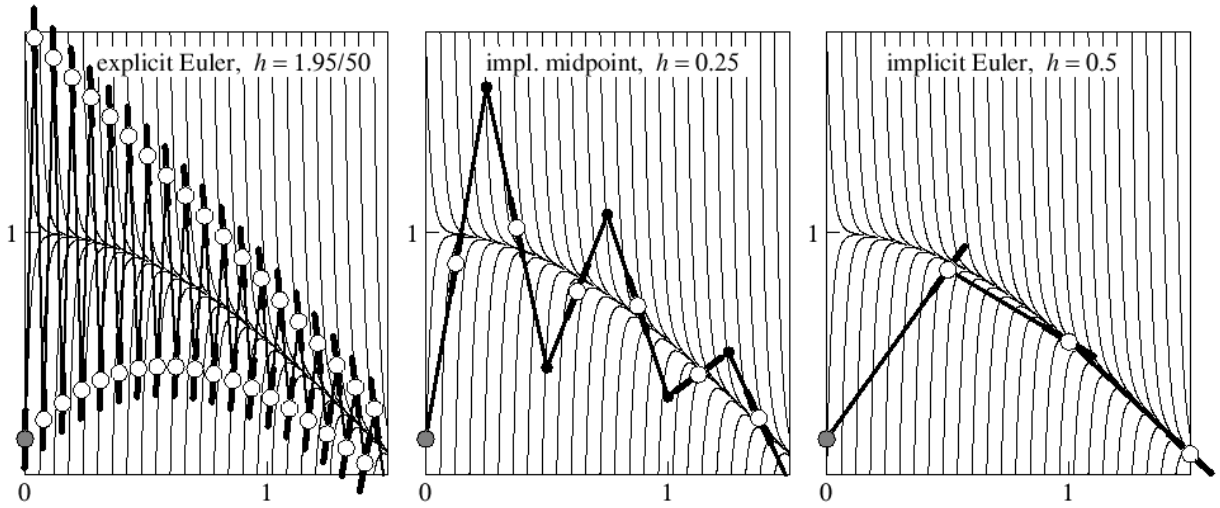
Le schéma conduit maintenant à

$$\left(1 + \frac{h}{\varepsilon}\right) y_{n+1} = y_n + \frac{h}{\varepsilon} \cos t_{n+1}.$$

dont la solution peut être écrite sous la forme

$$y_n = C \cdot \left(1 + \frac{h}{\varepsilon}\right)^{-n} + \cos t_n + \varepsilon \sin t_n + \Theta(h\varepsilon).$$

Cette fois–ci on n'a pas de restriction sur la longueur du pas, car $|(1 + h/\varepsilon)^{-1}| < 1$ pour $h > 0$. La figure suivante illustre bien la bonne approximation même si h est grand.



Il est bien évidement possible de panacher les deux méthodes en utilisant un schéma où l'évaluation de f est effectuée en $t_n + \theta h$ avec $0 \geq \Theta \geq 1$. La méthode pour $\theta = 1/2$ est connue comme *implicite Euler midpoint*.

Le calcul précédent a montré que ce n'est pas la solution particulière qui pose des difficultés à la méthode explicite, mais c'est l'approximation de la solution de l'équation homogène $\varepsilon y' = -y$. On considère donc le problème un peu plus général

$$y' = \lambda y$$

comme équation de test. Sa solution exacte est $y(t) = C \cdots e^{\lambda t}$ et elle est bornée pour $t \geq 0$ si $Re(\lambda \leq 0)$. La solution numérique d'une méthode de Runge–Kutta appliquée avec des pas constant ne dépend que de $h\lambda$. Il est alors intéressant d'étudier pour qu'elle valeur de $z = h\lambda$ la solution numérique reste bornée.

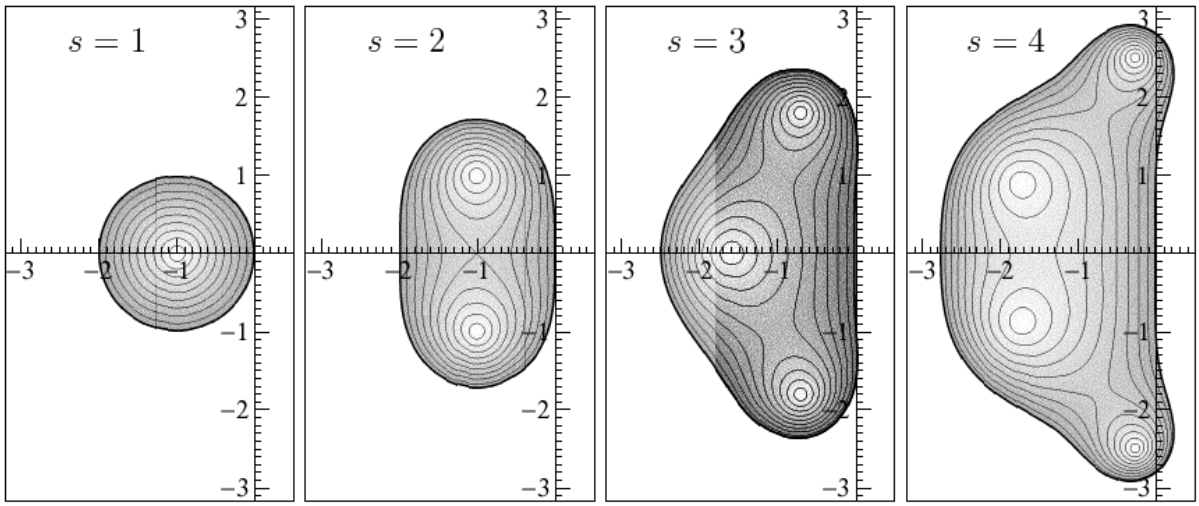
Définition (A–stabilité) : *Considérons une méthode dont la solution numérique $y_{n n\geq 0}$ pour l'équation de test $y' = \lambda y$ est une fonction de $z = h\lambda$. Alors, l'ensemble*

$$S = z \in \mathbb{C};\ y_{n n\geq 0} \quad \textit{est bornée}$$

s'appelle le domaine de stabilité de la méthode. On dit que la méthode est A-stable si le domaine de stabilité recouvre l'ensemble du demi-plan complexe défini par $Re(z) \leq 0$.

Pour la méthode d'Euler explicite le domaine de stabilité est le disque de rayon 1 est de centre -1. Pour la méthode d'Euler implicite il est l'extérieur du disque de rayon 1 est de centre $+1$. La méthode d'Euler implicite est A–stable, Elle est donc particulièrement adaptée au problème raide.

La figure suivante présente les domaines de stabilité des méthodes de Runge–Kutta explicite, Si le domaine de stabilité croit avec l'ordre, il ne recouvre jamais entièrement le demi–plan.



15.4 TP Équations différentielles

Mise en bouche Utiliser la méthode d'Euler pour déterminer la solution numérique du problème suivant

$$y' = t^2 + y^2, \quad y(-1.5) = -1.4$$

en utilisant $h = 1/n$.

Représenter graphiquement les différentes solutions ainsi que la convergence de la solution $y(1.5)$ en fonction de n.

Un programme simple Appliquer les cinq méthodes de Runge–Kutta vues jusqu'à maintenant et comparer leurs performances pour le problème de Van der Pol

$$\begin{aligned} y_1' &= y_2 & y_1(0) &= 2.00861986087484313650940188 \\ y_2' &= (1-y_1^2)y_2 - y_1 & y_2(0) &= 0. \end{aligned}$$

La valeur initiale est choisie pour que la solution soit périodique de période

$$T = 6.66328685932313018969968203 05$$

1. Vérifier la valeur de cette période
2. Subdiviser l'intervalle $[0, T]$ en n parties équidistantes et appliquer n fois la méthode.
3. Représenter l'erreur à la fin de l'intervalle en fonction du travail (nombre d'évaluation de f).
4. Déterminer l'ordre de chaque méthode.

Un programme à pas variables Pour résoudre un problème réaliste, un calcul à pas constant est en général inefficace. Mais comment choisir la division ? L'idée est de choisir les pas afin que l'erreur locale soit partout environ égale à Tol (fourni par l'utilisateur). Pour estimer l'erreur locale, on construit une deuxième méthode de Runge–Kutta avec $\hat{y}_i$ comme approximation numérique, et on utilise la différence $\hat{y}_i - y_i$ comme estimation de l'erreur locale du moins bon résultat.

Méthode emboîtée

Soit donnée une méthode d'ordre p à s étages (coefficients a_i, b_i, $c_{i,j}$). On cherche une approximation $\hat{y}_i$ d'ordre $\hat{p} < p$ qui utilise les mêmes évaluations de f, c–à–d. ,

$$\hat{y}_1 = y_0 + h\left(\hat{b}_1 k_1 + \cdots + \hat{b}_s k_s + \hat{b}_{s+1} f(x_1, y_1)\right)$$

Le dernier terme est rajouté pour plus de souplesse est demande une évaluation de f nécessaire au pas suivant.

Pour la méthode de Runge, on peut utiliser la méthode d'Euler comme méthode emboîtée. Montrer que l'expression $err = h(k_2 - k_1)$ est une approximation de l'erreur.

La première méthode emboîtée proposée est la *méthode de Merson (1957)*, dont voici le schéma

$$\begin{aligned} k_1 &= f(x_0, y_0) \\ k_2 &= f(x_0 + \frac{h}{3}, y_0 + \frac{h}{3}k_1) \\ k_3 &= f(x_0 + \frac{h}{3}, y_0 - \frac{h}{6}k_1 + \frac{h}{6}k_2) \\ k_4 &= f(x_0 + \frac{h}{2}, y_0 + \frac{h}{8}k_1 + 3\frac{h}{8}k_3) \\ k_5 &= f(x_0 + h, y_0 + \frac{h}{2}k_1 - 3\frac{h}{2}k_2 + 2hk_3) \\ y_1 &= y_0 + \frac{h}{6}(k_1 + 4k_2 + k_5) \\ \hat{y}_1 &= y_0 + \frac{h}{2}(k_1 - 3k_3 + 4k_4) \end{aligned}$$

L'erreur est donnée par $err = |y_1 - \hat{y}_1|$ est évaluer à chaque pas. C'est la méthode que l'on propose d'utiliser ici. La méthode divise le pas par 2 quand $err > tol$ et le multiplie par deux quand $err \leq tol/64$. La méthode de Merson est d'ordre 5 pour le calcul de la solution et d'ordre 4 pour le calcul de l'erreur.

1. Utiliser l'algorithme de Merson pour le problème suivant

$$\begin{aligned} y_1' &= 1 + y_1^2 y_2 - 4y_1 & y_1(0) &= 1.5 \\ y_2' &= 3y_1 - y_1^2 y_2 & y_2(0) &= 3 \end{aligned}$$

sur l'intervalle $[0, 20]$ pour une précision $tol = 10^{-4}$.

2. Représenter les deux composantes de la solution avec tous les pas acceptés
3. Représenter l'évolution de h avec les pas rejetés

Chapitre 16

Méthode numérique d'algèbre Linéaire

Sommaire

Il existe de nombreuses méthodes numériques pour l'algèbre linéaire répartie en deux grandes familles. Les méthodes de résolution d'un système linéaire et les méthodes de diagonalisation. Présenter l'ensemble de ces méthodes étant au-delà de l'objectif de ce cours, on focalise sur une famille de méthode très appréciée utilisant la décomposition matricielle dite QR par le procédé de Gram–Schmidt.

16.1 Décomposition QR

La décomposition QR d'une matrice A est une réécriture de la matrice A sous la forme d'un produit d'une matrice orthogonale Q (telle que $Q \cdot Q^t = I$) et d'une matrice triangulaire supérieure R telle que

$$A = Q \cdot R.$$

Notons d'emblée que la décomposition QR n'est pas unique, elle dépend de la base orthogonale composant la matrice Q. Elle n'est pas non plus réservée aux matrices carrées, car si Q est toujours une matrice carrée, R peut être rectangulaire.

Plusieurs méthodes permettent de réaliser une décomposition QR : citons par exemple celle de Householder, de Schmidt et de Givens. Chacune d'entre elles a ses avantages et ses inconvénients. La méthode ici présentée est celle de Schmidt moins pour son efficacité (celle de Givens est en ce sens meilleure) que pour sa similarité avec le procédé de Gram–Schimdt déjà présenté dans ce cours.

On considère le procédé de Gram-Schmidt appliqué aux colonnes de la $m \times n$–matrice A pris comme des vecteurs $A = [\overrightarrow{a}_1, \overrightarrow{a}_2, \ldots, \overrightarrow{a}_n]$, muni du produit scalaire $\langle \overrightarrow{u} \mid \overrightarrow{v} \rangle$ permettant de définir le projecteur orthogonal de $\overrightarrow{v}$ sur $\overrightarrow{u}$: $\Pi_{\overrightarrow{u}}(\overrightarrow{v})$.

À partir de la base des vecteurs $(\overrightarrow{a}_i)$ on construit *via* le procédé de Gram–Schimdt une base orthonormale $(\overrightarrow{e}_i)$. Ici on normalise au fur et à mesure et on projette sur les vecteurs normalisés, cela

simplifie l'étape suivante :

$$\begin{aligned}
\overrightarrow{f}_1 = \overrightarrow{a}_1 \quad &; \quad \overrightarrow{e}_1 = \frac{\overrightarrow{f}_1}{||\ \overrightarrow{f}_1\ ||} \\
\overrightarrow{f}_2 = \overrightarrow{a}_2 - \Pi_{\overrightarrow{e}_1}(\overrightarrow{a}_2) \quad &; \quad \overrightarrow{e}_2 = \frac{\overrightarrow{f}_2}{||\ \overrightarrow{f}_2\ ||} \\
\overrightarrow{f}_3 = \overrightarrow{a}_3 - \Pi_{\overrightarrow{e}_1}(\overrightarrow{a}_3) - \Pi_{\overrightarrow{e}_2}(\overrightarrow{a}_3) \quad &; \quad \overrightarrow{e}_3 = \frac{\overrightarrow{f}_3}{||\ \overrightarrow{f}_3\ ||} \\
\cdots & \\
\overrightarrow{f}_n = \overrightarrow{a}_n - \sum_{j=1}^{n-1} \Pi_{\overrightarrow{e}_j}(\overrightarrow{a}_j) \quad &; \quad \overrightarrow{e}_n = \frac{\overrightarrow{f}_n}{||\ \overrightarrow{f}_n\ ||}
\end{aligned}$$

L'orthonormalisation étant accomplie, on a $\langle \overrightarrow{e}_i \mid \overrightarrow{a}_i \rangle = ||\ \overrightarrow{f}_i\ ||$, on peut donc réarranger les équations de sorte que les $\overrightarrow{a}_i$ soient à gauche, en utilisant le fait que les $\overrightarrow{e}_i$ sont des vecteurs unitaires :

$$\begin{aligned}
\overrightarrow{a}_1 &= \langle \overrightarrow{e}_1, \overrightarrow{a}_1 \rangle \overrightarrow{e}_1 \\
\overrightarrow{a}_2 &= \langle \overrightarrow{e}_1, \overrightarrow{a}_2 \rangle \overrightarrow{e}_1 + \langle \overrightarrow{e}_2, \overrightarrow{a}_2 \rangle \overrightarrow{e}_2 \\
\overrightarrow{a}_3 &= \langle \overrightarrow{e}_1, \overrightarrow{a}_3 \rangle \overrightarrow{e}_1 + \langle \overrightarrow{e}_2, \overrightarrow{a}_3 \rangle \overrightarrow{e}_2 + \langle \overrightarrow{e}_3, \overrightarrow{a}_3 \rangle \overrightarrow{e}_3 \\
&\vdots \\
\overrightarrow{a}_k &= \sum_{j=1}^{k} \langle \overrightarrow{e}_j, \overrightarrow{a}_k \rangle \overrightarrow{e}_j
\end{aligned}$$

Ceci s'écrit matriciellement comme

$$A = Q \cdot R$$

avec

$$Q = [\overrightarrow{e}_1, \overrightarrow{e}_2, \ldots, \overrightarrow{e}_n] \qquad \text{et} \qquad R = \begin{pmatrix} \langle \overrightarrow{e}_1 \mid \overrightarrow{a}_1 \rangle & \langle \overrightarrow{e}_1 \mid \overrightarrow{a}_2 \rangle & \langle \overrightarrow{e}_1 \mid \overrightarrow{a}_3 \rangle & \ldots \\ 0 & \langle \overrightarrow{e}_2 \mid \overrightarrow{a}_2 \rangle & \langle \overrightarrow{e}_2 \mid \overrightarrow{a}_3 \rangle & \ldots \\ 0 & 0 & \langle \overrightarrow{e}_3 \mid \overrightarrow{a}_3 \rangle & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

Remarquons que lors de l'exécution du procédé de Gram–Schimdt, on calcul simultanément les éléments de la matrice R, le procédé fournit donc directement la décomposition QR sans coût numérique supplémentaire.

Exercice : Calculer la décomposition QR de la matrice $A = \begin{pmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{pmatrix}$.

Exercice : Montrer que pour une matrice carré de taille n, le nombre d'opérations requises par la méthode de Schmidt varie en n^3.

16.2 Résolution d'un système linéaire

Définition : *Soit A une $n \times n$–matrice à coefficient dans $\mathbb{R}$ ou $\mathbb{C}$. Soit $\overrightarrow{x} = (x_1, x_2, \ldots, x_n)$ un vecteur inconnu et $\overrightarrow{b} = (b_1, b_2, \ldots, b_n)$ un vecteur connu. On appelle système linéaire toute série d'équations pouvant se mettre sous la forme matricielle*

$$A \cdot \overrightarrow{x} = \overrightarrow{b}.$$

La résolution du système consiste à déterminer les valeurs x_i.

La décomposition $A = QR$ de la matrice permet de résoudre simplement le système linéaire.

$$A \cdot \overrightarrow{x} = QR \cdot \overrightarrow{x} = \overrightarrow{b}$$

en multipliant à gauche par Q^t et en utilisant l'orthogonalité de Q on obtient

$$R \cdot \overrightarrow{x} = \underbrace{Q^t \cdot \overrightarrow{b}}_{\overrightarrow{c}}.$$

Le vecteur $\overrightarrow{c} = (c_1, c_2, \ldots, c_n)$ se calcul par un simple produit matriciel. Il reste à résoudre le système triangulaire

$$\begin{pmatrix} r_{1,1} & r_{1,2} & \ldots & r_{1,n} \\ 0 & r_{2,2} & \ldots & r_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & r_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}.$$

La solution est donc

$$\begin{aligned} x_n &= \frac{c_n}{r_{n,n}} \\ x_i &= \frac{1}{r_{i,i}} \left(c_i - \sum_{j=i+1}^{n} r_{i,j} x_j \right) \quad \forall i \in [1, n-1]. \end{aligned}$$

La solution aussi peut être écrite sous forme matricielle :

$$\overrightarrow{x} = R^{-1} \ \overrightarrow{b}$$

mais la détermination des coefficients de R^{-1} n'est pas aisée

La solution peut donc être calculé (elle existe) si, et seulement si, tous les termes diagonaux de R sont non nul, c'est donc un critère préalable à la résolution du système. Cela n'est pas surprenant puisque l'on sait que A est inversible si $\det(A) \neq 0$, que le déterminant d'un produit QR est le produit des déterminants, que le déterminant d'une matrice orthogonale vaut ± 1 et enfin que le déterminant d'une matrice triangulaire est le produit des termes diagonaux.

La résolution de système linéaire est la méthode la plus sûr numériquement pour calculer l'inverse d'une matrice. Déterminer l'inverse d'une $n \times n$–matrice A, c'est déterminer la matrice B telle que $A \cdot B = I$. Cela revient à chercher les vecteurs colonnes de $B = \left(\overrightarrow{b}_1, \overrightarrow{b}_2, \ldots, \overrightarrow{b}_n \right)$ solution de

$$A \cdot \overrightarrow{b_i} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}.$$

Pour résoudre ce problème, on a besoin que d'une seule décomposition QR de la matrice A.

16.3 Diagonalisation d'une matrice par la décomposition QR

L"utilisation de décomposition QR permet de déterminer par une méthode itérative les valeurs propres d'une $n \times n$–matrice A. Le fondement de la méthode est que deux matrices semblables ont les mêmes valeurs propres.

Considérons une $n \times n$–matrice A munie d'une décomposition QR : $A = Q \cdot R$. Remarquons en premier lieu que la matrice $A' = R \cdot Q$ est semblable à la matrice A. En effet

$$A' = R \cdot Q == I \cdot R \cdot Q = Q^t \cdot Q \cdot R \cdot Q = Q^t \cdot A \cdot Q = Q^{-1} \cdot A \cdot Q$$

La matrice Q est la matrice de passage de A à A'. On peut donc construire la suite

$$\left\{\begin{array}{ccc} A_0 & = & A \\ Q_k, R_k & \text{tel que} & A_k = Q_k \cdot R_k \\ A_{k+1} & = & R_k \cdot Q_k \end{array}\right.$$

Cette suite tend vers une matrice triangulaire ayant les mêmes valeurs propres que A.

> **Théorème (admis) :** *Si A est une matrice inversible de valeurs propres réelles différentes, la suite de matrice A_k converge vers une matrice triangulaire supérieure dont la diagonale est constituée des valeurs propres de A.*

La méthode fonctionne si la matrice de départ A a des valeurs propres de modules différents $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. La vitesse de convergence dépend du rapport entre deux valeurs propres successives $\frac{|\lambda_{k+1}|}{|\lambda_k|}$. Cette méthode ne calcule pas directement les vecteurs propres.

Il faut encore déterminer un critère pour arrêter l'itération, on ne peut la conduire à l'infinie. En pratique elle est stoppée lorsque les éléments sous diagonaux de la matrice A_k sont "suffisamment petit" devant les éléments diagonaux.

Une fois les valeurs propres λ_i obtenues, les vecteurs propres associés $\overrightarrow{v}_i$ sont déterminés en résolvant les systèmes linéaires

$$(A - \lambda_i I) \cdot \overrightarrow{v}_i = 0.$$

Voici un exemple numérique d'itération montrant convergence vers une matrice triangulaire supérieure et la convergence des valeurs sur diagonale correspondant aux valeurs propres de la matrice rangées par module décroissant.

$$A_0 = \begin{pmatrix} 10.000000 & 2.0000000 & 3.0000000 & 5.0000000 \\ 3.0000000 & 6.0000000 & 8.0000000 & 4.0000000 \\ 0.0000000 & 5.0000000 & 4.0000000 & 3.0000000 \\ 0.0000000 & 0.0000000 & 4.0000000 & 3.0000000 \end{pmatrix}$$

$$A_5 = \begin{pmatrix} 13.939985 & 3.9935031 & 4.8753643 & 4.3772740 \\ 0.54256356 & 8.2253866 & -1.5317335 & 0.70956749 \\ 0.0000000 & 1.58412885E-02 & 2.0885437 & 2.6964869 \\ 0.0000000 & 0.0000000 & 0.89015508 & -1.2539154 \end{pmatrix}$$

$$A_{10} = \begin{pmatrix} 14.280842 & 3.4956751 & 5.6976051 & -3.0561123 \\ 3.02423276E-02 & 7.8797555 & -1.7986865 & -0.83473241 \\ 0.0000000 & 6.28821654E-05 & 2.7638412 & -1.6238439 \\ 0.0000000 & 0.0000000 & 0.18058813 & -1.9244391 \end{pmatrix}$$

$$A_{15} = \begin{pmatrix} 14.296490 & 3.4670315 & 5.5460892 & 3.3126199 \\ 1.52919802E-03 & 7.8640842 & -1.8593094 & 0.73819095 \\ 0.0000000 & 3.04792280E-07 & 2.6895354 & 1.8315634 \\ 0.0000000 & 0.0000000 & 2.71366145E-02 & -1.8501107 \end{pmatrix}$$

$$A_{20} = \begin{pmatrix} 14.297274 & 3.4655797 & 5.5683718 & -3.2744837 \\ 7.69680264E-05 & 7.8633018 & -1.8554347 & -0.75025022 \\ 0.0000000 & 1.45188372E-09 & 2.7021308 & -1.8001851 \\ 0.0000000 & 0.0000000 & 4.24155267E-03 & -1.8627062 \end{pmatrix}$$

$$A_{25} = \begin{pmatrix} 14.297314 & 3.4655063 & 5.5648303 & 3.2804687 \\ 3.87307909E-06 & 7.8632617 & -1.8563020 & 0.74822062 \\ 0.0000000 & 6.93884316E-12 & 2.7001970 & 1.8050843 \\ 0.0000000 & 0.0000000 & 6.58778648E-04 & -1.8607720 \end{pmatrix}$$

$$A_{30} = \begin{pmatrix} 14.297316 & 3.4655027 & 5.5653763 & -3.2795382 \\ 1.94893374E-07 & 7.8632598 & -1.8561807 & -0.74852794 \\ 0.0000000 & 3.31460764E-14 & 2.7004986 & -1.8043228 \\ 0.0000000 & 0.0000000 & 1.02418831E-04 & -1.8610731 \end{pmatrix}$$

$$A_{35} = \begin{pmatrix} 14.297316 & 3.4655027 & 5.5652909 & 3.2796824 \\ 9.80703518E-09 & 7.8632598 & -1.8562005 & 0.74847972 \\ 0.0000000 & 1.58347291E-16 & 2.7004519 & 1.8044411 \\ 0.0000000 & 0.0000000 & 1.59203883E-05 & -1.8610264 \end{pmatrix}$$

$$A_{40} = \begin{pmatrix} 14.297316 & 3.4655027 & 5.5653043 & -3.2796595 \\ 4.93489971E-10 & 7.8632598 & -1.8561975 & -0.74848735 \\ 0.0000000 & 7.56456618E-19 & 2.7004590 & -1.8044225 \\ 0.0000000 & 0.0000000 & 2.47478692E-06 & -1.8610337 \end{pmatrix}$$

$$A_{45} = \begin{pmatrix} 14.297316 & 3.4655027 & 5.5653024 & 3.2796626 \\ 2.48324156E-11 & 7.8632598 & -1.8561980 & 0.74848628 \\ 0.0000000 & 3.61375055E-21 & 2.7004578 & 1.8044250 \\ 0.0000000 & 0.0000000 & 3.84698410E-07 & -1.8610326 \end{pmatrix}$$

Il existe des méthodes très astucieuses (sans être forcément compliquées) pour accélérer drastiquement la convergence de la suite, mais elles dépassent la portée de ce cours.

Critères d'évaluation

Les deux listes suivantes correspondent aux compétences exigibles relativent à ce cours, et forment la base de l'évaluation.

Avant chaque calcul complexe (du type calculer les coefficients de Fourier d'une fonction donnée), il sera demandé une présentation écrite de la stratégie mise en oeuvre pour mener le calcul.

La **rédaction** de la copie, lors de l'évaluation, devra cibler un lecteur de votre niveau, et non un spécialiste (l'enseignant). Les phrases que vous auriez prononcées à l'oral pour expliquer un calcul ou une démarche doivent figurer explicitement sur la copie.

Pré-requis

- Calcul de dérivée de fonctions.
- Primitive des fonctions élémentaires.
- Calcul intégral élémentaire (IPP).
- Algèbre linéaire dans $\mathbb{R}^3$: base orthonormée, produit scalaire, norme.

Analyse numérique élémentaire

Compétences générales

- Représentation graphique d'une fonction, d'une méthode, d'un concept.
- Rigueur de la démarche mathématique.
- Présenter à l'écrit un résultat mathématique en indiquant par des phrases claires et concises les grandes étapes du calcul.
- Expression écrite claire et concise.
- Mener à son terme (sans faute) un calcul analytique comportant d'une dizaine d'étapes élémentaires.

Introduction aux probèmes numériques

- Comprendre et savoir expliquer les différents types d'erreurs numériques (Troncature, discrétisation, méthode).
- Savoir anticiper les erreurs numériques lors de la présentation d'un algorithme numérique.
- Comprendre et savoir expliquer la notion de vitesse de convergence d'une suite et de nombre de chiffres significatifs.
- Savoir comparer les performance de deux algorithmes numériques à partir de leurs sorties numériques.
- Savoir mettre en oeuvre une extrapolation de Richardson (par ex. pour les différences finies)

Résolution d'équation

- Comprendre et savoir expliquer le principe et les particularités (avantages et inconvénients) des méthodes de recherche de zéro : point fixe, bissection, fausse position, Newton, sécante.
- Représenter graphiquement les principes des méthodes sus-listées.
- Savoir utiliser un algorithme donné.

Il n'est pas demandé de connaître par coeur les relations de récurrence !

Integration numérique

- Comprendre et savoir expliquer les définitions de formule de quadrature, étage, ordre, poids et noeuds.
- Comprendre et savoir expliquer l'origine des formules de Newton–Cotes.
- Savoir utiliser les formules simples de Newton–Cotes pour évaluer $\int_a^b f(x)dx$.
- Savoir construire une formule composite à partir des formules simples.
- Identifier une méthode d'après son ordre sur une représentation graphique de l'erreur en fonction du coût numérique.
- Comprendre et savoir expliquer le principe de l'intégration de Gauss.
- Implémenter les diverses méthodes numériques dans un langage de programmation C.

Il n'est pas demandé d'apprendre par coeur les valeurs numériques des poids et noeuds.

Intégration numérique des équations différentielles

- Comprendre et savoir expliquer les méthodes de Runge-Kutta.
- Implémenter une méthode de Runge–Kutta à pas fixe et à pas variable (langage C)
- Identifier les difficultés liées aux équations raides.
- Implémenter une méthode d'Euler implicite pour les problèmes raides (langage C).

Méthodes numériques de l'algèbre linéaire

- Comprendre et savoir expliquer la décomposition QR à partir du procédé de Gramm-Schmidt.
- Mettre en oeuvre une décomposition QR pour résoudre un système linéaire (langage C).
- Mettre en oeuvre une méthode de diagonalisation de matrice par la décomposition QR (langage C).

Analyse Mathématique

Série de Fourier

- Comprendre et savoir expliquer le principe de la série de Fourier.
- Savoir calculer les coefficients c_k en distinguant c_0 (une IPP max).
- Savoir prévoir la loi de décroissance de $|c_k|$ à partir des discontinuités.
- Savoir retrouver les relations entre a_k, b_k et c_k.
- Savoir déterminer par un argument de parité, la nullité des a_k ou b_k.
- Représenter graphiquement le spectre d'une fonction.
- Comprendre et savoir expliquer la notion de variation bornée pour appliquer le théorème de Dirichlet.

Il n'est pas demandé de connaître par coeur les formules

Systèmes orthogonaux

- Savoir appliquer la notion de produit scalaire et de norme aux fonctions.
- Comprendre et savoir expliquer la notion de série de Fourier généralisée et la méthode de calcul des coefficients par le produit scalaire.
- Comprendre le lien avec l'algèbre dans $\mathbb{R}^3$.
- Savoir appliquer les résultats généraux pour retrouver l'expression des coefficients de Fourier d'une fonction périodique.
- Savoir démontrer l'orthogonalité d'une famille de fonction.

Série de Fourier

- Comprendre et savoir expliquer le principe de la série de Fourier.
- Savoir calculer les coefficients c_k en distinguant c_0 (une IPP max).
- Savoir prévoir la loi de décroissance de $|c_k|$ à partir des discontinuités.
- Savoir retrouver les relations entre a_k, b_k et c_k.
- Savoir déterminer par un argument de parité, la nullité des a_k ou b_k.
- Représenter graphiquement le spectre d'une fonction.
- Comprendre et savoir expliquer la notion de variation bornée pour appliquer le théorème de Dirichlet.

Il n'est pas demandé de connaître par coeur les formules

Systèmes orthogonaux

- Savoir appliquer la notion de produit scalaire et de norme aux fonctions.
- Comprendre et savoir expliquer la notion de série de Fourier généralisée et la méthode de calcul des coefficients par le produit scalaire.
- Comprendre le lien avec l'algèbre dans $\mathbb{R}^3$.
- Savoir appliquer les résultats généraux pour retrouver l'expression des coefficients de Fourier d'une fonction périodique.
- Savoir démontrer l'orthogonalité d'une famille de fonction.

Equations aux dérivées partielles

- Mettre en oeuvre la méthode de séparation des variables à deux et trois dimensions en géométrie cartésienne et cylindrique.
- Distinguer les conditions aux bords des conditions initiales.
- Déterminer la solution en utilisant les séries de Fourier généralisée.

Transformée de Fourier

- Comprendre et savoir expliquer le passage de la série à la transformée de Fourier
- Appliquer les propriétés générales (linéarité, changement d'échelle, translation, dérivation et intégration) pour calculer une transformée de Fourier.
- Comprendre et savoir expliquer la notion de Produit de convolution