Le Frido 2018,
volume 4
Laurent Claessens

Plusieurs extensions et versions de ce livre.

1. La version courante, régulièrement mise à jour et qui deviendra petit à petit le Frido 2019. Téléchargeable sur

   `https://laurent.claessens-donadello.eu/pdf/lefrido.pdf`

2. La version la plus complète, contenant des exercices ainsi que de la mathématique de niveau recherche sur

   `https://laurent.claessens-donadello.eu/pdf/mazhe.pdf`

3. Et bien entendu les sources LaTeX

   `https://tuleap.net/plugins/git/lefrido/lefrido`
   `https://github.com/LaurentClaessens/mazhe`

Copyright 2011-2018 Laurent Claessens, Carlotta Donadello, Lilian Besson, Bertrand Desmons, and many other contributors. A complete list could be retreivied from the git log.
Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the chapter entitled "GNU Free Documentation License".

(c) 2015 David Revoy pour les illustrations de couverture CC-BY,
`https://www.peppercarrot.com/`

ISBN : 979-10-97085-13-1

Chapitre 30

Séries de Fourier

Soit f une fonction ; nous définissons ses **coefficients de Fourier** par

$$c_n(f) = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-int} \tag{30.1}$$

avec plus de détails en 28.5.6.

30.1 Densité des polynômes trigonométriques

30.1.1 Convergence pour les fonctions continues (via Weierstrass)

Le résultat fondamental qui nous permet d'utiliser les polynômes trigonométriques comme base pour les fonctions *continues* périodiques est le suivant. Notons que pour les fonctions non continues, il y a encore du travail.

Lemme 30.1.
Si $f\colon \mathbb{R} \to \mathbb{C}$ est une fonction continue 2π-périodique et si $\epsilon > 0$, alors il existe un polynôme trigonométrique P tel que $\|f - P\|_\infty \leqslant \epsilon$.

Démonstration. Nous allons utiliser le théorème de Stone-Weierstrass 19.4. Soit le compact Hausdorff

$$S^1 = \{z \in \mathbb{C} \text{ tel que } |z| = 1\}, \tag{30.2}$$

et $C(S^1, \mathbb{C})$ l'algèbre des fonctions continues de S^1 vers $\mathbb{C}$. Il suffit de vérifier que les polynômes trigonométriques vérifient les hypothèse du théorème de Stone-Weierstrass. Un polynôme trigonométrique est un polynôme en z et $\bar{z}$ défini sur S^1.

(1) Le polynôme constant est dans l'algèbre, ok.

(2) Pour la séparation des points, le polynôme trigonométrique $x \mapsto e^{ix}$.

(3) Si P est un polynôme en z et $\bar{z}$, alors $\bar{P}$ l'est encore.

Donc si $\epsilon > 0$ et $\tilde{f} \in C(S^1, \mathbb{C})$ sont donnés, il existe un polynôme trigonométrique P tel que

$$\sum_t |\tilde{f}(e^{it}) - P(t)| < \epsilon. \tag{30.3}$$

Soit $f\colon \mathbb{R} \to \mathbb{C}$ une fonction continue 2π-périodique. Nous considérons $\tilde{f} \in C(S^1, \mathbb{C})$ donnée par $\tilde{f}(e^{it}) = f(t)$. Alors $\sup_t |f(t) - P(t)| \leqslant \epsilon$. □

30.1.2 Convergence pour les fonctions continues (via Fejér)

Si nous ne voulons pas passer par le gros théorème de Stone-Weierstrass pour prouver la densité des polynômes trigonométrique dans $\left(C^0_{2\pi}, \|.\|_\infty\right)$, nous pouvons passer par le gros théorème de Fejér. C'est ce que nous faisons maintenant.

Le **noyau de Dirichlet** est la fonction

$$D_n(t) = \sum_{k=-n}^{n} e^{int}. \tag{30.4}$$

Le **noyau de Fejér** est la moyenne de Cesaro des noyaux de Dirichlet :

$$F_n(t) = \frac{1}{n} \sum_{k=0}^{n-1} D_k(t). \tag{30.5}$$

Lemme 30.2.
Le noyau de Dirichlet s'exprime sous la forme

$$D_n(t) = \sum_{k=-n}^{n} e^{-ikt} = \frac{\sin\left(\frac{2n+1}{2}t\right)}{\sin(t/2)} \tag{30.6}$$

Note : ce noyau n'est pas positif.

Démonstration. Nous commençons par mettre en facteur le premier terme :

$$D_n(t) = \sum_{k=-n}^{n} e^{int} = e^{-int} \sum_{k=0}^{2n} e^{ikt}. \tag{30.7}$$

En utilisant la formule de la somme géométrique,

$$D_n(t) = e^{-int} \frac{1 - (e^{it})^{2n+1}}{1 - e^{it}} \tag{30.8a}$$

$$= e^{-int} \frac{1 - e^{(2n+1)it}}{1 - e^{it}} \tag{30.8b}$$

$$= e^{-int} \frac{e^{(2n+1)it/2}}{e^{i\frac{t}{2}}} \frac{e^{-(2n+1)it/2} - e^{(2n+1)it/2}}{e^{-it/2} - e^{it/2}} \tag{30.8c}$$

$$= \frac{(-2i)\sin\left(\frac{2n+1}{2}t\right)}{(-2i)\sin\left(\frac{t}{2}\right)}. \tag{30.8d}$$

□

Théorème 30.3 (Théorème de Dirichlet)**.**
Soit f une fonction 2π-périodique et C^1 par morceaux. Pour tout $x \in \mathbb{R}$ nous posons

$$s_n(x) = \sum_{k=-n}^{n} c_k(f) e^{ikx}. \tag{30.9}$$

Alors nous avons

$$\lim_{n\to\infty} s_n(x) = \frac{f(x^+) + f(x^-)}{2}. \tag{30.10}$$

Lemme 30.4.
Le noyau de Fejér s'exprime sous la forme

$$F_n(t) = \frac{1}{n} \left(\frac{\sin\frac{nt}{2}}{\sin\frac{t}{2}} \right)^2. \tag{30.11}$$

Note : ce noyau est positif. C'est important parce qu'on s'en sert dans la preuve du théorème de Fejér.

Démonstration. L'astuce est de noter $\sin(x) = \Im(e^{ix})$ et de repartir du résultat à propos du noyau de Dirichlet. En utilisant encore la formule de la série géométrique partielle,

$$F_n(t) = \frac{1}{n\sin(t/2)}\Im\sum_{k=0}^{n-1} e^{(2k+1)it/2} \tag{30.12a}$$

$$= \frac{1}{n\sin(t/2)}\Im e^{\frac{it}{2}}\sum_{k=0}^{n-1} \tag{30.12b}$$

$$= \frac{1}{n\sin(t/2)}\Im e^{\frac{it}{2}}\left(\frac{1-e^{nit}}{1-e^{it}}\right) \tag{30.12c}$$

$$= \frac{1}{n\sin(t/2)}\Im e^{it/2}\frac{e^{\frac{nit}{2}}\left(e^{-\frac{int}{2}}-e^{\frac{nit}{2}}\right)}{e^{\frac{it}{2}}\left(e^{-it/2}-e^{it/2}\right)} \tag{30.12d}$$

$$= \frac{1}{n\sin(t/2)}\underbrace{\Im e^{\frac{nit}{2}}}_{\sin(nt/2)}\frac{\sin\left(\frac{nt}{2}\right)}{\sin(\frac{t}{2})} \tag{30.12e}$$

$$= \frac{1}{n}\left(\frac{\sin\frac{nt}{2}}{\sin\frac{t}{2}}\right)^2. \tag{30.12f}$$

□

Théorème 30.5 (Fejèr)**.**
Soit $f\colon \mathbb{R}\to\mathbb{C}$ une fonction continue et 2π-périodique. Pour tout $k\in\mathbb{Z}$ nous notons

$$\begin{aligned} e_k\colon \mathbb{R} &\to \mathbb{C} \\ x &\mapsto e^{ikx}. \end{aligned} \tag{30.13}$$

Pour chaque $n\in\mathbb{N}$ nous posons

$$D_n = \sum_{k=-n}^{n} e_k \qquad \tilde{S}_n(f) = \sum_{k=-n}^{n} c_k(f)e_k \tag{30.14a}$$

$$F_n = \frac{D_0+\cdots+D_{n-1}}{n} \qquad \tilde{F}_n = \sigma_n(f) = \frac{1}{n}\sum_{k=0}^{n-1} S_k(f). \tag{30.14b}$$

Alors

1. *(1)* $\frac{1}{2\pi}\int_{-\pi}^{\pi} F_n(t)dt = 1$.
2. *(2) Pour tout $\alpha\in\,]0,\pi[$, F_n converge uniformément vers 0 sur $[-\pi,\pi]\backslash[-\alpha,\alpha]$.*
3. *(3) La suite $\tilde{F}_n$ converge uniformément sur $\mathbb{R}$ vers f.*
4. *(4) Le système trigonométrique $\{e_k\}_{k\in\mathbb{Z}}$ est total pour l'espace $\left(C^0(S^1),\|.\|_\infty\right)$ des fonctions continues 2π-périodiques.*

Démonstration. Un calcul usuel montre que

$$\int_{-\pi}^{\pi} e_l(t)dt = \begin{cases} 0 & \text{si } l\neq 0 \\ 2\pi & \text{si } l=0 \end{cases} \tag{30.15}$$

Nous avons alors

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} F_n(t)dt = \frac{1}{2\pi}\frac{1}{n}\sum_{k=0}^{n-1}\sum_{l=-k}^{k}\underbrace{\int_{-\pi}^{\pi} e_l(t)dt}_{2\pi\delta_l} = \frac{1}{n}\sum_{k=0}^{n-1} 1 = 1. \tag{30.16}$$

Cela prouve déjà le premier point.

Pour le second point, en partant de l'expression (30.11) et en considérant $x \in [-\pi, \pi,]\backslash[-\alpha, \alpha]$ (ce qui nous évite l'annulation du dénominateur),

$$|F_n(x)| \leqslant \frac{1}{(n+1)\sin^2(\alpha/2)}, \tag{30.17}$$

et donc $F_n \to 0$ uniformément sur l'ensemble considéré.

Nous passons maintenant à cette histoire de convergence uniforme de la moyenne de Cesaro vers f. Pour tout $n \in \mathbb{N}$ nous avons

$$\tilde{D}_n(x) = \frac{1}{2\pi} \sum_{k=-n}^{n} \left(\int_{-\pi}^{\pi} f(t) e^{-ikt} dt \right) e^{ikx} \tag{30.18a}$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \sum_{k=-n}^{n} e_k(x-t) \tag{30.18b}$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) D_k(x-t). \tag{30.18c}$$

Par conséquent, en effectuant le changement de variable $u = x - t$ et la périodicité,

$$\tilde{F}_n(x) = \int_{-\pi}^{\pi} f(t) F_n(x-t) dt \tag{30.19a}$$

$$= -\int_{x+\pi}^{x-\pi} f(x-u) F_n(u) du \tag{30.19b}$$

$$= \int_{-\pi}^{\pi} f(x-u) F_n(u) du. \tag{30.19c}$$

Nous prouvons à présent l'uniforme continuité. Soit $\epsilon > 0$; étant donné que f est continue et 2π-périodique, elle est uniformément continue et nous considérons $\delta > 0$ tel que $|x-y| < \delta$ implique $\big|f(x) - f(y)\big| < \epsilon$. Soit M un majorant de $|f|$ sur $\mathbb{R}$. L'équation (30.19) nous donne

$$\big|f(x) - \tilde{F}_n(x)\big| = \big|\frac{1}{2\pi} \int_{-\pi}^{\pi} \big(f(x-t) - f(x)\big) F_n(t) dt\big| \tag{30.20a}$$

$$\leqslant \frac{1}{2\pi} \int_{\delta \leqslant |t| \leqslant \pi} |2M F_n(t)| dt + \frac{1}{2\pi} \int_{-\delta}^{\delta} \epsilon |F_n(t)| dt \tag{30.20b}$$

$$\leqslant \frac{2M}{2\pi} \int_{\delta \leqslant |t| \leqslant \pi} F_n(t) dt + \epsilon' \tag{30.20c}$$

Pour obtenir (30.20a) nous avons pu rentrer $f(x)$ dans l'intégrale en utilisant le premier point. Pour obtenir (30.20c) nous avons d'abord utilisé la positivité de F_n (lemme 30.4) pour enlever les valeurs absolues, et nous avons ensuite utilisé le fait que son intégrale valait 2π.

Étant donné que $F_n \to 0$ uniformément sur $[-\pi, \pi,]\backslash[-\alpha, \alpha]$, il existe un N tel que

$$\int_{\delta \leqslant |t| \leqslant \pi} F_n(t) dt \leqslant \epsilon \tag{30.21}$$

dès que $n > N$. Le résultat découle.

Pour le point (4), il suffit de remarquer que chacun des $\tilde{F}_n$ est une combinaison finie d'éléments du système trigonométrique. □

30.1.3 Densité dans L^p

Nous venons de voir (de deux façons différentes) que les polynômes trigonométriques étaient dense dans $\big(C^0_{2\pi}(\mathbb{R}), \|.\|_\infty\big)$. Nous avons aussi déjà vu par le théorème 28.54 que ces polynômes trigonométriques étaient denses dans $L^p(S^1)$. Nous présentons à présent une autre façon de prouver cette dernière densité.

Théorème 30.6.
Les polynômes trigonométriques sont denses dans $L^p(S^1)$ pour $1 \leqslant p < \infty$.

Démonstration. Par les théorèmes 30.1 ou 30.5 (au choix), nous savons que les polynômes trigonométriques sont denses dans $\left(C^0_{2\pi}(S^1), \|.\|_\infty\right)$. Vu que S^1 est compact, la densité est également au sens L^p. En effet si $\|f_n - f\|_\infty \leqslant \epsilon$, alors

$$\|f_n - f\|_\infty = \int_0^{2\pi} |f_n - f|^p \leqslant \int_0^{2\pi} \epsilon^p = 2\pi\epsilon^p. \tag{30.22}$$

Donc les polynômes trigonométriques sont denses dans $\left(C^0_{2\pi}(S^1), \|.\|_p\right)$. Mais nous savons par (un a fortiori sur) le théorème 28.34 que les fonctions continues sont denses dans $L^p(S^1)$.

Par densité de la densité, les polynômes trigonométriques sont denses dans $L^p(S^1)$. □

30.1.4 Suite équirépartie, critère de Weyl

Définition 30.7.
Soit u une suite dans $[0,1]$. Pour $0 \leqslant a \leqslant b \leqslant 1$ nous posons

$$X_n(a,b) = \text{Card}\left\{k \in \{1,\dots,n\} \text{ tel que } u_k \in [a,b]\right\}. \tag{30.23}$$

*Nous disons que la suite u est **équirépartie** si pour tout $0 \leqslant a < b < 1$, on a*

$$\lim_{n\to\infty} \frac{X_n(a,b)}{n} = b - a. \tag{30.24}$$

Voir aussi la remarque 38.132 sur les nombres normaux.

Proposition 30.8 (Critère de Weyl[233, 4])**.**
Soit (x_n) une suite dans $[0,1[$. Les conditions suivantes sont équivalentes.

(1) La suite (x_n) est équirépartie.

(2) Pour toute fonction continue à valeurs réelles sur $[0,1]$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(x_k) = \int_0^1 f(x)dx. \tag{30.25}$$

(3) Pour tout $p \in \mathbb{N}^*$ nous avons

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} e^{2i\pi p x_k} = 0. \tag{30.26}$$

Démonstration. On pose

$$S_n(f) = \frac{1}{n} \sum_{k=1}^{n} f(x_k). \tag{30.27}$$

Une espèce de lemme Supposons connaitre un ensemble de fonctions A dense dans $C^0([0,1])$ pour toutes les fonctions duquel nous avons la limite (30.25). Alors la limite a lieu pour toute fonction de $C^0([0,1])$. En effet, soit $f \in C^0([0,1])$ et $g \in A$ tel que $\|f - g\|_\infty < \epsilon$. Alors

$$\left\| \frac{1}{n} \sum_{k=1}^{n} f(x_k) - \int_0^1 f(t)dt \right\| \leqslant \left\| \frac{1}{n} \sum_{k=1}^{n} \big(f(x_k) - g(x_k)\big) \right\| \tag{30.28a}$$

$$+ \left\| \frac{1}{n} \sum_{k=1}^{n} g(x_k) - \int_0^1 g(t)dt \right\| \tag{30.28b}$$

$$+ \left\| \int_0^1 g(t)dt - \int_0^1 f(t)dt \right\|. \tag{30.28c}$$

Le premier terme se majore par ϵ. Le troisième est la même majoration : $\int_0^1 \big(f(t)-g(t)\big)dt \leqslant \|f-g\|_\infty = \epsilon$. Par hypothèse sur l'espace A, le second terme se majore par ϵ lorsque n est grand.

(1)⇒(2) Nous supposons que la suite est équirépartie et nous commençons par montrer le résultat pour les fonctions en escalier. Soit donc la fonction en escalier $\eta(x) = c_j$ sur $a_{j-1} < x < a_j$. Sur le point a_j lui-même, la fonction η vaut soit c_j soit c_{j+1}. Nous avons

$$\frac{1}{n}\sum_{k=1}^{n}\eta(x_k) = \frac{1}{n}\left[\sum_{j=1}^{m} c_j X_n(a_j,a_{j+1}) - \sum_{j=1}^{m} c_j X_n(a_j,a_j) + \sum_{j=1}^{m}\eta(a_j)X_n(a_j,a_j)\right]. \tag{30.29}$$

À la limite $n\to\infty$, les deux derniers termes tombent [1] et il reste

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\eta(x_k) = \sum_{j=1}^{m} c_j(a_{j-1}-a_j). \tag{30.30}$$

Or par construction, pour une fonction en escalier,

$$\sum_{j=1}^{m} c_j(a_{j-1}-a_j) = \int_0^1 \eta. \tag{30.31}$$

Étant donné que les fonctions en escalier sont denses dans les fonctions continues, l'espèce de lemme plus haut conclut.

(2)⇒(1) Nous prouvons maintenant le sens inverse. C'est à dire que pour toute fonction continue sur $[0,1]$, nous avons

$$\int_0^1 f(x)dx = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} f(x_k). \tag{30.32}$$

Nous devons en déduire que (x_n) est équirépartie. Pour ce faire, soit $x\in[0,1[$ et $\epsilon>0$ tel que $x+\epsilon<1$. Nous considérons $\varphi = \mathbb{1}_{[x,1[}$ et

$$\varphi_{\epsilon(t)} = \begin{cases} 0 & \text{si } t\in[0,x[\\ \frac{t-x}{\epsilon} & \text{si } t\in[x,x+\epsilon[\\ 1 & \text{si } t\geqslant x+\epsilon. \end{cases} \tag{30.33}$$

Cela est une fonction continue, donc

$$\lim_{n\to\infty} S_n\big(\varphi_\epsilon(t)\big) = \int_0^1 \varphi_\epsilon(t)dt = \int_x^{x+\epsilon}\frac{t-x}{\epsilon}dt + \int_{x+\epsilon}^1 1dt = 1-x-\frac{\epsilon}{2}. \tag{30.34}$$

Mais $\varphi_\epsilon \leqslant \varphi$, donc $S_n(\varphi_\epsilon)\leqslant S_n(\varphi)$ et donc

$$\liminf_{n\to\infty} S_n(\varphi) \geqslant 1-x. \tag{30.35}$$

Notons que nous ne savons pas si la *vraie* limite de gauche existe ; c'est pourquoi nous prenons la limite inférieure, qui existe toujours.

Nous définissons aussi

$$\psi_\epsilon(t) = \begin{cases} 0 & \text{si } t\in[0,x-\epsilon[\\ \frac{t-x+\epsilon}{\epsilon} & \text{si } t\in[x-\epsilon,x[\\ 1 & \text{si } t>x. \end{cases} \tag{30.36}$$

C'est encore une fonction continue et nous trouvons [2]

$$\int_0^1 \psi_\epsilon(t)dt = 1-x+\frac{\epsilon}{2}. \tag{30.37}$$

1. J'en profite pour mentionner que mon équation (30.29) n'est pas la même que celle de [233] dans laquelle il me semble voir une faute ; quoi qu'il en soit, les termes litigieux tombent.

2. Je recommande chaudement de dessiner les fonctions φ_ϵ et ψ_ϵ pour avoir une idée de la situation.

Vu que $\psi_\epsilon \geqslant \varphi$, nous avons $S_n(\psi_\epsilon) \geqslant S_n(\varphi)$ et donc

$$\limsup_n S_n(\varphi) \leqslant 1 - x. \tag{30.38}$$

Nous avons déjà obtenu que

$$1 - x \leqslant \liminf S_n(\varphi) \leqslant \limsup S_n(\varphi) \leqslant 1 - x, \tag{30.39}$$

donc la limite existe et vaut

$$\lim_{n\to\infty} S_n(\varphi) = 1 - x. \tag{30.40}$$

Cela est pour la fonction caractéristique $\varphi = \mathbb{1}_{[x,1[}$. Si nous prenons une fonction caractéristique $\mathbb{1}_{[a,b]}$, nous avons la même chose parce que $\mathbb{1}_{[a,b[}$ est une combinaisons linéaire de fonctions du type $\mathbb{1}_{[x,1[}$.

Nous avons donc

$$\lim_{n\to\infty} S_n\big(\mathbb{1}_{[a,b]}\big) = b - a, \tag{30.41}$$

alors que le membre de gauche n'est autre que

$$S_n\big(\mathbb{1}_{[a,b]}\big) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{[a,b]}(x_k) = \frac{1}{n}N(n,a,b). \tag{30.42}$$

(2)⇒(3) Vu que $e^{2i\pi p x_k} = \cos(2\pi p x_k) + \sin(2\pi i x_k)$ est une fonction périodique, c'est immédiat.

(3)⇒(2) Par linéarité, le point (2) montre que si f est un polynôme trigonométrique, alors

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(x_k) = \int_0^1 f(t)dt. \tag{30.43}$$

Densité des polynômes trigonométriques Il nous reste à prouver que les polynômes trigonométriques sont denses dans les fonctions continues sur $[0,1]$. Soit une fonction continue sur $[0,1]$ avec $f(0) = f(1)$. Alors le théorème de Stone-Weierstrass dans sa version trigonométrique (lemme 30.1) nous donne la densité.

Si $f(1) \neq f(0)$ c'est pas très grave : on peut trouver une fonction g vérifiant $g(0) = g(1)$ et $\|f - g\|_\infty \leqslant \epsilon$. Ensuite un polynôme trigonométrique approxime très bien g. .

□

30.2 Fonctions de Dirichlet

Définition 30.9.
*Une fonction $f\colon \mathbb{R} \to \mathbb{C}$ est une **fonction de Dirichlet** si*

(1) elle est 2π-périodique,

(2) elle est continue par morceaux,

(3) pour tout $x \in \mathbb{R}$ nous avons

$$f(x) = \frac{f(x^+) + f(x^-)}{2}. \tag{30.44}$$

Nous notons $\mathcal{D}$ l'ensemble des fonctions de Dirichlet.

Lemme 30.10 ([364]).
L'ensemble $C^0(S^1)$ est dense dans $\big(\mathcal{D}, \|.\|_2\big)$.

Démonstration. Nous commençons par supposer que $f \in \mathcal{D}$ n'ait qu'un seul point de discontinuité, x_0. Alors nous considérons la fonction f_n qui est égale à f sur $S^1 \backslash B(x_0, \frac{1}{n})$ et qui sur $B(x_n, \frac{1}{n})$ est le segment de droite joignant $f(x_0 - \frac{1}{n})$ et $f(x_0 + \frac{1}{n})$. Cela est une fonction continue, et de plus nous avons

$$|f_n(x)| \leqslant \|f\|_\infty \tag{30.45}$$

pour tout x. En effet si x est en dehors de $B(x_0, \frac{1}{n})$ c'est évident, et si $x \in B(x_0, \frac{1}{n})$, alors $|f_n(x)|$ est majoré soit par $f(x_0 - \frac{1}{n})$ soit par $f(x_0 + \frac{1}{n})$ suivant que le raccord affin soit croissant ou décroissant. Avec ça nous avons

$$\|f_n - f\|_2^2 = \int_{x_0-1/n}^{x_0+1/n} |f(x) - f_n(x)|^2 dx \leqslant \int_{x_0-1/n}^{x_0+1/n} 4\|f\|_\infty = \frac{8\|f\|_\infty}{n}. \tag{30.46}$$

Et nous voyons que $\|f_n - f\|_2 \to 0$.

Si f contient plusieurs points de continuité, on fait le même coup autour de chaque point, en prenant n assez grand pour que si x_0 est un point de discontinuité, $B(x_0, \frac{1}{n})$ n'en contienne pas d'autres. □

Notons que la densité de $C^0(S^1)$ dans $\big(\mathcal{D}, \|.\|_\infty\big)$ est impossible parce qu'une limite uniforme de fonctions continues est continue.

Théorème 30.11.
Le système trigonométrique $\{e_n\}_{n\in\mathbb{Z}}$ est total dans $\big(\mathcal{D}, \|.\|_2\big)$.

Démonstration. Soit $f \in \mathcal{D}$. Si elle est continue, le théorème de Fejèr 30.5 nous donne convergence uniforme sur S^1 d'une suite de polynômes trigonométriques vers f. Cette convergence est également une convergence L^2 parce que S^1 est compact.

Prenons donc $f \in \mathcal{D}$ non continue et $\epsilon > 0$[3]. Par le lemme 30.10, il existe une fonction $g \in C^0(S^1)$ telle que

$$\|g - f\|_2 \leqslant \epsilon. \tag{30.47}$$

Le théorème de Fejèr donne aussi un polynôme trigonométrique P tel que $\|P - g\|_2 < \epsilon$; nous avons alors

$$\|P - f\|_2 \leqslant \|P - g\|_2 + \|g - f\|_2 \leqslant 2\epsilon. \tag{30.48}$$

□

Notons que cette histoire de fonctions de Dirichlet n'a pas attaquée le vrai fond du problème de la densité des polynômes trigonométriques dans $L^2(S^1)$ parce que nous restons avec une hypothèse de continuité, alors que les représentants des éléments de $L^2(S^1)$ n'ont strictement aucune régularité a priori.

30.3 Coefficients et série de Fourier

Définition 30.12.
*La **série de Fourier** associée à f est*

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n(f) e^{2\pi i \frac{n}{T} x}. \tag{30.49}$$

Cette expression est pour l'instant purement formelle. Cela ne présume ni de la convergence de la série, ni, au cas où elle serait convergente, que la limite soit f.

Pour la suite nous allons considérer des fonctions périodiques de période 2π, et les coefficients de Fourier de f (quand ils existent) sont alors

$$c_n(f) = \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-int} dt \tag{30.50}$$

3. Par exemple $\epsilon = 0.4$, mais ce n'est qu'un exemple hein. Si vous en voulez un autre, prenez p, un nombre premier puis calculez $\epsilon = 1/p$.

Proposition 30.13 ([365]).
Soit f une fonction continue et périodique telle que sa série de Fourier converge uniformément. Alors la convergence est vers f.

Démonstration. Notons d'abord que f étant continue sur $[0,2\pi]$, elle y est bornée et L^2. Par conséquent Parseval nous enseigne que

$$\|S_N(f)-f\|_{L^2}\to 0. \tag{30.51}$$

Cela signifie que

$$\lim_{N\to\infty}\frac{1}{2\pi}\int_0^{2\pi}|f(t)-S_N(t)|^2dt=0. \tag{30.52}$$

L'hypothèse de convergence uniforme nous dit que la fonction $|f(t)-S_N(t)|^2$ converge uniformément vers la fonction $|f(t)-S(t)|^2$ où nous avons écrit S la limite de S_N. En permutant la limite et l'intégrale,

$$\frac{1}{2\pi}\int_0^{2\pi}|f(t)-S(t)|^2dt=0, \tag{30.53}$$

ce qui signifie que la fonction $t\mapsto|f(t)-S(t)|^2$ est la fonction nulle. Nous en déduisons que $f=S$. □

Proposition 30.14.
Soit f une fonction 2π-périodique. Si $\sum_{n\in\mathbb{Z}}|c_n(f)|<\infty$, alors pour tout $x\in\mathbb{R}$ nous avons

$$f(x)=\sum_{n\in\mathbb{Z}}c_n(f)e^{inx}. \tag{30.54}$$

De plus, la suite (S_nf) converge uniformément vers f.

Démonstration. Nous posons

$$g(x)=\sum_{n\in\mathbb{Z}}c_n(f)e^{inx}. \tag{30.55}$$

Étant donné les hypothèses, la série de droite converge absolument, la fonction g est continue sur $\mathbb{R}$. Nous avons

$$\big|g(x)-(S_nf)(x)\big|\leqslant\sum_{|k|>n}|c_k(f)|, \tag{30.56}$$

mais le terme de droite tend vers zéro lorsque $n\to\infty$ parce que c'est le reste d'une série convergente. Cela signifie que S_nf converge uniformément vers g.

Par ailleurs nous savons que dans L^2 nous avons la convergence $S_nf\to f$ (parce que f est continue sur le compact $[0,2\pi]$ et donc y est bornée et L^2), ce qui signifie que $g=f$ presque partout au sens L^2. Ces deux fonctions étant continues, elles sont égales partout. □

Théorème 30.15.
Soit f, une fonction C^1 et 2π-périodique. Nous notons $(c_n)_{n\in\mathbb{Z}}$ la suite de ses coefficients de Fourier. Alors $(c_n)\in\ell^1(\mathbb{Z})$ et pour tout $x\in\mathbb{R}$ nous avons

$$f(x)=\sum_{n\in\mathbb{Z}}c_n(f)e^{inx}. \tag{30.57}$$

Démonstration. Soit $n\in\mathbb{Z}$. Nous posons $g(t)=f(t)e^{-int}$. Nous avons

$$0=g(2\pi)-g(0)=\int_0^{2\pi}g'(t)dt=\int_0^{2\pi}\big[f'(t)e^{-int}-inf(t)e^{-int}\big]. \tag{30.58}$$

Du coup, $c_n(f')=inc_n(f)$. La fonction f' étant bornée (parce que continue sur $[0,2\pi]$), elle est de carré intégrable sur $[0,2\pi]$ et par les inégalités de Parseval (théorème 27.44) nous avons

$$\sum_{n\in\mathbb{Z}}|c_n(f')|^2<\infty. \tag{30.59}$$

Par conséquent $(c_n(f')) \in \ell^2(\mathbb{Z})$ et a fortiori $(c_n(f'))_{n\in\mathbb{N}} \in \ell^2(\mathbb{N})$. L'inégalité de Cauchy-Schwarz nous indique alors

$$\sum_{n\in\mathbb{N}} |c_n(f)| = \sum_{n\in\mathbb{N}} \frac{1}{n}|c_n(f')| \leqslant \left(\sum_n \frac{1}{n^2}\right)^{1/2} \left(\sum_n |c_n(f')|^2\right)^{1/2} < \infty. \tag{30.60}$$

Nous procédons de même pour $n < 0$. Cela prouve que

$$\sum_{n\in\mathbb{Z}} |c_n(f)| < \infty. \tag{30.61}$$

□

Corollaire 30.16.
Soient f, g deux fonctions continues et 2π-périodiques. Si $c_n(f) = c_n(g)$ alors $f = g$.

Démonstration. Dans le cas de fonctions continues, le théorème de Fejér nous enseigne que si nous posons

$$S_n(x) = \sum_{k=-n}^{n} c_k(f)e^{ikx} \tag{30.62}$$

alors nous avons la convergence

$$\frac{1}{N+1}\sum_{n=0}^{N} S_n(f)(x) \to f(x). \tag{30.63}$$

C'est à dire qu'une fonction continue est déterminée par ses coefficients de Fourier. □

Exemple 30.17
Considérons la fonction

$$f(x) = 1 - \frac{x^2}{\pi^2} \tag{30.64}$$

sur $[-\pi, \pi]$. Nous la développons en série trigonométrique, et étant paire il n'y a pas de sinus. Un calcul montre que

$$a_0 = \frac{4}{3} \tag{30.65}$$

et

$$a_n = (-1)^{n+1}\frac{4}{n^2\pi^2}, \tag{30.66}$$

de telle sorte que

$$f(x) = \frac{2}{3} - \frac{4}{\pi^2}\sum_{n=1}^{\infty}(-1)^n\frac{\cos(nx)}{n^2}. \tag{30.67}$$

Nous avons $f(\pi) = 0$, mais vu le développement,

$$f(\pi) = \frac{2}{3} - \frac{4}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}, \tag{30.68}$$

donc

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}. \tag{30.69}$$

△

30.3.1 Le contre-exemple que nous attendions tous

Nous montrons maintenant que la continuité et la périodicité ne sont pas suffisantes pour avoir convergence de la série de Fourier.

Proposition 30.18 ([4]).
Soit $C^0_{2\pi}$ l'ensemble des fonctions continues muni de la norme uniforme. Nous définissons

$$S_n(f)(x) = \sum_{k=-n}^{n} c_k(f)e^{ikx}. \tag{30.70}$$

Alors il existe $f \in C^0_{2\pi}$ tel que la suite $n \mapsto S_n(f)(0)$ soit divergente. En particulier f n'est pas la somme de sa série de Fourier.

Démonstration. Nous considérons la forme linéaire

$$\begin{aligned} l_n\colon C^0_{2\pi} &\to \mathbb{C} \\ f &\mapsto S_n(f)(0) = \sum_{k=-n}^{n} c_k(f). \end{aligned} \tag{30.71}$$

La forme est continue Nous montrons d'abord que $\|l_n\|$ est continue en montrant que $\|l_n\| < \infty$ et en utilisant la proposition 12.21. Pour cela nous calculons un peu :

$$l_n(f) = \sum_{k=-n}^{n} \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-ikt}dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\sum_{k=-n}^{n} e^{-ikt}dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)D_n(t)dt \tag{30.72}$$

où $D_n(t)$ est le noyaux de Dirichlet dont nous savons une formule par le lemme 30.2. Nous avons donc

$$|l_n(f)| \leqslant \frac{1}{2\pi}\int_{-\pi}^{\pi} |D_n(t)|\|f\|_\infty dt. \tag{30.73}$$

En prenant $\|f\|_\infty = 1$ nous avons la borne suivante pour la norme de l_n :

$$\|l_n\| \leqslant \frac{1}{2n}\int_{-\pi}^{\pi} |D_n(t)|dt < \infty. \tag{30.74}$$

Notons que la convergence de l'intégrale vient de la continuité de la fonction

$$t \mapsto \frac{\sin\left(\frac{2n+1}{2}t\right)}{\sin\left(\frac{t}{2}\right)} \tag{30.75}$$

qui, elle même, se prouve avec une règle de l'Hospital :

$$\lim_{t\to 0}\frac{\sin(at)}{\sin(t)} = \lim_{t\to 0}\frac{a\cos(at)}{\cos(t)} = a. \tag{30.76}$$

Donc $D_n(t)$ a une limite bien définie pour $t \to 0$ et est alors une fonction continue sur le compact $[-\pi, \pi]$.

La norme de l_n (début) Nous avons prouvé que $\|l_n\| \leqslant \frac{1}{2\pi}\int_{-\pi}^{\pi}|D_n(t)|dt$. Nous allons à présent prouver que cela est effectivement la norme de l_n. Pour $\epsilon > 0$ nous considérons la fonction

$$\begin{aligned} f_\epsilon\colon \mathbb{R} &\to \mathbb{C} \\ x &\mapsto \frac{D_n(x)}{|D_n(x)| + \epsilon}. \end{aligned} \tag{30.77}$$

C'est une fonction continue et 2π-périodique satisfaisant $\|f_\epsilon\| \leqslant 1$ parce que le dénominateur est toujours plus grand que le numérateur. Nous nous proposons de calculer

$$l_n(f_\epsilon) = \sum_{k=-n}^{n}\frac{1}{2\pi}\int_{-\pi}^{\pi} f_\epsilon(t)e^{-ikt}dt. \tag{30.78}$$

Vu que $f_\epsilon(t)e^{-ikt}$ vaut en norme $|f_\epsilon(t)|$ qui est une fonction intégrable (ne dépendant pas de k) sur $[-\pi,\pi]$, le théorème de la convergence dominée 16.140 nous permet de permuter la somme et l'intégrale :

$$l_n(f_\epsilon)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{D_n(t)}{|D_n(t)|+\epsilon}\underbrace{\sum_{k=-n}^{n}e^{-ikt}}_{=D_n(t)}dt=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{|D_n(t)|^2}{|D_n(t)|+\epsilon}dt. \tag{30.79}$$

Nous avons donc

$$\lim_{\epsilon\to0}l_n(f_\epsilon)=\frac{1}{2\pi}\int_{-\pi}^{\pi}|D_n(t)|dt. \tag{30.80}$$

Mais vue l'inégalité (30.74) nous avons

$$\|l_n\|=\frac{1}{2\pi}\int_{-\pi}^{\pi}|D_n(t)|dt. \tag{30.81}$$

Notre tâche est maintenant de donner une valeur à cette intégrale.

Norme de l_n tend vers ∞ D'abord nous écrivons

$$\|l_n\|=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\left|\sin\left(\frac{2n+1}{2}t\right)\right|}{\left|\sin(t/2)\right|}dt, \tag{30.82}$$

ensuite nous nous souvenons que $|\sin(x)|\leqslant|x|$ pour tout x, ce qui nous permet de changer le dénominateur :

$$\|l_n\|\geqslant\frac{2}{\pi}\int_0^{\pi}\frac{\left|\sin\left(\frac{2n+1}{2}t\right)\right|}{|t|}dt \tag{30.83}$$

Nous y effectuons le changement de variable $u=\frac{2n+1}{2}t$ qui donne

$$\|l_n\|\geqslant\frac{2}{\pi}\int_0^{(n+\frac{1}{2})\pi}\frac{\left|\sin(u)\right|}{|u|}. \tag{30.84}$$

Nous y reconnaissons l'intégrale (19.459) du sinus cardinal que nous savons diverger. Cela donne

$$\lim_{n\to\infty}\|l_n\|=\infty. \tag{30.85}$$

La conclusion L'espace $\left(C^0_{2\pi},\|.\|_\infty\right)$ est complet [4], donc le théorème de Banach-Steinhaus 28.7 s'applique. Par rapport aux notations de l'énoncé de Banch-Steinhaus, nous posons

$$E=\left(C^0_{2\pi},\|.\|_\infty\right) \tag{30.86a}$$
$$F=\mathbb{R} \tag{30.86b}$$
$$H=\{l_n\}_{n\in\mathbb{N}}. \tag{30.86c}$$

Vu que la suite $(\|l_n\|)$ n'est pas bornée, il existe $f\in C^0_{2\pi}$ tel que

$$\sup_n\|l_n(f)\|=\infty. \tag{30.87}$$

Pour cette fonction nous avons

$$\sup_{n\geqslant0}S_n(f)(0)=\infty, \tag{30.88}$$

et donc la série de Fourier de f ne converge pas en zéro.

□

4. Parce qu'une limite uniforme de fonctions continues est continue, théorème 14.314.

30.3.2 Inégalité isopérimétrique

Définition 30.19.
*Une **courbe de Jordan** dans le plan est une application $\gamma\colon S^1 \to \mathbb{R}^2$ qui est continue et injective.*

Une telle courbe peut évidemment être vue comme une application $\gamma\colon [0,2\pi] \to \mathbb{R}^2$ telle que $\gamma(0) = \gamma(2\pi)$. En particulier il n'est jamais mauvais de se rappeler qu'on peut choisir une paramétrisation normale par la proposition 23.44.

Théorème 30.20 (Théorème de Jordan[366])**.**
Si γ est une courbe de Jordan, alors l'ensemble $\mathbb{R}^2\backslash\gamma$ a exactement deux composantes connexes. L'une est bornée, l'autre non. Les deux ont γ comme frontière.

Le théorème suivant dit que parmi les courbes C^1, le cercle a la plus grande surface possible à périmètre donné.

Théorème 30.21 (Inégalité isopérimétrique[4])**.**
Soit $f\colon S^1 \to \mathbb{C}$ une courbe de Jordan de classe C^1. Nous notons L sa longueur et S l'aire contenue de la surface délimitée[5] *par f. Alors*

*(1) Nous avons l'**inégalité isopérimétrique** : $L^2 \geqslant 4\pi S$.*

(2) Nous avons l'égalité $L^2 = 4\pi S$ si et seulement si la courbe donnée par f est un cercle.

Démonstration. Nous commençons par considérer un chemin dont la longueur est 2π et nous en considérons sa paramétrisation normale. Nous allons exprimer l'aire S en utilisant le théorème de Green, et plus particulièrement la formule de surface (22.255).

Si $f(s) = x(s) + iy(s)$, nous devons intégrer $y'x - x'y$, qui n'est rien d'autre que la partie imaginaire de $f'(s)\overline{f(s)}$. Donc

$$S = \frac{1}{2}\operatorname{Im}\int_0^{2\pi} f'(s)\overline{f(s)}ds \tag{30.89}$$

Nous considérons les coefficients de Fourier de f donnés par la formule (30.50) :

$$c_n(f) = \frac{1}{2\pi}f(s)e^{-ins}. \tag{30.90}$$

Ceux de f' (qui est aussi continue sur le compact S^1 et donc tout autant L^2) sont donnés par

$$c_n(f') = inc_n(f). \tag{30.91}$$

D'autre part en vertu du théorème 23.9, la longueur de γ s'exprime en terme de l'intégrale de la norme de sa dérivée :

$$2\pi = L = \int_0^{2\pi} |f'(s)|ds = \int_0^{2\pi} |f'(s)|^2 ds \tag{30.92}$$

parce que nous avons choisi une paramétrisation normale qui vérifie automatiquement $|f'(s)| = 1$ pour tout s. L'identité de Parseval sous sa forme (27.104) appliquée à f' nous enseigne que

$$L = 2\pi = \int_0^{2\pi} |f'(s)|^2 ds = 2\pi\langle f', f'\rangle = 2\pi\sum_{n=-\infty}^{\infty} |c_n(f')|^2 = 2\pi\sum_n n^2|c_n(f)|^2. \tag{30.93}$$

Par ailleurs le système trigonométrique étant une base hilbertienne, et les fonctions f et f' étant dans $L^2\big([0,2\pi]\big)$ (parce que continues sur un compact), elles sont égales à leurs séries de Fourier

5. C'est la partie connexe bornée de $\mathbb{C}\backslash\gamma$ dont l'existence est donnée par le théorème de Jordan 30.20.

(au sens L^2), c'est à dire que nous avons l'égalité (28.205). Nous avons alors

$$\langle f', f\rangle_{L^2} = \langle \sum_{n\in\mathbb{Z}} c_n(f')e_n, \sum_{m\in\mathbb{Z}} c_m(f)e_m\rangle \tag{30.94a}$$
$$= \sum_m \sum_n c_n(f')\overline{c_m(f)} \underbrace{\langle e_n, e_m\rangle}_{\delta_{m,n}} \tag{30.94b}$$
$$= \sum_{n\in\mathbb{Z}} c_n(f')\overline{c_n(f)} \tag{30.94c}$$
$$= \sum_n in|c_n(f)|^2 \tag{30.94d}$$

où nous avons utilisé la continuité du produit scalaire pour sortir les sommes. Avec cela nous pouvons exprimer l'aire (30.89) en termes de coefficients de Fourier :

$$S = \frac{1}{2}\operatorname{Im} 2\pi\langle f', f\rangle = \pi \sum_{n\in\mathbb{Z}} n|c_n(f)|^2. \tag{30.95}$$

En utilisant les expressions (30.93) et (30.95) pour L et S, et en écrivant $L = 2\pi L$, nous avons

$$L^2 - 4\pi S = 4\pi^2 \sum_{n\in\mathbb{Z}} (n^2 - n)|c_n(f)|^2 \geqslant 0. \tag{30.96}$$

Cela prouve l'inégalité demandée dans le cas où $L = 2\pi$.

Si γ n'est pas de longueur 2π mais L, alors nous considérons le chemin $\sigma(t) = \frac{2\pi\gamma(t)}{L}$. Sa longueur est 2π et son aire, au vu de la formule de Green (30.89), son aire est $4\pi^2\frac{S}{L^2}$. L'inégalité isopérimétrique appliquée au chemin σ donne alors $L^2 \geqslant 4\pi S$.

Le cas d'égalité s'obtient uniquement si $c_n = 0$ pour tout n différent de 0 ou 1. Dans ce cas nous avons

$$f(s) = c_0(f) + c_1(f)e^{is}, \tag{30.97}$$

qui est un cercle de centre $c_0(f)$ et de rayon $|c_1(f)|$. □

30.3.3 À propos des coefficients

Nous considérons l'application

$$\begin{aligned} c\colon \big(L^1_{2\pi}, \|.\|_1\big) &\to \big(C_0, \|.\|_\infty\big) \\ f &\mapsto (c_n(f))_{n\in\mathbb{Z}} \end{aligned} \tag{30.98}$$

qui à une fonction 2π-périodique fait correspondre la suite (bornée) de ses coefficients de Fourier. Nous rappelons la définition

$$c_n(f) = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-int}. \tag{30.99}$$

Nous allons montrer que cette application est linéaire, continue, injective et non surjective. Pour la continuité, par la linéarité il suffit de la montrer en 0. Nous devons donc montrer que si nous avons une suite de fonctions f_k qui tend vers 0 au sens L^1, alors $c(f_k) \to 0$ au sens de la norme $\|.\|_\infty$ sur l'ensemble des suites.

Si nous posons $r_k = \int_0^{2\pi} |f_k(t)|dt$, alors $r_k = \|f_k\|_1$ et nous avons $r_k \to 0$. Mais par définition

$$|c_n(f_k)| \leqslant r_k, \tag{30.100}$$

et donc $\|c(f_k)\|_\infty \leqslant r_k$. L'application c est donc continue. L'injectivité est donnée par le corollaire 30.16.

Si nous supposons que l'application c est continue, alors le théorème d'isomorphisme de Banach (28.2) nous dit que cela devrait être un homéomorphisme, c'est à dire que c^{-1} serait également continue. Nous allons montrer qu'il n'en est rien.

Nous considérons la suite de suite

$$(c_n)_k = \begin{cases} 1 & \text{si } k < n \\ 0 & \text{sinon.} \end{cases} \tag{30.101}$$

Ici $(c_n)_k$ est le terme numéro k de la suite n. Par injectivité de l'application qui à une fonction fait correspondre la suite de ses coefficients de Fourier, la seule fonction qui possède ces coefficients est

$$f_n(t) = \sum_{k\in\mathbb{N}} c_{n,k} e^{ikt}. \tag{30.102}$$

Étant donné que $\|f_n\|_1 = n$, la suite $(\|f_n\|_1)$ n'est pas bornée alors que a suite de suites (30.101) est bornée dans l'ensemble des suites parce que $\|c_n\|_\infty = 1$.

Chapitre 31

Transformation de Fourier

Définition 31.1.
Soit une fonction f sur $\mathbb{R}^d$, dont nous ne précisons pas la régularité. Sa ***transformée de fourier*** *est la fonction $\hat{f}$ définie par*

$$\hat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-i\xi \cdot x} dx \tag{31.1}$$

si elle existe.

Ce qui est bien avec cette définition est que si la formule (31.1) ne définit pas $\hat{f}$ (parce que l'intégrale n'existe pas, par exemple), nous nous réservons le droit de définir tout de même $\hat{f}$ par d'autres biais. Ce sera d'ailleurs l'objet du théorème 31.28 qui définira $\hat{f}$ pour tout $f \in L^2(\mathbb{R}^d)$ alors que la formule (31.1) ne fonctionne pas sur toutes ces fonctions.

Une bonne partie de ce qui va suivre aura pour objet de déterminer des espaces de fonctions sur lesquels la transformée est bien définie, et sur lesquels elle a de bonnes propriétés.

Nous allons par ailleurs utiliser indifféremment les notations $\mathcal{F}(f)$ ou $\hat{f}$ pour la transformée de Fourier de f. La notation $\mathcal{F}$ est pratique pour les transformées de looooooongues expressions ainsi que pour parler de l'application « transformée de Fourier » d'un espace de fonction vers un autre.

31.2.
Nous verrons dans le théorème 31.28 que la transformée de Fourier n'est pas une isométrie de L^2. Pour avoir une isométrie, il aurait fallu choisir des coefficients moins simples.

31.1 Transformée de Fourier sur $L^1(\mathbb{R}^d)$

Nous rappelons que les espaces L^p sont des ensembles de classes de fonctions, définition 28.10. La transformée de Fourier, comme presque tout ce qui a trait aux intégrales, passe aux classes.

Lemme 31.3.
Soit une fonction $f : \mathbb{R}^d \to \mathbb{C}$ telle que $\mathcal{F}(f)$ existe. Alors pour toute fonction $g \in [f]$ la transformée $\mathcal{F}(g)$ existe et $\hat{g} = \hat{f}$.

Démonstration. Par définition des classes, il existe une fonction $s : \mathbb{R}^d \to \mathbb{C}$ presque partout nulle telle que $g = f + s$. Soit $\xi \in \mathbb{R}^d$ fixé. La fonction $x \mapsto s(x)e^{-i\xi x}$ est presque partout nulle et donc intégrable d'intégrale null. La proposition 16.129 nous permet alors d'affirmer que $f + s$ est intégrable et que

$$\mathcal{F}(f+s)(\xi) = \int_{\mathbb{R}^d} (f+s)(x)e^{-i\xi x}dx = \int_{\mathbb{R}^d} f(x)e^{-i\xi x}dx + \int_{\mathbb{R}^d} s(x)e^{-i\xi x}dx = \mathcal{F}(f)(x). \tag{31.2}$$

□

À partir de maintenant, lorsque nous parlons de transformée de Fourier d'une fonction dans L^p, nous parlons indifféremment d'une vraie fonction ou d'une classe.

Lemme 31.4.
Si $f \in L^1(\mathbb{R}^d)$, alors $\hat{f}$ existe.

Démonstration. Par définition, si $f \in L^1(\mathbb{R}^d)$, alors l'intégrale $\int_{\mathbb{R}^d} |f|$ existe et est finie. Alors la fonction qui arrive dans la transformée de Fourier en ξ, la fonction $s\colon x \to f(x)e^{-i\xi x}$ est également dans $L^1(\mathbb{R}^d)$ parce que $|s| = |f|$. □

<++>

Lemme 31.5.
Si $f \in L^1(\mathbb{R}^d)$ et si $g(x) = f(\lambda x)$ alors

$$\hat{g}(\xi) = \lambda^{-d}\hat{f}(\xi/\lambda). \tag{31.3}$$

Démonstration. Il s'agit de faire le changement de variable $y = \lambda x$ dans l'intégrale

$$\hat{g}(\xi) = \int_{\mathbb{R}^d} f(\lambda x)e^{-i\xi x}dx. \tag{31.4}$$

Dans le changement de variables, vient le coefficient $dx = \lambda^{-d}dy$. □

Proposition 31.6.
La transformée de Fourier est un morphisme vis-à-vis de la convolution sur $L^1(\mathbb{R}^n)$:

$$\widehat{f * g} = \hat{f}\hat{g}. \tag{31.5}$$

Démonstration. Nous devons étudier l'intégrale

$$\widehat{f * g}(\xi) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} f(y)g(t-y) \right] e^{-it\xi}dt. \tag{31.6}$$

Ici nous avons choisi des représentants f et g dans les classes de L^1. Montrons que f est borélienne. D'abord $f(x) = f_+(x) - f_-(x)$ où f_+ et f_- sont des fonctions positives. Afin d'alléger les notations nous supposons un instant que f est positive et nous posons

$$f_n(x) = \sum_{k=1}^{2^n} \frac{k}{n} \mathbb{1}_{f(x)\in[\frac{k}{n},\frac{k+1}{n}[}. \tag{31.7}$$

Le fait que f soit dans L^1 implique que chacune des fonctions f_n est borélienne [1] et donc que f l'est aussi en tant que limite ponctuelle de fonctions boréliennes [2].

Nous allons appliquer le théorème de Fubini 16.238 à la fonction

$$\phi(x, y) = f(x)g(y)e^{-i\xi(x+y)} \tag{31.8}$$

qui est borélienne en tant que produit et composé de fonctions boréliennes. Nous avons

$$\int_{\mathbb{R}} \left(\int_{\mathbb{R}} |f(x)e^{-i\xi x}||g(y)e^{-i\xi y}|dy \right) dx = \int_{\mathbb{R}} \left(|f(x)| \int_{\mathbb{R}} |g(y)|dy \right) dx \tag{31.9a}$$

$$= \int_{\mathbb{R}} |f(x)|\|g\|_1 \tag{31.9b}$$

$$= \|f\|_1\|g\|_1 < \infty. \tag{31.9c}$$

Le théorème est donc applicable. D'abord nous avons :

$$\hat{f}(\xi)\hat{g}(\xi) = \left(\int_{\mathbb{R}} f(x)e^{-i\xi x}dx \right) \left(\int_{\mathbb{R}} g(y)e^{-i\xi y}dy \right) \tag{31.10a}$$

$$= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(x)g(y)e^{-i\xi(x+y)}dy \right) dx \tag{31.10b}$$

$$= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(x)g(t-x)e^{-i\xi t} \right) dx. \tag{31.10c}$$

1. Ceci demanderait plus de justification. Dites moi si vous savez comment justifier que les f_n soient boréliennes.
2. Le fait que f soit borélienne est une conséquence du théorème 28.96.

Jusqu'ici nous n'avons pas utilisé Fubini. Nous avons seulement introduit le nombre $\int_{\mathbb{R}} g(y)e^{-i\xi y}dy$ dans l'intégrale par rapport à x et effectué le changement de variables $y \mapsto t = x + y$. Maintenant nous appliquons le théorème de Fubini pour inverser l'ordre des intégrales :

$$\hat{f}(\xi)\hat{g}(\xi) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(x)g(t-x)e^{-it\xi}dx \right) dy \tag{31.11a}$$

$$= \int_{\mathbb{R}} e^{-it\xi} \left(\int_{\mathbb{R}} f(x)g(t-x)dx \right) dt \tag{31.11b}$$

$$= \int_{\mathbb{R}} e^{-it\xi}(f * g)(t)dt \tag{31.11c}$$

$$= \widehat{f * g}(\xi). \tag{31.11d}$$

□

Proposition 31.7.
Soit une fonction $f \in L^1(\mathbb{R}^d)$. Alors sa transformée de Fourier est continue.

Démonstration. Nous considérons une fonction f définie sur $\mathbb{R}^d$ et à valeurs dans $\mathbb{R}$ ou $\mathbb{C}$. Sa transformée de Fourier est donnée par

$$\hat{f}(\xi) = \int_{\mathbb{R}^d} e^{-i\xi x} f(x)dx. \tag{31.12}$$

Pour montrer que cette fonction $\hat{f}$ est continue en ξ_0 nous considérons une suite $(\xi_n) \to \xi_0$ et nous voulons montrer que $\hat{f}(\xi_n) \to \hat{f}(\xi_0)$. Pour cela nous considérons les fonctions

$$g_n(x) = e^{-i\xi_n x} f(x) \tag{31.13}$$

qui convergent simplement vers $g(x) = e^{-i\xi x} f(x)$. Étant donné que

$$|g_n(x)| < |f(x)|, \tag{31.14}$$

le théorème de la convergence dominée donne alors

$$\lim_{n\to\infty} \int g_n(x) = \int \lim_{n\to\infty} g_n(x), \tag{31.15}$$

c'est à dire $\lim_{n\to\infty} \hat{f}(\xi_n) = \hat{f}(\xi)$. La fonction $\hat{f}$ est donc continue. □

Lemme 31.8.
Pour tout $f \in L^1(\mathbb{R}^n)$ nous avons $\|\hat{f}\|_\infty \leqslant \|f\|_1$.

Démonstration. Cela est une simple vérification :

$$\hat{f}(\xi) = \int_{\mathbb{R}^n} f(x)e^{-ix\xi}dx, \tag{31.16}$$

nous avons, pour tout ξ,

$$|\hat{f}(\xi)| \leqslant \int_{\mathbb{R}} |f(x)|dx, \tag{31.17}$$

ce qui signifie exactement $\|\hat{f}\|_\infty \leqslant \|f\|_1$. □

Lemme 31.9 (Lemme de Riemann-Lebesgue[367])**.**
Si f est une fonction $L^1(\mathbb{R})$ alors $\lim_{\xi\to\pm\infty} \hat{f}(\xi) = 0$.

Démonstration. Nous commençons par prouver le résultat dans le cas d'une fonction g en escalier, et plus précisément par une fonction caractéristique d'un compact $K = [a, b]$. Au niveau de la transformée de Fourier nous avons

$$\hat{\mathbb{1}}_K(\xi) = \int_a^b e^{-i\xi x} dx = -\frac{1}{i\xi}(e^{-ib\xi} - e^{-ia\xi}). \tag{31.18}$$

Par conséquent

$$|\hat{\mathbb{1}}_K(\xi)| \leqslant \frac{2}{|\xi|}. \tag{31.19}$$

Plus généralement si $g = \sum_{i=1}^N c_i \mathbb{1}_{K_i}$, alors

$$|\hat{g}(\xi)| \leqslant \frac{2}{|\xi|} \sum_{i=1}^N |c_i|, \tag{31.20}$$

et donc nous avons effectivement $\lim_{\xi\to\pm\infty} |\hat{g}(\xi)| = 0$.

Nous passons maintenant au cas général $f \in L^1(\mathbb{R})$. Étant donné que les fonctions L^1 en escalier sont denses dans L^1, nous considérons une fonction $g \in L^1(\mathbb{R})$ en escalier telle que $\|f - g\|_1 < \epsilon$. Nous avons donc

$$\|\hat{f} - \hat{g}\|_\infty \leqslant \|f - g\|_1 < \epsilon. \tag{31.21}$$

Donc

$$\|\hat{f}(\xi)\| \leqslant \|\hat{f}(\xi) - \hat{g}(\xi)\|_| \hat{g}(\xi)|. \tag{31.22}$$

Le premier terme est plus petit que ϵ. Il nous reste à voir que

$$\lim_{\xi\to\infty} |\hat{g}(\xi)| = 0, \tag{31.23}$$

mais cela est le résultat de la première partie de la preuve. □

Corollaire 31.10.
La transformée de Fourier d'une fonction $L^1(\mathbb{R})$ est bornée.

Démonstration. Par le corollaire 31.7, la transformée de Fourier d'une fonction L^1 est continue. Le lemme de Riemann-Lebesgue 31.9 impliquant qu'elle tend vers zéro en $\pm\infty$, elle doit être bornée. □

31.1.1 Formule sommatoire de Poisson

Proposition 31.11 (Formule sommatoire de Poisson)**.**
Soit $f : \mathbb{R} \to \mathbb{C}$ une fonction continue et $L^1(\mathbb{R})$. Nous supposons que

(1) il existe $M > 0$ et $\alpha > 1$ tels que

$$|f(x)| \leqslant \frac{M}{(1 + |x|)^\alpha}, \tag{31.24}$$

(2) $\sum_{n=-\infty}^{\infty} |\hat{f}(2\pi n)| < \infty$.

Alors nous avons

$$\sum_{n=-\infty}^{\infty} f(n) = \sum_{n=-\infty}^{\infty} \hat{f}(2\pi n). \tag{31.25}$$

Démonstration. **Convergence normale** Nous commençons par montrer qu'il y a convergence normale sur tout compact séparément des séries sur les $n \geqslant 0$ et sur les $n < 0$.

Soit K un compact de $\mathbb{R}$ contenu dans $[-A, A]$ et $n \in \mathbb{Z}$ tel que $|n| \geqslant 2A$. Pour $x \in K$ nous avons

$$|x + n| \geqslant |n| - |x| \geqslant |n| - A \geqslant \frac{|n|}{2}. \tag{31.26}$$

Du coup nous avons un $\alpha > 1$ tel que

$$|f(x+n)| \leqslant \frac{M}{\big(1+|x+n|\big)^{\alpha}} \leqslant \frac{M}{\Big(1+\frac{|n|}{2}\Big)^{\alpha}}. \tag{31.27}$$

Lorsque n est grand, cela a le comportement de $M/|n|^{\alpha}$ et donc la série

$$\sum_{n=0}^{\infty} f(x+n) \tag{31.28}$$

est une série convergent normalement. Les deux séries (usuelles)

$$a_- = \sum_{n\leqslant 0} f(x+n) \tag{31.29a}$$

$$a_- = \sum_{n>0} f(x+n) \tag{31.29b}$$

convergent normalement.

Convergence commutative Au sens de la définition 11.191 nous avons

$$\sum_{n\in\mathbb{Z}} f(x+n) = a_+ + a_-. \tag{31.30}$$

En effet si nous prenons $J_0' \subset \mathbb{N}$ fini tel que $|\sum_{\mathbb{N}\setminus J_0} f(x+n) - a_+| \leqslant \epsilon$ et $J_1' \in -\mathbb{N}$ tel que $|\sum_{n\in-\mathbb{N}\setminus J_1'} f(x+n)| - a_- < \epsilon$, et si nous posons $J_0 = J_0' \cup J_1'$ alors si K est un ensemble fini de $\mathbb{Z}$ contenant J_0 nous avons

$$|\sum_{n\in K} f(n+x) - (a_+ + a_-)| \leqslant |\sum_{n\in K^+} f(n+x) - a_+| + |\sum_{n\in K^-} f(n+x) - a_-| \leqslant 2\epsilon \tag{31.31}$$

où K^+ sont les éléments positifs de K et K^- sont les *strictement* négatifs. Maintenant que la famille $\{f(n+x)\}_{n\in\mathbb{Z}}$ est une famille sommable, nous savons qu'elle est commutativement sommable et que la proposition 11.196 nous permet de sommer dans l'ordre que l'on veut. Nous pouvons donc écrire sans ambigüité l'expression $\sum_{n\in\mathbb{Z}} f(x+n)$ ou $\sum_{n=-\infty}^{\infty} f(x+n)$.

re-convergence normale Nous posons donc sans complexes la série

$$F(x) = \sum_{n\in\mathbb{Z}} f(x+n) \tag{31.32}$$

qui converge tant commutativement que normalement. Notons que nous pouvons maintenant dire que la série sur $\mathbb{Z}$ converge normalement ; pas seulement les deux séries séparément.

Continuité, périodicité Étant donné que chacune des fonctions $f(x+n)$ est continue, la convergence normale nous assure que F est continue.

De plus F est périodique parce que

$$F(x+1) = \sum_{n=-\infty}^{\infty} f(x+1+n) = \sum_{p=-\infty}^{\infty} f(x+p) \tag{31.33}$$

où nous avons posé $p = 1+n$.

Coefficients de Fourier En vertu de la définition (28.168) et de la périodicité de F,

$$\begin{aligned}
c_n(F) &= \int_{-1/2}^{1/2} F(t)e^{-2\pi int}dt && (31.34a)\\
&= \int_0^1 F(t)e^{-2\pi int}dt && (31.34b)\\
&= \int_0^1 \sum_{n\in\mathbb{Z}} f(t+n)e^{-2i\pi nt}dt && (31.34c)\\
&= \sum_{n\in\mathbb{Z}} \int_n^{n+1} f(u)e^{-2\pi i(u-n)t}du && (31.34d)\\
&= \int_{-\infty}^{\infty} f(u)e^{-2\pi inu}du && (31.34e)\\
&= \hat{f}(2\pi n). && (31.34f)
\end{aligned}$$

où nous avons effectué le changement de variables $u = t + n$, et permuté l'intégrale et la somme en vertu du fait que la somme converge normalement.

Conclusion Étant donné l'hypothèse $\sum_{n\in\mathbb{Z}} |\hat{f}(n)| < \infty$ la proposition 30.14 nous dit que

$$F(x) = \sum_{n\in\mathbb{Z}} c_n(F)e^{2\pi inx}, \tag{31.35}$$

c'est à dire que

$$\sum_{n-\infty}^{\infty} f(x+n) = \sum_{n=-\infty}^{\infty} \hat{f}(2\pi n)e^{2\pi inx}. \tag{31.36}$$

En écrivant cette égalité en $x = 0$ nous trouvons le résultat :

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \hat{f}(2\pi n). \tag{31.37}$$

□

Exemple 31.12

La formule sommatoire de Poisson peut être utilisée pour calculer des sommes dans l'espace de Fourier plutôt que dans l'espace direct. Nous allons montrer dans cet exemple l'égalité

$$\sum_{n=-\infty}^{\infty} e^{-\alpha n^2} = \sum_{n=-\infty}^{\infty} \sqrt{\frac{\pi}{\alpha}} e^{-\pi^2 n^2/\alpha}. \tag{31.38}$$

Si α est grand, alors la somme de gauche est plus rapide, tandis que si α est petit, c'est le contraire.

Nous appliquons la formule sommatoire de Poisson à la fonction

$$f(x) = e^{-\alpha x^2}. \tag{31.39}$$

Nous avons

$$\begin{aligned}
\hat{f}(x) &= \int_{\mathbb{R}} e^{-\alpha t^2 - ixt}dt && (31.40a)\\
&= e^{-x^2/4\alpha} \int_{\mathbb{R}} e^{-(\sqrt{\alpha}t + \frac{ix}{2\sqrt{\alpha}})^2} && (31.40b)\\
&= e^{-x^2/4\alpha} \frac{1}{\sqrt{\alpha}} \int_{\mathbb{R}+\frac{ix}{2\sqrt{\alpha}}} e^{-u^2}du. && (31.40c)
\end{aligned}$$

Pour traiter cette intégrale nous utilisons la proposition 29.6 en considérant le chemin rectangulaire fermé qui joint les points $-R$, R, $R+ai$, $-R+ai$ et $f(z)=e^{-z^2}$. Calculons l'intégrale sur les deux côtés verticaux. Nous posons

$$\gamma_R(t)=R+tia \tag{31.41}$$

avec $t\colon 0\to 1$. Nous avons

$$\int_{\gamma_R} f=\int_0^1 f\big(\gamma_R(t)\big)\|\gamma_R'(t)\|dt \tag{31.42a}$$

$$=ae^{-R^2}\int_0^1 e^{-2tRia+at^2}dt, \tag{31.42b}$$

donc en module nous avons

$$|\int_{\gamma_R} f|\leqslant ae^{-R^2}\int_0^1 e^{at^2}dt\leqslant Me^{-R^2}, \tag{31.43}$$

où M est une constante ne dépendant pas de R. Lorsque $R\to\infty$, la contribution des chemins verticaux s'annule et nous trouvons que

$$\int_{\mathbb{R}+ai} e^{-u^2}du=\int_{\mathbb{R}} e^{-u^2}du, \tag{31.44}$$

que nous pouvons utiliser pour continuer le calcul (31.40). Nous avons

$$\hat{f}(x)=\frac{e^{-x^2/4\alpha}}{\sqrt{\alpha}}\int_R e^{-u^2}du=\sqrt{\frac{\pi}{\alpha}}e^{-x^2/4\alpha} \tag{31.45}$$

où nous avons utilisé la formule (16.732). Par conséquent ce qui rentre dans la formule sommatoire de Poisson est

$$\hat{f}(2\pi n)=\sqrt{\frac{\pi}{\alpha}}e^{-\pi^2n^2/\alpha}. \tag{31.46}$$

△

31.2 Transformée de Fourier dans l'espace de Schwartz

La définition de la transformée de Fourier de $\varphi\in\mathscr{S}(\mathbb{R}^d)$ est

$$\hat{\varphi}(\xi)=\int_{\mathbb{R}^n}\varphi(x)e^{-ix\cdot\xi}dx. \tag{31.47}$$

Si α est un multiindice de taille m, nous notons

$$(M_\alpha f)(x)=x_{\alpha_1}\dots x_{\alpha_m}f(x). \tag{31.48}$$

Lemme 31.13 (Lemme de transfert)**.**
Si $\varphi\in\mathscr{S}(\mathbb{R}^d)$ et si α est un multiindice, alors

$$\partial^\alpha\hat{\varphi}=(-i)^{|\alpha|}\widehat{M_\alpha\varphi}. \tag{31.49}$$

et

$$\widehat{\partial^\alpha\varphi}(\xi)=(-i)^{|\alpha|}\xi^\alpha\hat{\varphi}(\xi). \tag{31.50}$$

Démonstration. Nous considérons la fonction $h(x,\xi)=\varphi(x)e^{-ix\cdot\xi}$ dont la dérivée par rapport à ξ_i est donnée par $-i(M_i\varphi)(x)e^{x\cdot\xi}$. Cette fonction est majorée en norme par

$$G(x)=M_i\varphi(x), \tag{31.51}$$

qui est encore une fonction à décroissance rapide et donc parfaitement intégrable sur $\mathbb{R}^d$. Le théorème 19.27 nous dit donc que la dérivée de $\hat{\varphi}$ par rapport à ξ_i existe et vaut

$$\frac{\partial\hat{\varphi}}{\partial\xi_i}(\xi) = -i\int_{\mathbb{R}^n} x_i\varphi(x)e^{-i\xi\cdot x} = -i\widehat{M_i\varphi}(\xi). \tag{31.52}$$

En appliquant ce résultat en chaîne, nous trouvons la première formule annoncée.

Nous passons à la seconde formule annoncée. Étant donné que $\varphi\in\mathscr{S}$, ses dérivées le sont aussi et par conséquent, il n'y a pas de problèmes pour écrire

$$\widehat{\partial_{x_k}\varphi}(\xi) = \int_{\mathbb{R}^d}\frac{\partial\varphi}{\partial x_k}(x)e^{-ix\cdot\xi}dx. \tag{31.53}$$

Étant donné que

$$\frac{\partial}{\partial x_k}\Big(\varphi(x)e^{-ix\cdot\xi}\Big) = \frac{\partial\varphi}{\partial x_k}(x)e^{-ix\cdot\xi} - i\xi_k\varphi(x)e^{-ix\cdot\xi}, \tag{31.54}$$

notre tâche sera de prouver que

$$\int_{\mathbb{R}^d}\frac{\partial}{\partial x_k}\Big(\varphi(x)e^{-ix\cdot\xi}\Big)\,dx = 0. \tag{31.55}$$

Autrement dit, nous voulons montrer que le terme au bord d'une intégration par partie s'annule. D'abord le fait que φ soit à décroissance rapide nous assure que l'intégrale (31.55) converge. Pour chaque ξ, la fonction

$$f(x,\xi) = \frac{\partial}{\partial x_k}\Big(\varphi(x)e^{-ix\cdot\xi}\Big) \tag{31.56}$$

est intégrable par rapport à x. De plus, f est dans $\mathscr{S}(\mathbb{R})$ pour chacune de ses variables (les autres étant fixées). Le théorème de Fubini 16.239 nous permet alors de décomposer l'intégrale en

$$\int_{\mathbb{R}^d} f(x,\xi)dx = \int_{\mathbb{R}}\dots\int_{\mathbb{R}} f(x_1,\dots,x_d)dx_1\dots dx_d. \tag{31.57}$$

De plus nous pouvons intégrer dans l'ordre de notre choix et nous choisissons évidemment d'intégrer d'abord par rapport à x_k. Étudions donc l'intégrale

$$\int_{\mathbb{R}}\frac{\partial}{\partial x}\Big(\varphi(x)e^{-ix\xi}\Big)\,dx = \lim_{A\to\infty}\int_{-A}^{A}\frac{\partial}{\partial x}\Big(\varphi(x)e^{-ix\xi}\Big)\,dx \tag{31.58}$$

dans laquelle nous avons un peu allégé les notations. Une primitive de ce qui est intégré est toute trouvée : c'est $\varphi(x)e^{-ix\xi}$, et nous pouvons utiliser le théorème fondamental du calcul intégral pour écrire que

$$\int_{-A}^{A}\Big(\varphi(x)e^{-ix\xi}\Big)'\,dx = \Big[\varphi(x)e^{-ix\xi}\Big]_{x=-A}^{x=A}. \tag{31.59}$$

Vu que φ est dans $\mathscr{S}$, la limite $A\to\infty$ donne zéro.

En substituant maintenant (31.54) dans (31.53) et en tenant compte du terme que nous venons de montrer s'annuler, nous avons

$$\widehat{\partial_k\varphi}(\xi) = -i\xi_k\int_{\mathbb{R}^d}\varphi(x)e^{-ix\cdot\xi} = -i\xi_k\hat{\varphi}(\xi). \tag{31.60}$$

En recommençant la procédure $|\alpha|$ fois nous trouvons la seconde formule annoncée. □

Proposition 31.14 ([216]).
L'espace de Schwartz est stable par transformée de Fourier. De plus l'application

$$\mathcal{F}\colon\mathscr{S}(\mathbb{R}^d)\to\mathscr{S}(\mathbb{R}^d) \tag{31.61}$$

est une bijection linéaire et continue.

Démonstration. La linéarité découle de celle de l'intégrale. La difficulté est de prouver que pour $\varphi \in \mathscr{S}(\mathbb{R}^d)$ nous avons bien que $\hat{\varphi} \in \mathscr{S}(\mathbb{R}^d)$ et que cette association est continue[3].

Stabilité Nous devons prouver que pour tout multiindices α et β, nous avons $p_{\alpha,\beta}(\hat{\varphi}) < \infty$. Nous avons

$$\xi^\beta \partial^\alpha \hat{\varphi}(\xi) = \xi^\beta (-i)^{|\alpha|} \widehat{M_\alpha \varphi}(\xi) = (-i)^{|\alpha|+|\beta|} \widehat{\partial^\beta M_\alpha \varphi}(\xi). \tag{31.62}$$

Ensuite nous nous souvenons que $\|\hat{f}\|_\infty \leqslant \|f\|_1$ parce que

$$|\hat{f}(\xi)| \leqslant \int_{\mathbb{R}^d} \left| f(x) e^{-ix \cdot \xi} \right| = \int_{\mathbb{R}^d} |f(x)| dx = \|f\|_1. \tag{31.63}$$

Donc

$$p_{\alpha,\beta}(\hat{\varphi}) = \|\widehat{\partial^\beta M_\alpha \varphi}\|_\infty \leqslant \|\partial^\beta M_\alpha \varphi\|_1. \tag{31.64}$$

Du fait que φ soit dans $\mathscr{S}$, la dernière expression est finie. Cela prouve déjà que

$$\mathcal{F}\big(\mathscr{S}(\mathbb{R}^d)\big) \subset \mathscr{S}(\mathbb{R}^d). \tag{31.65}$$

Continuité Nous supposons avoir une suite $\varphi_n \xrightarrow{\mathscr{S}} \varphi$, et nous devons prouver que $\hat{\varphi}_n \xrightarrow{\mathscr{S}} \hat{\varphi}$. Pour alléger les notations, nous posons $f_n = \varphi_n - \varphi$. Nous avons

$$\begin{aligned}
\|\hat{f}\|_{\alpha,\beta} &= \|\xi^\beta \partial^\alpha \hat{f}\|_\infty && \text{(31.66a)}\\
&= \|\widehat{\partial^\beta M_\alpha f}\|_\infty \text{ lemme 31.13.} && \text{(31.66b)}\\
&\leqslant \|\partial^\beta M_\alpha f\|_1 && \text{(31.66c)}
\end{aligned}$$

La convergence $f_n \xrightarrow{\mathscr{S}} 0$ nous dit ente autres que $\partial^\beta M_\alpha f_n \xrightarrow{\mathscr{S}} 0$; en particulier la proposition 28.95 nous dit que $\partial^\beta M_\alpha f_n \xrightarrow{L^1} 0$, ce qui signifie, par les majorations (31.66) que

$$\|\hat{f}_n\|_{\alpha,\beta} \leqslant \|\partial^\beta M_\alpha f_n\|_1 \to 0, \tag{31.67}$$

ce qui prouve la continuité de transformée de Fourier dans $\mathscr{S}(\mathbb{R}^d)$.

Bijection Une preuve peut être trouvée dans [368].

□

Proposition 31.15 ([1])**.**
Soit $\varphi \in \mathscr{S}(\mathbb{R}^n \times \mathbb{R}^m)$ et la transformée de Fourier partielle

$$\tilde{\varphi}(x,k) = \int_{\mathbb{R}^m} e^{-iky} \varphi(x,y) dy. \tag{31.68}$$

Alors $\tilde{\varphi} \in \mathscr{S}(\mathbb{R}^n \times \mathbb{R}^m)$.

Démonstration. Il s'agit de reprendre les étapes de la partie correspondante de la preuve de la proposition 31.14. Soient des multiindices α, α', β et β' où α et β se réfèrent à la variable x tandis que α' et β' se réfèrent à la variable k.

Vu que la multiplication par $k^{\beta'}$ commute avec ∂^α nous avons

$$x^\beta k^{\beta'} \partial^\alpha \partial^{\alpha'} \tilde{\varphi}(x,k) = x^\beta k^{\beta'} \partial^\alpha (-i)^{|\alpha'|} \widetilde{M_{\alpha'} \varphi}(x,k) = (-i)^{|\alpha'|+|\beta'|} x^\beta \partial^\alpha \widetilde{\partial^{\beta'} M_{\alpha'} \varphi}(x,k). \tag{31.69}$$

D'autre part nous avons $\partial^\alpha \tilde{\varphi} = \widetilde{\partial^\alpha \varphi}$ parce que la fonction $\partial_x \varphi$ étant Schwartz, la fonction

$$G(y) = \sup_{x \in \mathbb{R}^n} |(\partial_x \varphi)(x,y)| \tag{31.70}$$

est dans $L^1(\mathbb{R}^m)$ par le corollaire 28.91. Par conséquent le théorème 19.27 permet de permuter la dérivée et l'intégrale dans

$$\frac{\partial}{\partial x} \tilde{\varphi}(x,k) = \frac{\partial}{\partial x} \int_{\mathbb{R}^m} e^{-iky} \varphi(x,y) dy. \tag{31.71}$$

3. Pour rappel, en dimension infinie, il n'est pas garanti qu'une application linéaire soit continue.

Dans le même ordre d'esprit mais dans difficultés de permutation de limites nous avons $M_\beta\tilde{\varphi} = \widetilde{M_\beta\varphi}$.

D'autre part nous avons encore $\|\tilde{\varphi}\|_\alpha < \infty$ parce que

$$|\tilde{\varphi}(x,k)| \leqslant \int_{\mathbb{R}^m} |\varphi(x,y)|dy \leqslant \sup_x \int_{\mathbb{R}^m} |\varphi(x,y)|dy \leqslant \int_{\mathbb{R}^m} |\sup_x \varphi(x,y)|dy < \infty \tag{31.72}$$

parce que φ est Schwartz et le corollaire 28.91 donne l'intégrabilité.

Donc nous avons

$$p_{(\alpha\alpha'),(\beta\beta')(\tilde{\varphi})} = \|\partial^{\beta'}\widetilde{M_{\alpha'}M_\beta}\partial^\alpha\varphi\|_\infty < \infty. \tag{31.73}$$

Cela prouve que $\tilde{\varphi}$ est Schwartz. □

31.2.1 Quelques transformées de Fourier

Exemple 31.16([4])
Soit la fonction $g_\epsilon(x) = e^{-\epsilon x^2}$. Sa transformée de Fourier sera déduite dans le lemma 31.17 en utilisant le lemme de transfert 31.13. Nous nous proposons ici de déduire de façon directe l'équation différentielle vérifiée par la transformée de Fourier de g_ϵ.

Nous posons

$$I(k) = \int_{\mathbb{R}} e^{-ikx}e^{-\epsilon x^2}dx. \tag{31.74}$$

et nous considérons la fonction

$$f(k,x) = e^{-ikx}e^{-\epsilon x^2}. \tag{31.75}$$

Elle est de classe C^1 par rapport à k, et intégrable en x pour chaque k. De plus sa dérivée

$$(\partial_k f)(k,x) = -ixe^{-ikx}e^{-\epsilon x^2} \tag{31.76}$$

vérifie $|\partial_k f| \leqslant xe^{-\epsilon x^2}$. La dérivée est donc majorée (uniformément en k) par une fonction intégrable. Le théorème 19.27 permet de permuter la dérivée et l'intégrale :

$$\begin{aligned}
I'(k) &= \int_{\mathbb{R}} -ixe^{-ikx}e^{-\epsilon x^2}dx && (31.77a)\\
&= i\int_{\mathbb{R}} e^{-ikx}\frac{1}{2\epsilon}\frac{d}{dx}\left(e^{-\epsilon x^2}\right)dx && (31.77b)\\
&= \frac{-i}{2\epsilon}\int_{\mathbb{R}} \frac{d}{dx}\left(e^{-ikx}\right)e^{-\epsilon x^2}dx \qquad \text{par partie} && (31.77c)\\
&= \frac{-k}{2\epsilon}\int_{\mathbb{R}} e^{-ikx}e^{-\epsilon x^2}dx && (31.77d)\\
&= \frac{-k}{2\epsilon}I(k). && (31.77e)
\end{aligned}$$

D'où l'équation différentielle $I'(k) = -\frac{k}{2\epsilon}I(k)$. △

Lemme 31.17 (Transformée de Fourier de la Gausienne [369])**.**
La transformée de Fourier de

$$\begin{aligned} g_\epsilon\colon \mathbb{R}^d &\to \mathbb{R}\\ x &\mapsto e^{-\epsilon\|x\|^2}\end{aligned} \tag{31.78}$$

est donnée par

$$\hat{g}_\epsilon(\xi) = \left(\frac{\pi}{\epsilon}\right)^{d/2} e^{-\|\xi\|^2/4\epsilon} \tag{31.79}$$

Démonstration. Nous commençons par la fonction $g(x) = e^{-\|x\|^2/2}$ et nous prouvons que sa transformée de Fourier est $\hat{g}(\xi) = (2\pi)^{d/2}g(\xi)$.

Réduction à la dimension 1 La fonction g est dans l'espace de Schwartz. Par le théorème de Fubini,

$$\hat{g}(\xi) = \int_{\mathbb{R}^d} \prod_{k=1}^{d} e^{-x_k^2} e^{-i\xi_k x_k} dx = \prod_{k=1}^{d} \int_{\mathbb{R}} e^{-t^2/2} e^{-\xi_k x} dt = \prod_{k=1}^{d} \hat{f}(\xi_k) \tag{31.80}$$

où f est la fonction d'une variable

$$f(x) = e^{-x^2/2}. \tag{31.81}$$

Notons que $f \in \mathscr{D}(\mathbb{R})$.

Une équation différentielle Voyons l'équation différentielle satisfaite par la transformée de Fourier $\hat{f}$ de la fonction (31.81). Grâce au lemme 31.13 nous trouvons l'équation différentielle [4]

$$\xi \hat{f}(\xi) + (\hat{f})'(\xi) = 0. \tag{31.82}$$

C'est le moment d'utiliser le théorème de Cauchy-Lipschitz (19.49), appliqué à la fonction $f(t, y) = -ty$ qui est Lipschitz et continue au au problème

$$\begin{cases} y' + ty = 0 & \text{(31.83a)} \\ y(0) = y_0 & \text{(31.83b)} \end{cases}$$

possède une unique solution maximale, en l'occurrence $y(x) = y_0 e^{-x^2/2}$. En ce qui concerne la condition initiale nous avons

$$\hat{f}(0) = \int_{\mathbb{R}} e^{-x^2/2} dx = \sqrt{2\pi}. \tag{31.84}$$

par l'exemple 16.242. Donc

$$\hat{f}(\xi) = \sqrt{2\pi} e^{-\xi^2/2}. \tag{31.85}$$

En reformant le produit (31.80) nous concluons.

Nous passons maintenant à la fonction g_ϵ. Nous pouvons écrire g_ϵ sous la forme

$$g_\epsilon(x) = g(\sqrt{2\epsilon}x). \tag{31.86}$$

Utilisant successivement la transformée de Fourier de g que nous venons de calculer et 31.5 (facteur d'échelle) nous trouvons

$$\begin{aligned} \hat{g}(\xi) &= (2\pi)^{d/2} g(\xi) && \text{(31.87a)} \\ \hat{g}_\epsilon(\xi) &= (2\epsilon)^{-d/2} \hat{g}\big(\xi/\sqrt{2\epsilon}\big) && \text{(31.87b)} \\ &= \left(\frac{\pi}{\epsilon}\right)^{d/2} e^{-|\xi|^2/4\epsilon}.. && \text{(31.87c)} \end{aligned}$$

Nous voyons que $\hat{g}_\epsilon \in \mathscr{S}(\mathbb{R}^d)$ (c'était gagné d'avance par la proposition 31.14). □

31.3 Suite régularisante

Définition 31.18.
*Une **suite régularisante** est une suite (ρ_n) dans $L^1(\mathbb{R}^d)$ telle que*

(1) pour tout n, $\rho_n \geqslant 0$ et $\int_{\mathbb{R}^d} \rho_n = 1$;

(2) pour tout $\alpha > 0$,

$$\lim_{n\to\infty} \int_{|t|>\alpha} \rho_n = 0. \tag{31.88}$$

4. Une façon directe de déduire cette équation différentielle est donnée dans l'exemple 31.16.

Une telle suite est régularisante parce que souvent $\rho_n \in \mathscr{D}(\mathbb{R}^d)$, ce qui donne $f * \rho_n \in C^\infty$ par le corollaire 28.43.

Proposition 31.19 ([370, 371]).
Soit une suite régularisante $\rho_n \in L^1(\mathbb{R}^d)$. Alors :

(1) Si f est continue à support compact, nous avons la convergence uniforme sur $\mathbb{R}^d$:

$$f * \rho_n \xrightarrow{unif} f. \tag{31.89}$$

(2) Si $g \in L^p$ $(1 \leqslant p < \infty)$ alors

$$g * \rho_n \xrightarrow{L^p} g. \tag{31.90}$$

Démonstration. Si f est continue à support compact, elle est uniformément continue [5], et elle est bornée. Soit $\epsilon > 0$ et $\alpha > 0$ tel que pour tout x, y tels que $\|x-y\| < \alpha$ nous ayons $|f(x)-f(y)| < \epsilon$. Nous prenons de plus n suffisamment grand pour avoir $\int_{B(0,\alpha)^c} \rho_n < \epsilon$. Nous avons alors

$$|f(x) - (f*\rho_n)(x)| = |\int_{\mathbb{R}^d} \big(f(x)-f(y)\big)\rho_n(x-y)dy| \tag{31.91a}$$

$$\leqslant \int_{B(x,\alpha)} \underbrace{|f(x)-f(y)|}_{\leqslant\epsilon} \rho_n(x-y)dy + \int_{B(x,\alpha)^c} \underbrace{|f(x)-f(y)|}_{\leqslant 2\|f\|_\infty} \rho_n(x-y)dy \tag{31.91b}$$

$$\leqslant \epsilon(1+2\|f\|_\infty). \tag{31.91c}$$

Nous avons prouvé que pour tout $\epsilon > 0$, il existe N tel que $n > N$ implique $\big|f(x)-(f*\rho_n)(x)\big| \leqslant \epsilon$. Cela prouve l'uniforme convergence sur $\mathbb{R}^d$ de $f * \rho_n$ vers f.

Pour le point (2) nous considérons $g \in L^1(\mathbb{R}^d)$ et $\phi \in \mathscr{D}(\mathbb{R}^d)$. Nous avons la majoration

$$\|g*\rho_n - g\|_p \leqslant \|g*\rho_n - \phi*\rho_n\|_p + \|\phi*\rho_n - \phi\|_p + \|\phi - g\|_p \tag{31.92}$$

En ce qui concerne le premier terme ;

$$\|(g-\phi)*\rho_n\|_p \leqslant \|g-\phi\|_p \tag{31.93}$$

par la proposition 28.41. Donc

$$\|g*\rho_n - g\|_p \leqslant 2\|g-\phi\|_p + \|\phi*\rho_n - \phi\|_p. \tag{31.94}$$

Par la densité de $\mathscr{D}$ dans L^p (théorème 28.34(5)) nous pouvons considérer une suite $\phi_i \xrightarrow{L^p} g$ dans $\mathscr{D}(\mathbb{R}^d)$. Pour tout i nous avons

$$\|g*\rho_n - g\|_p \leqslant 2\|g-\phi_i\|_p + \|\phi_i*\rho_n - \phi\|_p. \tag{31.95}$$

Nous effectuons la limite sur $n \to \infty$:

$$\lim_{n\to\infty} \|g*\rho_n - g\|_p \leqslant 2\|g-\phi_i\| + \underbrace{\lim_{n\to\infty} \|\phi_i*\rho_n - \phi_i\|_p}_{=0} \tag{31.96}$$

parce que le point (1) s'applique à ϕ_i. Nous effectuons ensuite la limite sur $i \to \infty$ dans

$$\lim_{n\to\infty} \|g*\rho_n - g\| \leqslant 2\|g-\phi_i\| \to 0. \tag{31.97}$$

□

5. Théorème de Heine 14.104.

Lemme 31.20.
Si $g_\epsilon(x) = e^{-\epsilon\|x\|^2}$ alors la suite

$$\rho_n = \frac{1}{(2\pi)^d}\hat{g}_{1/n} \tag{31.98}$$

est une suite régularisante (définition 31.18).

Démonstration. Nous savons déjà la transformée de Fourier de g_ϵ par le lemme 31.17. Nous montrons que la suite ρ_n est régularisante. Nous avons $\hat{g}_\epsilon \in L^1(\mathbb{R}^d)$ et $\hat{g}_\epsilon \geqslant 0$ ainsi que $\lim_{\epsilon\to 0}\int_{B(0,\alpha)}\hat{g}_\epsilon = 0$ pour tout α. Il y a seulement un couac avec la norme. Nous calculons $\int_{\mathbb{R}^d}\hat{g}_\epsilon(\xi)d\xi$ avec la forme (31.87c). En utilisant sauvagement Fubini [6] pour séparer les intégrales et en effectuant le changement de variable $u = t/(2\sqrt{\epsilon})$ nous calculons :

$$\begin{aligned}
\int_{\mathbb{R}^d} e^{-|\xi|^2/4\epsilon}d\xi &= \prod_{k=1}^{d}\int_{\mathbb{R}} e^{-t^2/4\epsilon}dt && (31.99a)\\
&= 2\sqrt{\epsilon}\prod_{k=1}^{d}\int_{\mathbb{R}} e^{-u^2}du && (31.99b)\\
&= \prod_{k=1}^{d} 2\sqrt{\epsilon}\sqrt{\pi} && (31.99c)\\
&= 2^d(\pi\epsilon)^{d/2}. && (31.99d)
\end{aligned}$$

Nous avons utilisé l'exemple 16.242 pour le calcul de l'intégrale gaussienne. Avec tout cela nous avons

$$\int_{\mathbb{R}^d}\hat{g}_\epsilon = (2\pi)^d. \tag{31.100}$$

Donc $\frac{1}{(2\pi)^d}\hat{g}_{1/n}$ est une suite régularisante. □

Le corollaire suivant regroupe les résultats à propos des suites régularisantes, leur utilité et leur existence.

Corollaire 31.21.
Si la suite régularisante ρ_n est dans $L^1(\mathbb{R}^d)\cap C^\infty(\mathbb{R}^d)$ alors pour $f \in L^p(\mathbb{R}^d)$ en posant $f_n = \rho_n * f$ nous avons

(1) $f_n \in C^\infty(\mathbb{R}^d)\cap L^p(\mathbb{R}^d)$

(2) $f_n \xrightarrow{L^p} f$

De plus, de telles suites existent.

Démonstration. Le fait que f_n soit de classe C^∞ est le corollaire 28.43, et la convergence est la proposition 31.19(2).

De telles suites existent, par exemple celle donnée par le lemme 31.20. □

31.3.1 Formule d'inversion

Proposition 31.22 (Formule d'inversion de Fourier[4])**.**
Si $f \in \mathscr{S}(\mathbb{R})$, alors nous avons la formule d'inversion

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{ikx}\hat{f}(k)dk. \tag{31.101}$$

6. Le pauvre !

Cette formule peut d'écrire de plusieurs autres façons :

$$\mathcal{F}\big(\mathcal{F}(f)\big)(x) = 2\pi f(-x), \tag{31.102a}$$

$$\mathcal{F}^{-1}(f)(x) = \frac{1}{2\pi}\hat{f}(-x), \tag{31.102b}$$

$$f(x) = \frac{1}{2\pi}\mathcal{F}(\hat{f})(-x). \tag{31.102c}$$

Démonstration. Pour $\epsilon > 0$ nous posons

$$f_\epsilon(k) = e^{-\epsilon k^2} e^{ikx}\hat{f}(k). \tag{31.103}$$

Nous allons calculer

$$\lim_{\epsilon\to 0}\int_{\mathbb{R}} e^{-\epsilon k^2} e^{ikx}\hat{f}(k)dk \tag{31.104}$$

de deux façons.

D'abord en utilisant directement le théorème de la convergence dominée 16.140. La fonction $\hat{f}$ est dans $\mathscr{S}(\mathbb{R})$ (théorème 31.14) et par conséquent $f_\epsilon \in L^1(\mathbb{R})$ parce que le facteur $e^{-\epsilon k^2}$ ne va certainement pas empêcher de converger. De plus $|f_\epsilon| \leqslant |\hat{f}|$ et $\hat{f} \in L^1$. Le théorème est de la convergence dominée est applicable et

$$\lim_{\epsilon\to 0}\int_{\mathbb{R}} e^{-\epsilon k^2} e^{ikx}\hat{f}(k)dk = \int_{\mathbb{R}} e^{ikx}\hat{f}(k)dk. \tag{31.105}$$

Pour le deuxième calcul nous allons faire appel à Fubini [7] pour la fonction

$$\begin{aligned} u\colon \mathbb{R}\times\mathbb{R} &\to \mathbb{R} \\ (k,y) &\mapsto e^{ik(x-y)}e^{-\epsilon k^2}f(y). \end{aligned} \tag{31.106}$$

D'abord nous nous assurons que $u \in L^1(\mathbb{R}\times\mathbb{R})$ par le corollaire 16.238, et ensuite nous utilisons le théorème de Fubini 16.239 pour manipuler les intégrales (et en particulier les inverser). Dans un premier temps nous avons :

$$\int_{\mathbb{R}}\int_{\mathbb{R}} |e^{ik(x-y)}e^{-\epsilon k^2}f(y)|dy\,dk \leqslant \int_{\mathbb{R}} e^{-\epsilon k^2}\Big[\int_{\mathbb{R}}|f(y)|dy\Big]dk < \infty \tag{31.107}$$

parce que f étant dans $\mathscr{S}(\mathbb{R})$, l'intégrale intérieure se réduit à un nombre. Nous savons maintenant que $u \in L^1(\mathbb{R}\times\mathbb{R})$. Nous pouvons alors calculer un peu ...

$$\int_{\mathbb{R}} e^{ikx}e^{-\epsilon k^2}\hat{f}(k)dk = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{ikx}e^{-\epsilon k^2}e^{-iky}f(y)dy\,dk \tag{31.108a}$$

$$= \int_{\mathbb{R}}\Big[\int_{\mathbb{R}} e^{ik(x-y)}e^{-\epsilon k^2}f(y)dk\Big]dy \tag{31.108b}$$

$$= \int_{\mathbb{R}} f(y)\Big[\int_{\mathbb{R}} e^{ik(x-y)}e^{-\epsilon k^2}dk\Big]dy \tag{31.108c}$$

$$= \int_{\mathbb{R}} f(y)\hat{g}_\epsilon(y-x)dy \tag{31.108d}$$

$$= \sqrt{\frac{\pi}{\epsilon}}\int_{\mathbb{R}} f(y)e^{-(y-x)^2/4\epsilon}dy \tag{31.108e}$$

$$= 2\sqrt{\epsilon}\sqrt{\frac{\pi}{\epsilon}}\int_{\mathbb{R}} f(x+2\sqrt{\epsilon}t)e^{-t^2}dt \tag{31.108f}$$

$$= 2\sqrt{\pi}\int_{\mathbb{R}} f(x+2\sqrt{\epsilon}t)e^{-t^2}dt \tag{31.108g}$$

$$\tag{31.108h}$$

Justifications :

7. Parce qu'il est toujours plus simple de refiler le boulot aux autres que de le faire soi-même... pauvre Fubini !

— La fonction g_ϵ est la gaussienne dont la transformée de Fourier est a été l'objet du lemme 31.17.
— Nous avons effectué le changement de variables $t = (y-x)/(2\sqrt{\epsilon})$ qui donne $dt = dy/2\sqrt{\epsilon}$.

La fonction f étant Schwartz (en particulier bornée), dans la dernière intégrale, nous pouvons effectuer la majoration

$$f(x + 2\sqrt{\epsilon}t)e^{-t^2} \leqslant \|f\|_\infty e^{-t^2}, \tag{31.109}$$

qui est une fonction intégrable. Nous pouvons donc permuter la limite et l'intégrale. Dans l'égalité

$$\lim_{\epsilon\to 0} \int_{\mathbb{R}} e^{ikx} e^{-\epsilon k^2} \hat{f}(k)dk = \lim_{\epsilon\to 0} 2\sqrt{\pi} \int_{\mathbb{R}} f(x + 2\sqrt{\epsilon}t)e^{-t^2}dt \tag{31.110}$$

À gauche nous avons déjà la limite depuis (31.105), et à droite nous obtenons

$$\lim_{\epsilon\to 0} 2\sqrt{\pi} \int_{\mathbb{R}} f(x + 2\sqrt{\epsilon}t)e^{-t^2}dt = \int_{\mathbb{R}} f(x)e^{-t^2}dt = 2\sqrt{\pi}f(x)\sqrt{\pi} = 2\pi f(x) \tag{31.111}$$

où nous avons utilisé l'intégrale gaussienne faite dans l'exemple 16.242.

En remettant tout ensemble,

$$2\pi f(x) = \lim_{\epsilon\to 0} \int_{\mathbb{R}} e^{-\epsilon k^2} e^{ikx} \hat{f}(k)dk = \int_{\mathbb{R}} e^{ikx} \hat{f}(k)dk, \tag{31.112}$$

ce qu'il fallait prouver. □

Corollaire 31.23.
Nous avons la formule

$$\int_{\mathbb{R}} \int_{\mathbb{R}} e^{-ikx} f(x)dx\, dk = 2\pi f(0). \tag{31.113}$$

Démonstration. Poser $x = 0$ dans l'équation (31.101). □

31.24.
Les physiciens qui n'ont que rarement peur écrivent souvent la formule (31.113) sous la forme

$$\int_{\mathbb{R}} e^{-ikx} dk = \delta(x) \tag{31.114}$$

où δ serait la fonction de Dirac qui vaut zéro partout sauf en $x = 0$ où elle vaudrait l'infini, mais pas n'importe quel infini ; juste celui qu'il faut pour que sont intégrale valle 1.

Lemme 31.25.
Si $\phi \in \mathscr{S}(\mathbb{R} \times \mathbb{R}^n)$, alors

$$\partial_t \hat{\phi} = \widehat{\partial_t \phi} \tag{31.115}$$

où le chapeau dénote la transformée de Fourier par rapport à la variable dans $\mathbb{R}^n$ et non par rapport à celle dans $\mathbb{R}$. Le t par contre est la variable dans $\mathbb{R}$.

Démonstration. Par définition de la transformée de Fourier nous avons

$$(\partial_t \hat{\phi})(t, \xi) = \frac{\partial}{\partial t} \int_{\mathbb{R}^n} \phi(t, x)e^{-ix\xi}dx. \tag{31.116}$$

Notre but est de permuter l'intégrale et la dérivée en utilisant le théorème 19.27. Il nous faut une fonction $G\colon \mathbb{R}^n \to \mathbb{R}$ qui soit intégrable sur $\mathbb{R}^n$ et telle que

$$\left| \frac{\partial \phi}{\partial t} \phi(t, x) \right| \leqslant G(x) \tag{31.117}$$

pour tout $t \in B(t_0, \delta)$. Étant donné que la fonction $\partial_t \phi$ est tout autant Schwartz que ϕ elle-même nous pouvons alléger les notations et chercher une fonction G qui convient pour toute fonction $\varphi \in \mathscr{S}(\mathbb{R} \times \mathbb{R}^n)$. Soit la fonction

$$G(x) = \sup_{t \in B(t_0,\delta)} |\varphi(t, x)|. \tag{31.118}$$

Pour tout multiindice α nous avons alors

$$\sup_{x \in \mathbb{R}^n} \left|x^\alpha G(x)\right| \leqslant \sup_{(t,x) \in \mathbb{R} \times \mathbb{R}^n} \left|x^\alpha \varphi(t, x)\right| \leqslant M_\alpha \in \mathbb{R}. \tag{31.119}$$

Grâce à la proposition 28.90, cela signifie que φ décroît plus vite que n'importe quel polynôme ; G est donc intégrable sur $\mathbb{R}^n$. □

31.4 Transformée de Fourier sur $L^2(\mathbb{R}^d)$

La théorie des transformées de Fourier est intéressante sur $L^2(\mathbb{R}^d)$ parce qu'elle y donne une isométrie. Nous allons la donner avec des fonctions à valeurs dans $\mathbb{C}$.

Remarque 31.26.
Une remarque qui vaut ce qu'elle vaut, mais si u est une classe de fonction pour la relation $u \sim v$ si et seulement si $u(x) = v(x)$ pour presque tout v alors l'intégrale

$$\hat{u}(\xi) = \int_{\mathbb{R}^d} u(x) e^{ix\xi} dx \tag{31.120}$$

ne dépend pas du choix du représentant. Nous pouvons donc parfaitement parler de transformée de Fourier d'une classe de fonctions.

31.4.1 Le problème

Nous avons défini en général la transformée de Fourier d'une fonction $f : \mathbb{R} \to \mathbb{C}$ par la formule

$$\hat{f}(\xi) = \int_{\mathbb{R}} e^{i\xi x} f(x) dx \tag{31.121}$$

tant que cette intégrale existe.

Il se fait que cette intégrale n'existe pas toujours pour des fonctions dans $L^2(\mathbb{R})$. Donc nous devons faire mieux pour définir la transformée de Fourier sur L^2.

Exemple 31.27([1])
Prenons la fonction

$$f(x) = \begin{cases} 0 & \text{si } x < 1 \\ \frac{1}{x} & \text{si } x \geqslant 1. \end{cases} \tag{31.122}$$

Vu que l'intégrale $\int_1^\infty \frac{1}{x^2} dx$ existe et est finie (proposition 16.227(2)), la fonction f est dans $L^2(\mathbb{R})$.

Cependant l'intégrale (31.121) n'existe pas. Pour nous convaincre de cela, nous pouvons simplement nous souvenir de la définition d'une intégrale à valeurs dans un espace vectoriel (définition 16.132). Nous fixons $\xi \in \mathbb{R}$ et nous posons $g(x) = f(x)e^{i\xi x}$.

Bien évidemment, $|f(g)| = \frac{1}{x}$ sur $]1, \infty[$. Donc $\int_{\mathbb{R}} |g| = \infty$, et la fonction g n'est pas intégrable. Fin de l'histoire.

Nous pouvons toujours essayer de comprendre mieux. Vu que $\int_{\mathbb{R}} |g| = \infty$, la proposition 16.133 nous dit qu'au moins une des intégrales parmi

$$\int f_{re}^+, \int f_{im}^+, \int f_{re}^-, \int f_{im}^- \tag{31.123}$$

est égale à $+\infty$.

Note qu'en travaillant un peu, on se convainc qu'en réalité, elles divergent toutes les quatre.
△

31.4.2 Extension de $L^1 \cap L^2$ vers L^2

Théorème 31.28 (Extention de la transformée de Fourier vers $L^2(\mathbb{R}^d)$[369]).
Soit $f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. Alors

(1) Nous avons $\mathcal{F}(f) \in L^2$ et $\|\hat{f}\|_{L^2} = (2\pi)^d \|f\|_{L^2}$.

(2) L'application $\mathcal{F}\colon L^1 \cap L^2 \to L^2$ peut être étendue en une application $\mathcal{F}\colon L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ vérifiant

$$\|\hat{f}\|_{L^2} = (2\pi)^d \|f\|_{L^2} \tag{31.124}$$

pour tout $f \in L^2(\mathbb{R}^d)$.

Démonstration. Le fait que $f \in L^1$ implique $\|\mathcal{F}(f)\|_\infty \leqslant \|f\|_1$ (c'est le lemme 31.8). En particulier, $|\mathcal{F}(f)(\xi)|^2$ est majoré et l'intégrale

$$\clubsuit = \int_{\mathbb{R}^d} |\hat{f}|^2 e^{-\epsilon\xi^2} d\xi \tag{31.125}$$

existe et est finie.

Découper l'intégrale Dans un premier temps nous développons les intégrales. Dans les égalités suivantes, $x\xi$ est le produit scalaire $x \cdot \xi$ dans $\mathbb{R}^d$.

$$\clubsuit = \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \overline{f(x)} e^{ix\xi} dx \right) \left(\int_{\mathbb{R}^d} f(y) e^{-y\xi} \right) e^{-\epsilon|\xi|^2} d\xi \tag{31.126a}$$

$$= \int_{\mathbb{R}^d} \left[\int_{\mathbb{R}^d \times \mathbb{R}^d} \overline{f(x)} f(y) e^{i\xi(x-y)} dxdy \right] e^{-\epsilon|\xi|^2} d\xi. \tag{31.126b}$$

Nous avons utilisé le théorème de Fubini pour regrouper les intégrales [8]. Vu que $f \in L^1(\mathbb{R}^d)$, la fonction $(x, y, \xi) \mapsto f(x)f(y)e^{-\epsilon|\xi|^2}$ est dans $L^1(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ et le théorème de Fubini 16.239 avec $\Omega_1 = \mathbb{R}^d \times \mathbb{R}^d$ et $\Omega_2 = \mathbb{R}^d$ nous permet de permuter les intégrales pour avoir

$$\clubsuit = \int_{\mathbb{R}^d \times \mathbb{R}^d} \overline{f(x)} f(y) \left[\int_{\mathbb{R}^d} e^{i\xi(x-y)} e^{-\epsilon|\xi|^2} d\xi \right] dxdy. \tag{31.127}$$

Discuter de cette gaussienne En posant

$$g(x) = e^{-|x|^2/2} \tag{31.128a}$$

$$g_\epsilon(x) = g(\sqrt{2\epsilon}x) = e^{-\epsilon|x|^2} \tag{31.128b}$$

nous avons $g_\epsilon \in \mathscr{S}(\mathbb{R}^d)$ et le lemme 31.20 nous autorise à écrire

$$\hat{g}(\xi) = (2\pi)^{d/2} g(\xi) \tag{31.129a}$$

$$\hat{g}_\epsilon(\xi) = \left(\frac{\pi}{\epsilon}\right)^{d/2} e^{-|\xi|^2/4\epsilon} \tag{31.129b}$$

Nous voyons que $\hat{g}_\epsilon \in \mathscr{S}(\mathbb{R}^d)$ (c'était gagné d'avance par la proposition 31.14) et que $\hat{g}_\epsilon$ est une fonction paire (encore une fois, c'était gagné d'avance parce que la transformée de Fourier d'une fonction paire est paire).

Tout cela pour dire que l'intégrale entre crochet dans (31.127) est $\hat{g}_\epsilon(y-x) = \hat{g}_\epsilon(x-y)$, et donc

$$\clubsuit = \int_{\mathbb{R}^d \times \mathbb{R}^d} \overline{f(x)} f(y) \hat{g}_\epsilon(x-y) dxdy. \tag{31.130}$$

Encore une fois le théorème de Fubini permet de séparer les intégrales et de calculer l'intégrale sur y en premier. Vu que $f \in L^1$ et que $\hat{g}_\epsilon \in \mathscr{S}(\mathbb{R}^d)$, le produit de convolution $f * \hat{g}_\epsilon$ est un élément de $\mathscr{S}(\mathbb{R}^d)$ par la proposition 28.99. Nous avons donc

$$\clubsuit = \int_{\mathbb{R}^d} \overline{f(x)} (f * \hat{g}_\epsilon)(x) dx. \tag{31.131}$$

8. Dans la suite nous allons encore utiliser Fubini quelques fois pour regrouper et dégrouper des intégrales.

Là, nous reconnaissons un produit scalaire dans $L^2(\mathbb{R}^d)$, et donc

$$\int_{\mathbb{R}^d} |\hat{f}|^2 e^{-\epsilon\xi^2} d\xi = \langle f, f * \hat{g}_\epsilon \rangle_{L^2(\mathbb{R}^d)}. \tag{31.132}$$

Notons que tout a un sens : $f \in L^2(\mathbb{R}^d)$ et $f * \hat{g}_\epsilon \in \mathscr{S}(\mathbb{R}^d) \subset L^2(\mathbb{R}^d)$.

Suite régularisante Nous prenons la suite régularisante du lemme 31.20 donnée par

$$\rho_n = \frac{1}{(2\pi)^d} \hat{g}_{1/n}. \tag{31.133}$$

Première conclusion Nous reprenons (31.132)

$$\int_{\mathbb{R}^d} |\hat{f}|^2 e^{-|\xi^2|/n} d\xi = \langle f, f * \hat{g}_{1/n} \rangle_{L^2(\mathbb{R}^d)} = (2\pi)^d \langle f, f * \rho_n \rangle. \tag{31.134}$$

En prenant la limite $n \to \infty$ nous trouvons

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} |\hat{f}|^2 e^{-\epsilon\xi^2} d\xi = (2\pi)^d \|f\|^2. \tag{31.135}$$

Pour effectuer la limite du membre de gauche nous devons remarquer qu'en posant

$$g_n(\xi) = |\hat{f}(\xi)| e^{-|\xi|^2/n}, \tag{31.136}$$

nous avons une suite décroissante de fonction (c'est à dire que à ξ fixé, c'est décroissant en n). Par ailleurs ces fonctions sont toujours à valeurs dans $[0, \infty]$ et nous pouvons utiliser le théorème de la convergence monotone 16.116 pour permuter la limite et l'intégrale. Au final :

$$\|\hat{f}\|_{L^2} = (2\pi)^d \|f\|_{L^2}. \tag{31.137}$$

En ce qui concerne l'extension, soit $f \in L^2(\mathbb{R}^d)$ et une suite (f_n) dans $L^1 \cap L^2$ telle que $f_n \xrightarrow{L^2} f$.

Existence d'une telle suite Si $f \in L^2(\mathbb{R}^d)$, alors nous pouvons poser

$$f_n(x) = f(x) e^{-|x|^2/n^2}. \tag{31.138}$$

Par l'inégalité de Hölder (28.51) nous avons $f_n \in L^1(\mathbb{R}^d)$; de plus $f_n \in L^2(\mathbb{R}^d)$ parce que pour tout x nous avons $|f_n(x)| \leqslant |f(x)|$. Montrons que $f_n \xrightarrow{L^2} f$. Nous avons

$$\|f_n - f\|_{L^2}^2 = \int_{\mathbb{R}^d} |f(x)(1 - e^{-|x^2|/n^2})|^2 dx. \tag{31.139}$$

Nous voulons prendre la limite $n \to \infty$. Pour ce faire à à droite nous remarquons que $e^{-|x|^2/n^2}$ est majoré par 1 ; ce qui se trouve dans l'intégrale est donc majoré (uniformément en n) par $|f(x)|^2$, qui est une fonction L^1 parce que f est L^2. Le théorème de la convergence dominée 16.140 nous permet alors de permuter la limite et l'intégrale, ce qui donne

$$\lim_{n\to\infty} \|f_n - f\|_2^2 = \int_{\mathbb{R}^d} \lim_{n\to\infty} |f(x)(1 - e^{-|x|^2/n^2})|^2 dx = 0. \tag{31.140}$$

Définition de $\mathcal{F}: L^2 \to L^2$ La suite (f_n) est une suite convergence dans L^2, et elle est donc de Cauchy. De plus pour chaque n, m nous avons

$$\|\hat{f}_n - \hat{f}_m\| = (2\pi)^d \|f_n - f_m\|. \tag{31.141}$$

Nous voyons donc que la suite $(\hat{f}_n)$ est également de Cauchy, dans l'espace $L^2(\mathbb{R}^d)$ qui est complet (lemme 28.58). Nous posons

$$\hat{f} = \lim_{n\to\infty} \hat{f}_n. \tag{31.142}$$

Indépendance aux choix Nous devons montrer que la définition de $\hat{f}$ ne dépend pas de la suite approximant f dans $L^1 \cap L^2$. Soient dans deux suites $f_n \xrightarrow{L^2} f$ et $g_n \xrightarrow{L^2} f$ telles que $\hat{f}_n \xrightarrow{L^2} F$ et $\hat{g}_n \xrightarrow{L^2} G$. Alors

$$\|\hat{f}_n - \hat{g}_n\| = (2\pi)^d \|f_n - g_n\| \leqslant (2\pi)^d \|f_n - f\| + (2\pi)^d \|g_n - f\| \to 0. \tag{31.143}$$

Par conséquent $(\hat{f}_n - \hat{g}_n)_n$ est une suite qui converge vers zéro. Par unicité de la limite, $F = G$.

□

Remarque 31.29.
Une autre suite possible, à la place de (31.138), est

$$f_n(x) = f(x)\mathbb{1}_{|x|<n}. \tag{31.144}$$

C'est à dire la fonction f limitée à une boule de rayon n autour de 0.

Chapitre 32

Distributions

Nous donnons ici une partie de la théorie sur les distributions. L'utilisation des distributions dans le cadre des équations différentielles est mise dans le chapitre sur les équations différentielles, section 34.11.

32.1 Topologie

32.1.1 Dérivée partielle au sens faible

Lemme-définition 32.1.
Soit $f \in L^p(I)$ où I est l'intervalle ouvert $]a,b[$. Il existe au maximum[1] *une fonction g telle que*

$$\int_I f\varphi' = -\int_I g\varphi \tag{32.1}$$

pour tout $\varphi \in \mathscr{D}(I)$. Lorsqu'une telle fonction existe, nous la nommons ***dérivée faible*** *de f.*

Démonstration. Soient $g, h \in L^2$ tels que

$$\int_I u\varphi' = -\int_I g\varphi = -\int_I h\varphi \tag{32.2}$$

pour tout $\varphi \in C_c^\infty(I)$. Nous avons alors

$$\int_I (g-h)\varphi = 0. \tag{32.3}$$

Cela implique que $g - h = 0$ presque partout par la proposition 28.76 [2]. □

Exemple 32.2(Dérivée faible de $\mathbb{1}_{\mathbb{Q}}$)
Vu que $\mathbb{Q}$ est de mesure nulle dans $\mathbb{R}$, nous avons

$$\int_{\mathbb{R}} \mathbb{1}_{\mathbb{Q}}\varphi' = 0 \tag{32.4}$$

pour tout $\varphi \in \mathscr{D}(\mathbb{R})$. Pour $g = 0$ nous avons aussi $\int_{\mathbb{R}} g\varphi = 0$. Donc $g = 0$ est la dérivée faible de $\mathbb{1}_{\mathbb{Q}}$.

Cela n'est pas étonnant du fait qu'en théorie de l'intégration, les parties de mesure nulle ne comptent pas. De ce point de vue, $\mathbb{1}_{\mathbb{Q}} = 0$. D'ailleurs cette égalité est vraie dans L^p (les classes et tout ça). △ <++>

1. En réalité, c'est une classe au sens de l'égalité presque partout.
2. Ou alors par le lemme 28.44 qui est moins général mais tout aussi bien pour ici.

Exemple 32.3(La fonction de Heaveside n'a pas de dérivée faible)
Nous montrons que la fonction

$$H(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geqslant 0 \end{cases} \tag{32.5}$$

n'a pas dérivée faible. Nous nommons g une hypothétique fonction vérifiant les conditions pour être la dérivée faible de H.

Soit une fonction $\varphi \in \mathscr{D}(\mathbb{R})$, dont le support est contenu dans $]0, \infty[$. Sur le support de φ, et donc aussi de φ', nous avons $H(x) = 1$ et donc

$$\int_{\mathbb{R}} \varphi' = -\int_{\mathbb{R}} g\varphi. \tag{32.6}$$

Vu que φ est à support compact, $\int_{\mathbb{R}} \varphi' = 0$. En effet, si le support de φ est contenu dans $[-M, M]$, alors en utilisant le théorème fondamental de l'analyse 16.212, nous trouvons $\int_{\mathbb{R}} \varphi = \int_{-M}^{M} \varphi' = \varphi(M) - \varphi(-M) = 0 - 0 = 0$.

Donc g doit satisfaire

$$\int_{\mathbb{R}} g(x)\varphi(x) = 0 \tag{32.7}$$

pour tout $\varphi \in \mathscr{D}(x > 0)$. La proposition 28.1 nous dit que $g = 0$ presque partout sur $]0, \infty[$.

Le même raisonnement dit que $g = 0$ presque partout sur les négatifs. Que g soit maintenant nulle ou non en $x = 0$ ne change pas le fait que $g = 0$ presque partout sur $\mathbb{R}$.

Par conséquent, $\int_{\mathbb{R}} g\varphi = 0$ pour toute $\varphi \in \mathscr{D}(\mathbb{R})$. Hélas, nous avons d'autre part

$$\int_{\mathbb{R}} H(x)\varphi'(x)dx = \int_0^{\infty} \varphi'(x)dx, \tag{32.8}$$

qui n'est pas forcément nul. Notons que pour avoir un exemple de φ qui donne $\int_{\mathbb{R}} H\varphi \neq 0$, il faut chercher des fonctions dont le support contient des négatifs et des positifs.

La fonction H n'a donc pas de dérivée faible. Notons cependant que cela ne présume en rien la possibilité d'accepter une dérivée au sens des distributions. △

Exemple 32.4(Dérivée faible de la valeur absolue)
L'exemple de base de fonction continue qui n'est pas dérivable est la valeur absolue $f(x) = |x|$ prise en $x = 0$. Nous allons montrer ici que la fonction

$$H(x) = \begin{cases} 1 & \text{si } x < 0 \\ -1 & \text{si } x > 0 \end{cases} \tag{32.9}$$

est la dérivée faible de f.

Commençons par noter que H peut valoir la valeur qu'on veut en zéro ; de toutes façons la dérivée faible n'est définie qu'à partie de mesure nulle près.

Soit $\varphi \in \mathscr{D}(\mathbb{R})$ et $M > 0$ tel que le support de φ soit contenu dans $[-M, M]$. Nous avons d'une part

$$\int_{\mathbb{R}} |x|\varphi'(x)dx = -\int_{-M}^{0} x\varphi'(x)dx + \int_0^{M} x\varphi'(x)dx \tag{32.10a}$$

$$= -[\varphi x]_{-M}^{0} + \int_{-M}^{0} \varphi + [x\varphi]_0^M - \int_0^M \varphi \tag{32.10b}$$

où nous avons utilisé l'intégration par partie de la proposition 22.132 en posant

$$u = x \qquad v' = \varphi' \tag{32.11a}$$

$$u' = 1 \qquad v = \varphi. \tag{32.11b}$$

Tout cela pour dire que

$$\int_{\mathbb{R}} |x|\varphi'(x)dx = \int_{-M}^{0} \varphi - \int_{0}^{M} \varphi. \tag{32.12}$$

D'autre part, l'égalité

$$\int_{\mathbb{R}} H(x)\varphi(x)dx = \int_{-M}^{0} \varphi - \int_{0}^{M} \varphi \tag{32.13}$$

est immédiate.

Nous en déduisons que H est bien la dérivée faible de $x \mapsto |x|$. △

Remarque 32.5.
La dérivée faible ne doit pas être confondue avec la dérivée au sens des distributions qui sera définie en 32.20. Nous avons donc trois notions distinctes de dérivation pour une fonction :

— la dérivée usuelle,
— la dérivée au sens des distributions,
— la dérivée faible.

La dérivée faible d'une fonction reste une fonction, tandis que la dérivée distributionnelle d'une fonction est une distribution.

Je vous mets en garde contre l'idée que l'existence de l'une impliquerait trop facilement l'existence d'une autre[3].

32.1.2 Dérivée faible partielle

La notion de dérivée partielle faible est la même que l'autre. Histoire de nous mettre dans le bain, nous écrivons la définition avec les notation du produit scalaire au lieu de l'intégrale.

Définition 32.6.
*Si $i = 1, \dots, n$, la **dérivée faible** de v dans la direction e_i est l'application*[4] *notée $\partial_i v$ définie par*

$$\langle \partial_i v, \phi \rangle = -\langle v, \partial_i \phi \rangle \tag{32.14}$$

pour tout $\phi \in C_c^\infty(\Omega)$.

Lemme 32.7.
Si $v \in L^2$ admet une dérivée faible, alors cette dernière est unique.

Démonstration. Supposons f, g telles que $\langle g, \phi \rangle$ et $\langle f, \phi \rangle$ soient tous deux égaux à $-\langle v, \partial_i \phi \rangle$. En particulier pour tout $\phi \in \mathscr{D}(\Omega)$ nous avons $\langle (f - g), \phi \rangle = 0$.

Cela donne $f - g = 0$ par la proposition 28.76. □

Soit Ω un ouvert de $\mathbb{R}^d$. Le but de notre histoire est de définir une distribution comme étant un élément de l'espace dual (topologique, voir définition 12.26) de l'espace $\mathscr{D}(\Omega)$ des fonctions C^∞ à support compact dans Ω. Pour ce faire nous devons voir un peu de topologie sur différents espaces de fonctions. Notons que l'espace $\mathscr{D}(\Omega)$ n'est pas réduit à la fonction nulle comme en témoigne l'exemple donné par l'équation (17.452).

Pour chaque K compact dans Ω et multiindice $\alpha \in \mathbb{N}^d$ nous considérons sur $C^\infty(\Omega)$ la seminorme suivante :

$$p_{K,m}(f) = \sum_{|\mu| \leqslant m} \|\partial^\mu f\|_{K,\infty}. \tag{32.15}$$

En particulier,

$$p_{K,0}(f) = \sup_{x \in K} |f(x)| = \|f\|_{\infty,K}. \tag{32.16}$$

3. Wikipédia cite l'exemple de la fonction de Cantor qui est dérivable presque partout au sens usuel, mais qui n'est pas faiblement dérivable. Écrivez-moi si vous connaissez des théorèmes qui lient les trois notions de dérivée.
4. En fait c'est une classe au sens de l'égalité presque partout.

32.1.3 Topologie, convergence

Définition 32.8.
Les topologies que nous allons considérer sont :

(1) Sur $C^\infty(\Omega)$, la topologie des semi-normes $p_{K,m}$ (avec K et m comme paramètres).

(2) Sur $\mathscr{D}(K)$, la topologie des semi-normes $p_{K,m}$ (avec seulement m comme paramètre).

(3) Sur $\mathscr{D}(\Omega)$, la topologie induite de $C^\infty(\Omega)$.

Cela n'est pas très explicite, mais heureusement nous n'aurons souvent pas besoin de plus que de la notion de convergence dans $\mathscr{D}'(\Omega)$. Rappelons que la topologie d'un espace donne la notion de convergence par la définition 8.25.

Lemme 32.9 (Convergence dans $\mathscr{D}(K)$)**.**
Si α est un multiindice et si $\varphi_n \xrightarrow{\mathscr{D}(K)} \varphi$, alors nous avons

$$\partial^\alpha \varphi_n \xrightarrow{unif} \partial^\alpha \varphi. \tag{32.17}$$

Démonstration. Quitte à considérer la suite $\varphi_n - \varphi$ nous pouvons supposer $\varphi_n \xrightarrow{\mathscr{D}(K)} 0$. Nous avons

$$\|\partial^\alpha \varphi_n\| \leqslant \sum_{\mu \leqslant \alpha} \|\partial^\mu \varphi_n\|_{K,\infty}. \tag{32.18}$$

Vu que le membre de droite tend vers zéro, nous avons

$$\lim_{n\to\infty} \|\partial^\alpha \varphi_n\|_{K,\infty} \to 0, \tag{32.19}$$

ce qui revient à dire que $\partial^\alpha \varphi_n$ converge uniformément sur K vers $\partial^\alpha \varphi$. □

Lemme 32.10.
Si une fonction $f : \mathscr{D}(\Omega) \to \mathbb{R}$ est continue sur chacun des $\mathscr{D}(K)$ pour tout K compact dans Ω alors est continue sur $\mathscr{D}(\Omega)$.

Démonstration. Soit I ouvert dans $\mathbb{R}$; nous devons trouver un ouvert $\mathcal{O}$ dans $C^\infty(\Omega)$ tel que $f^{-1}(I) = \mathscr{D}(\Omega) \cap \mathcal{O}$. Vu que f est continue sur chacun des $\mathscr{D}(K)$ avec K compact dans Ω, pour tout tel compact nous avons un ouvert $\mathcal{O}_K$ dans $\mathscr{D}(K)$ tel que $f^{-1}(I) \cap \mathscr{D}(K) = \mathcal{O}_K$. En tant qu'union d'ouverts [5], l'ensemble

$$\mathcal{O} = \bigcup_{K \text{ compact de } \Omega} \mathcal{O}_K \tag{32.20}$$

est ouvert dans $C^\infty(\Omega)$. Si $\phi \inf^{-1}(I)$, nous avons $\phi \in \mathscr{D}(K)$ pour un certain K compact de Ω, donc $f^{-1}(I) \subset \mathcal{O}$. A forciori nous avons $f^{-1}(I) \subset \mathcal{O} \cap \mathscr{D}(\Omega)$.

Dans l'autre sens, si $\phi \in \mathcal{O}$, alors ϕ est dans un des $\mathcal{O}_K$ et donc dans $f^{-1}(I)$. Nous avons donc bien $f^{-1}(I) = \mathscr{D}(\Omega) \cap \mathcal{O}$. □

Théorème 32.11 (Convergence dans $\mathscr{D}(\Omega)$[104])**.**
Soit $(\varphi_n)_{n\in\mathbb{N}}$ une suite dans $\mathscr{D}(\Omega)$ et $\varphi \in \mathscr{D}(\Omega)$. Nous avons $\varphi_n \xrightarrow{\mathscr{D}(\Omega)} \varphi$ si et seulement s'il existe K compact dans Ω tel que $\varphi_n \in \mathscr{D}(K)$ pour tout n et $\varphi_n \xrightarrow{\mathscr{D}(K)} \varphi$.

Démonstration. Supposons que $\varphi_n \xrightarrow{\mathscr{D}(\Omega)} \varphi$ et qu'il n'existe pas de compacts contenant tous les supports des φ_n. Alors pour tout compact de Ω il existe un n tel que le support de φ_n ne soit pas dans K. Nous considérons une suite de compacts (K_i) tels que $\text{Int}(K_n) \subset K_{n+1}$ et $\Omega = \bigcup_n K_n$. Une telle suite existe par le lemme 10.59. Ensuite nous construisons des sous-suites de la façon suivante. D'abord $L_1 = K_1$ et $n_1 \in \mathbb{N}$ est choisi de telle sorte que φ_{n_1} ait un support non contenu dans L_1. Ensuite L_i est un compact de la suite (K_n) choisi plus loin que L_{i-1} et tel que $\varphi_{n_{i-1}} \in \mathscr{D}(L_i)$.

5. Voir définition 8.1.

Le nombre n_i est alors choisitplus grand que n_{i-1} de telle sorte que $\varphi_{n_i} \notin \mathscr{D}(L_i)$. Ce faisant, en posant $\phi_i = \varphi_{n_i}$ nous avons

$$\phi_i \in \mathscr{D}(L_{i+1}) \backslash \mathscr{D}(L_i) \tag{32.21}$$

et $\text{Int}(L_n) \subset L_{n+1}$ et $\Omega = \bigcup_n L_n$. Étant donné que (ϕ_i) et une sous-suite de (φ_i) nous avons encore $\phi_i \xrightarrow{\mathscr{D}(\Omega)} \varphi$.

Soit $i \in \mathbb{N}$. Nous allons utiliser le résultat 28.83, aka la seconde forme géométrique du théorème de Hahn-Banach pour séparer les parties $\{\phi_i\}$ (compact) et $\mathscr{D}(L_i)$ (fermé) dans $\mathscr{D}(\Omega)$. Nous avons $f_i \in \mathscr{D}'(\Omega)$ telle que

$$\begin{cases} f_i(\phi_i) > \alpha & \text{(32.22a)} \\ f\big(\mathscr{D}(L_i)\big) < \alpha. & \text{(32.22b)} \end{cases}$$

Nous redéfinissons immédiatement f_i de façon à avoir

$$\begin{cases} f_i(\phi_i) = 0 & \text{(32.23a)} \\ f\big(\mathscr{D}(L_i)\big) < 0. & \text{(32.23b)} \end{cases}$$

Nous introduisons la fonction définie sur $\mathscr{D}(\Omega)$ par

$$p(\phi) = \sum_{i=1}^{\infty} i \frac{f_i(\phi)}{|f_i(\phi_i)|}. \tag{32.24}$$

Si $\phi \in L_k$, alors $f_k(\phi) = 0$ et même $f_l(\phi) = 0$ pour tout $l \geqslant k$. Donc pour chaque k, la somme définissant p est finie sur $\mathscr{D}(L_k)$. Nous en déduisons que p est continue sur chacun des $\mathscr{D}(L_k)$ et donc sur $\mathscr{D}(\Omega)$ par le lemme 32.10.

L'image de la suite convergente $\phi_k \xrightarrow{\mathscr{D}(\Omega)} \varphi$ par p doit être bornée parce que p est continue. Mais dans la somme (32.24), tous les termes sont positifs et en particulier le terme $i = k$ vaut k, donc $p(\phi_k) \geqslant k$, ce qui contredit le fait que l'image de la suite soit bornée. Nous en déduisons donc l'existence d'un compact K tel que $\varphi_n \in \mathscr{D}(K)$ pour tout n.

Nous devons encore prouver que $\varphi_n \xrightarrow{\mathscr{D}(K)} \varphi$ pour ce choix de K. Vu que $\varphi_n \xrightarrow{\mathscr{D}(\Omega)} \varphi$, le lemme 8.28 nous dit que nous avons aussi $\varphi_n \xrightarrow{C^\infty(\Omega)} \varphi$, ce qui signifie que pour tout K et m nous avons

$$p_{K,m}(\varphi_n - \varphi) \to 0. \tag{32.25}$$

En particulier pour le K fixé plus haut nous avons $p_m(\varphi_n - \varphi) \to 0$, c'est à dire que $\varphi_n \xrightarrow{\mathscr{D}(K)} \varphi$. □

Proposition 32.12.
Si K est compact dans Ω, l'espace $\mathscr{D}(K)$ est métrique et complet.

Démonstration. Nous allons d'abord montrer que $\mathscr{D}(K)$ est complet. Ensuite nous allons montrer que sa topologie peut être donnée par une distance.

Complet Nous considérons une suite de Cauchy (φ_n) dans $\mathscr{D}(K)$ au sens de la définition 10.19. Soient $\epsilon > 0$ et $i \in \mathbb{N}$; si k et l sont assez grands nous avons

$$\varphi_k - \varphi_l \in B_i(0, \epsilon). \tag{32.26}$$

En particulier pour $i = 0$ nous avons l'inégalité

$$\|\varphi_k - \varphi_l\|_\infty \leqslant \epsilon, \tag{32.27}$$

La suite (φ_n) est donc de Cauchy dans $\big(C(K), \|.\|_\infty\big)$ et y converge donc par complétude, proposition 14.316. Il existe donc une fonction $\varphi \in C(K)$ telle que

$$\varphi_n \xrightarrow{unif} \varphi. \tag{32.28}$$

Notre jeu à présent est de prouver que $\varphi \in \mathscr{D}(K)$, c'est à dire qu'elle est de classe C^∞. Soit un multiindice $\alpha = \mu_1, \dots, \mu_n, i$. Si k et l sont assez grands nous avons

$$\|\partial^\alpha(\varphi_k - \varphi_l)\|_\infty \leqslant \epsilon, \tag{32.29}$$

c'est à dire que

$$\|\partial_i(\partial^\mu\varphi_k) - \partial_i(\partial^\mu\varphi_l)\|_\infty \leqslant \epsilon. \tag{32.30}$$

Si nous notons $\psi_k = \partial^\mu\varphi_k$ cela signifie que $(\partial_i\psi_n)$ est une suite de Cauchy dans $\big(C(K), \|.\|_\infty\big)$. Elle y converge donc et il existe une fonction $g_i \in C(K)$ telle que

$$\partial_i\psi_n \xrightarrow{unif} g_i. \tag{32.31}$$

Dans ce cas le théorème 16.220 nous indique que ψ est de classe C^1, c'est à dire que $\varphi \in C^{n+1}(K)$.

Métrique La proposition 10.79 nous dit que la topologie donnée par l'écart

$$d(\varphi_1, \varphi_2) = \sup_{k\geqslant 1} \min\{\frac{1}{k}, p_{k-1}(\varphi_1 - \varphi_2)\} \tag{32.32}$$

est la même que celle de $\mathscr{D}(K)$. Il reste à montrer que cette formule est bien une distance au sens de la définition 8.75.

(1) Nous avons bien $d(\varphi_1, \varphi_2) \geqslant 0$ parce que tous les éléments du supremum et du minimum sont positifs.

(2) Si $d(\varphi_1, \varphi_2) = 0$ alors pour tout k nous devons avoir $p_{k-1}(\varphi_1 - \varphi_2) = 0$; en particulier pour $k = 1$ cela donne $\varphi_1 = \varphi_2$.

(3) Nous avons

$$p_k(\varphi_1 - \varphi_2) = p_k\big(-(\varphi_2 - \varphi_1)\big) = p_l(\varphi_2 - \varphi_1) \tag{32.33}$$

en utilisant la propriété (2) de la définition 10.73 de semi-norme.

(4) Nous avons

$$p_k(\varphi_1 - \varphi_2) = p_k(\varphi_1 - \varphi_3 + \varphi_3 - \varphi_2) \leqslant p_k(\varphi_1 - \varphi_3) + p_k(\varphi_3 - \varphi_2) \tag{32.34}$$

en utilisant la propriété (3) de la définition 10.73.

□

Notons que la proposition 10.79 nous dit que $\mathscr{D}(K)$ est complet tout autant pour la topologie des semi-normes que pour celle de la distance que nous venons de décrire. Ces deux topologies sont les mêmes. Étant métrique et complet, l'espace $\mathscr{D}(\Omega)$ et donc de Baire par le théorème 10.86. Ce qui est bien avec ces deux topologies identiques c'est qu'on peut utiliser la propriété de Baire même en ne parlant que des semi-normes.

32.2 Distributions

Si Ω est un ouvert de $\mathbb{R}^d$, alors l'ensemble $\mathscr{D}(\Omega)$ est contenu dans $C^\infty(\Omega)$. Nous allons commencer par définir une topologie sur $C^\infty(\Omega)$ et ensuite donner à $\mathscr{D}(\Omega)$ la topologie induite[6].

Définition 32.13 (Distribution)**.**
Une ***distribution*** *sur un ouvert Ω de $\mathbb{R}^d$ est une forme linéaire continue sur $\mathscr{D}(\Omega) = C_c^\infty(\Omega)$. C'est donc un élément de $\mathscr{D}'(\Omega)$.*

Le théorème suivant donne quelques façons de vérifier qu'une forme linéaire soit continue. En particulier il nous dit que pour prouver qu'une forme linéaire est une distribution il suffi de prouver la continuité séquentielle.

6. Définition 8.10.

Théorème 32.14 ([104, 372]).
Soit T une forme linéaire sur $\mathscr{D}(\Omega)$. Nous avons équivalence entre les points suivants.

(1) T est continue.

(2) Pour tout compact $K \subset \Omega$ il existe $m \in \mathbb{N}$ et $C \geqslant 0$ tel que pour tout $\varphi \in \mathscr{D}(K)$ nous ayons

$$\big|T(\varphi)\big| \leqslant C p_{m,K}(\varphi) \tag{32.35}$$

où $p_{m,K}$ est la semi-norme donnée en (32.15).

(3) T est séquentiellement continue sur $\mathscr{D}(\Omega)$.

(4) T est séquentiellement continue en 0.

(5) Pour tout compact $K \subset \Omega$, la restriction de T à $\mathscr{D}(K)$ est continue.

Un certain nombre d'ouvrages prennent le point (2) comme la définition d'une distribution.

Définition 32.15 (Topologie sur $\mathscr{D}'(\Omega)$).
*Nous munissons l'espace $\mathscr{D}'(\Omega)$ de la **topologie $*$-faible**, c'est à dire celle de la famille de semi-normes*

$$\begin{aligned} p_\varphi \colon \mathscr{D}'(\Omega) &\to \mathbb{R} \\ T &\mapsto \big|T(\varphi)\big| \end{aligned} \tag{32.36}$$

avec $\varphi \in \mathscr{D}(\Omega)$.

Oui, c'est bien une famille de semi-normes indicée par l'ensemble $\mathscr{D}(\Omega)$. Il n'y en a donc a priori pas du tout une quantité dénombrable.

Proposition 32.16 (Convergence au sens des distributions).
Nous avons $T_n \xrightarrow{\mathscr{D}'(\Omega)} T$ si et seulement si $T_n(\varphi) \to T(\varphi)$ pour tout $\varphi \in \mathscr{D}(\Omega)$.

Démonstration. La convergence $T_n \xrightarrow{\mathscr{D}'(\Omega)} T$ signifie que l'on ait $p_\varphi(T_n - T) \to 0$ pour tout $\varphi \in \mathscr{D}(\Omega)$, ce qui en retour signifie que

$$\big|(T_n - T)(\varphi)\big| \to 0. \tag{32.37}$$

□

Cette proposition suppose que l'on ait une distribution T qui vérifie $T_n(\varphi) \to T(\varphi)$ et conclut qu'on a une convergence dans les distributions. Le théorème suivant est plus fort : il va seulement supposer que $T_n(\varphi)$ converge dans $\mathbb{C}$ et va conclure que $T \colon \varphi \mapsto \lim_{n\to\infty} T_n(\varphi)$ est une distribution.

Théorème 32.17 ([373]).
Soit (T_n) une suite dans $\mathscr{D}'(\Omega)$ et nous supposons que pour tout $\varphi \in \mathscr{D}(\Omega)$ la suite $\big(T_n(\varphi)\big)$ converge dans $\mathbb{C}$. Alors il existe $T \in \mathscr{D}'(\Omega)$ telle que $T_n \xrightarrow{\mathscr{D}'(\Omega)} T$.

Proposition 32.18.
L'application

$$\begin{aligned} i \colon L^2(\Omega) &\to \mathscr{D}'(\Omega) \\ f &\mapsto T_f \end{aligned} \tag{32.38}$$

est une injection continue.

Démonstration. Le fait que ce soit une injection est le fait que si $T_f = T_g$ alors pour tout $\phi \in \mathscr{D}(\Omega)$ nous avons $\langle f - g, \phi \rangle = 0$, et cela implique que $f - g$ est nulle presque partout en tant que fonction et est simplement nulle en tant que classe de fonction dans L^2.

En ce qui concerne la continuité, il suffit de la prouver en zéro (par linéarité). Soit donc $f_n \xrightarrow{L^2} 0$ et montrons que $T_{f_n} \xrightarrow{\mathscr{D}'(\Omega)} T_0$. Pour prouver cela, la proposition 32.16 nous indique qu'il est suffisant de tester $T_n(\phi) \to 0$ pour tout $\phi \in \mathscr{D}(\Omega)$.

Notons que si $\phi \in \mathscr{D}$ a fortiori $\phi \in L^2$. Nous avons

$$T_{f_n}(\phi) = \int_\Omega f_n\phi \leqslant \|f_n\phi\|_{L^1} \leqslant \|f_n\|_2\|\phi\|_2 \to 0 \tag{32.39}$$

où nous avons utilisé l'inégalité de Hölder de la proposition 28.23. □

Cette proposition permet de donner un sens à des phrases du type « Soit une distribution T. Si $T \in L^2$, alors … ». Cela signifie qu'il existe $u \in L^2$ tel que $T = T_u$. Notons que dans ce cas, la distribution est définie sur L^2 et non seulement sur $\mathscr{D}$.

32.2.1 Multiplication d'une distribution par une fonction

Définition 32.19.
Si $T \in \mathscr{D}'(\Omega)$ et si $f \in C^\infty(\Omega)$ nous définissons la distribution fT par

$$(fT)(\varphi) = T(f\varphi). \tag{32.40}$$

Souvent écrit sous la forme plus compacte $\langle fu, \phi\rangle = \langle u, f\phi\rangle$.

Cela a un sens parce que si $\varphi \in \mathscr{D}(\Omega)$ alors $f\varphi$ est aussi dans $\mathscr{D}(\Omega)$.

Cette définition est motivée par ce que l'on ferait pour une distribution à densité. Si T est une distribution de densité notée également T, nous avons $T(\phi) = \int T(x)\phi(x)$ et donc

$$(fT)(\phi) = \int (fT)(x)\phi(x) = \int T(x)f(x)\phi(x) = \int T(x)(f\phi)(x) = T(f\phi). \tag{32.41}$$

En ce qui concerne les distributions tempérées, nous pouvons définir le produit avec une fonction $f \in \mathscr{S}(\Omega)$ par la même formule : si $f, \varphi \in \mathscr{S}(\Omega)$ alors le produit $f\varphi$ est encore Schwartz. Notons toutefois que nous ne pouvons pas définir fT dans $\mathscr{S}'(\Omega)$ si f est seulement dans $C^\infty(\Omega)$.

32.2.2 Dérivée de distribution

Proposition-définition 32.20.
Soit T une distribution sur Ω et $\alpha \in \mathbb{N}^d$. Alors la formule

$$(\partial^\alpha T)(\varphi) = (-1)^{|\alpha|}T(\partial^\alpha\varphi) \tag{32.42}$$

définit une distribution $\partial^\alpha T$.

*Cette distribution $\partial^\alpha T$ sera la **dérivée distributionnelle** de T. Notons que le même résultat est encore valide pour des distributions tempérées, et la démonstration est la même.*

Démonstration. La forme linéaire $\partial^\alpha T$ sera continue si elle est séquentiellement continue par le théorème 32.14. Nous considérons donc une suite $\varphi_n \xrightarrow{\mathscr{D}(\Omega)} \varphi$ et nous vérifions que

$$\lim_{n\to\infty}(\partial^\alpha T)(\varphi_n) = (\partial^\alpha T)(\varphi). \tag{32.43}$$

D'abord T étant une distribution (et donc continue) nous pouvons la permuter avec la limite :

$$\lim_{n\to\infty}(\partial^\alpha T)(\varphi_n) = \lim_{n\to\infty}(-1)^{|\alpha|}T(\partial^\alpha\varphi_n) = (-1)^{|\alpha|}T\Big(\lim_{n\to\infty}\partial^\alpha\varphi_n\Big). \tag{32.44}$$

Notons qu'à gauche la limite est une limite dans $\mathbb{R}$ tandis qu'à droite c'est une limite dans $\mathscr{D}(\Omega)$. Ensuite le lemme 32.9 nous dit que l'hypothèse $\varphi_n \xrightarrow{\mathscr{D}(\Omega)} \varphi$ signifie en particulier que nous avons un compact $K \subset \Omega$ contenant tous les supports des φ_n et que $\partial^\alpha\varphi_n$ converge uniformément (sur K et donc sur Ω) vers $\partial^\alpha\varphi$. Donc

$$\lim_{n\to\infty}(\partial^\alpha T)(\varphi_n) = (-1)^{|\alpha|}T\Big(\lim_{n\to\infty}\partial^\alpha\varphi_n\Big) = (-1)^{|\alpha|}T\big(\partial^\alpha\varphi\big) = (\partial^\alpha T)(\varphi), \tag{32.45}$$

ce qui est la relation demandée. □

Le lemme suivant montre une compatibilité entre la dérivée des distributions, la dérivée faible et l'injection de L^2 dans l'espace des distributions.

Lemme 32.21.
Soit Ω un ouvert bornée de $\mathbb{R}^n$ et $f \in L^2(\Omega)$. Alors nous avons

$$\partial_i(T_f) = T_{\partial_i f} \tag{32.46}$$

où la dérivée à droite est la dérivée faible définie en 32.6.

Démonstration. En utilisant la définition de la dérivation de distribution, pour tout $\phi \in \mathscr{D}$ nous avons

$$\partial_i(T_f)\phi = -T_f(\partial_i\phi) = -\langle f, \partial_i\phi\rangle = \langle \partial_i f, \phi\rangle = T_{\partial_i f}(\phi). \tag{32.47}$$

Nous avons utilisé la définition (32.14) de la dérivée faible. □

32.2.3 Ordre et support d'une distribution

Définition 32.22 (support d'une distribution[6])**.**
*Soit T une distribution. Le **support** de T est le complémentaire de l'union des ouverts $\mathcal{O}$ tels que $T(\varphi) = 0$ pour tout φ à support dans $\mathcal{O}$.*

Définition 32.23.
*Si T est une distribution sur Ω, nous disons que T est d'**ordre** inférieur ou égal à $p \in \mathbb{N}$ si pour tout compact K de Ω, il existe $C_K \in \mathbb{R}$ tel que pour tout $\varphi \in \mathscr{D}(K)$,*

$$\big|\langle T, \varphi\rangle\big| \leqslant C_K \max_{|\alpha|\leqslant p} \|\partial^\alpha \varphi\|_\infty. \tag{32.48}$$

Ici α est un multiindice.

La distribution T est d'ordre p si elle est d'ordre inférieur ou égal à p mais pas à $p-1$.

Pour la proposition suivante, on peut se remémorer la définition 32.8 de la topologie sur $C^\infty(\Omega)$.

Proposition 32.24 ([374])**.**
Restriction entre C^∞ et $\mathscr{D}$.

(1) *Si $T \in C^\infty(\Omega)'$, alors la restriction de T à $\mathscr{D}(\Omega)$ est une distribution à support[7] compact.*
(2) *Si T est une distribution à support compact alors elle se prolonge de façon unique en une forme linéaire continue sur $C^\infty(\Omega)$.*

Proposition 32.25 ([104])**.**
Une distribution à support compact est d'ordre fini.

Lemme 32.26 ([144])**.**
Soit $u \in \mathscr{D}'(\mathbb{R})$ et $\phi \in \mathscr{D}(\mathbb{R})$ tels que $\operatorname{supp}(u) \cap \operatorname{supp}(\phi) = \varnothing$. Alors $\langle u, \phi\rangle = 0$.

Démonstration. Soit $x \notin \operatorname{supp}(u)$. Alors il existe un voisinage V_x de x tel que $\langle u, \psi\rangle = 0$ pour tout $\psi \in \mathscr{D}(V_x)$. En particulier, si $x \in \operatorname{supp}(\phi)$, alors x n'est pas dans le support de u et les ensembles $\{V_x$ tel que $x \in \operatorname{supp}(\phi)\}$ recouvrent $\operatorname{supp}(\phi)$. Cependant ϕ est à support compact et nous pouvons extraire un sous-recouvrement fini de $\operatorname{supp}(\phi)$: il existe $x_1, \dots, x_p$ tels que

$$\operatorname{supp}(\phi) \subset \bigcup_{i=1}^{p} V_{x_i}. \tag{32.49}$$

7. Définition 32.22.

Nous prenons une partition de l'unité[8] subordonnée à ce recouvrement. C'est à dire des fonctions $\chi_i \in \mathscr{D}(V_{x_i})$ telles que pour tout $x \in \text{supp}(\phi)$,

$$\sum_{i=1}^{p} \chi_i(x) = 1. \tag{32.50}$$

En particulier nous avons $\sum_i \chi_i(x)\phi(x) = \phi(x)$, et donc

$$\langle u, \phi\rangle = \langle u, \sum \chi_i\phi\rangle = \sum \langle u, \chi_i\phi\rangle = 0 \tag{32.51}$$

parce que $\text{supp}(\chi_i\phi) \subset V_{x_i}$. □

Lemme 32.27 ([144]).
Si u est une distribution d'ordre fini N sur $\mathbb{R}$, si $\text{supp}(u) = \{x_0\}$ et si

$$\phi(x_0) = \ldots = \phi^{(N)}(x_0) = 0 \tag{32.52}$$

alors $\langle u, \phi\rangle = 0$.

Démonstration. Les fonctions plateaux dont nous avons parlé dans la section 17.14.1 nous permettent de considérer une fonction $\chi \in \mathscr{D}(\mathbb{R})$ vérifiant

$$\chi(x) = \begin{cases} 1 & \text{si } x \in \overline{B(0,1)} \\ 0 & \text{si } |x| > 2 \end{cases} \tag{32.53}$$

Ensuite nous posons $\chi_n(x) = \chi\big(n(x-x_0)\big)$. Par conséquent $\chi_n(x_0) = \chi(0) = 1$ et même $\chi_n(x_0+\epsilon) = \chi(\epsilon) = 1$ tant que ϵ est plus petit que disons $\frac{1}{2}$ pour être sur. Nous en déduisons que la fonction $1-\chi_n$ s'annule sur un voisinage de x_0 et que donc x_0 n'est pas dans le support de $1-\chi_n$. Donc $\text{supp}(1-\chi_n) \cap \text{supp}(u) = \varnothing$ et le lemme 32.26 est utilisable : $\langle u, (1-\chi_n\phi)\rangle = 0$, ou encore :

$$\langle u, \phi\rangle = \langle u, \chi_n\phi\rangle \tag{32.54}$$

pour tout n. Vu que le but est de prouver que $\langle u, \phi\rangle = 0$, nous allons prouver que

$$|\langle u, \chi_n\phi\rangle| \overset{n\to\infty}{\longrightarrow} 0. \tag{32.55}$$

Dans ce dessin nous posons

$$\|f\|_n = \sup_{x\in\overline{B(0,\frac{2}{n})}} \|f(x)\| \tag{32.56}$$

et

$$\|f\|_{(p)} = \sup_{i\leqslant p} \|\partial^i f\|_\infty. \tag{32.57}$$

La distribution u est d'ordre fini N, et nous en écrivons la définition 32.23 en prenant $\text{supp}(\chi_n\phi)$ en tant que K :

$$\big|\langle u, \chi_n\phi\rangle\big| \leqslant C \max_{k\leqslant N} \|\partial^k(\chi_n\phi)\|_\infty. \tag{32.58}$$

En remplaçant le maximum par une somme de $k=0$ à $k=N$, nous majorons. De plus le support de χ_n étant contenu dans $B_n = B(x_0, 2/n)$ nous ne changeons rient en utilisant $\|.\|_n$ au lieu de $\|.\|_\infty$. Donc

$$|\langle u, \chi_n\phi\rangle| \leqslant C \sum_{k=0}^{N} \|\partial^k(\chi_n\phi)\|_n \leqslant C \sum_{k=0}^{N} \binom{k}{i} \|\partial^i\chi_n\|_n \|\partial^{k-i}\phi\|_n. \tag{32.59}$$

Notons que la seconde inégalité est une inégalité du type $\|fg\| \leqslant \|f\|\|g\|$. En dérivant un petit peu nous trouvons que

$$(\partial^i\chi_n)(x) = n^i(\partial^i\chi)\big(n(x-x_0)\big). \tag{32.60}$$

8. Lemme 22.20.

Donc [9]

$$\|\partial^i\chi_n\|_n = \sup_{x\in B_n} n^i\big|(\partial^i\chi)\big(n(x-x_0)\big)\big| = n^i \sup_{y\in[-2,2]}\big|(\partial^i\chi)(y)\big| = n^i\|\partial^i\chi\|_\infty. \tag{32.61}$$

Nous pouvons donc remplacer $\|\partial^u\chi_n\|_n$ par $n^i\|\partial^i\chi\|_\infty$.

D'autre part nous voulons majorer $\|\partial^{k-i}\phi\|_n$ par quelque chose ne dépendant ni de k ni de i. Nous faisons le théorème des accroissements finis 12.123 : $\|\partial^l\phi\|_n \leqslant \frac{2}{n}\|\partial^{l+1}\phi\|_n$. Ce n au dénominateur est salutaire parce que nous avions un n^i apparu à cause du remplacement (32.61). Nous faisons donc $i+1$ fois le théorème des accroissements finis :

$$\|\partial^{k-i}\phi\|_n \leqslant \left(\frac{2}{n}\right)^{i+1}\|\partial^{k+1}\phi\|_n. \tag{32.62}$$

Toutes ces majorations donnent

$$\big|\langle u,\chi_n\phi\rangle\big| \leqslant C\sum_{k=0}^{N}\sum_{i=0}^{k}\binom{i}{k}n^i\underbrace{\|\partial^i\chi\|_\infty}_{\leqslant\|\chi\|_{(N)}}\left(\frac{2}{n}\right)^{i+1}\underbrace{\|\partial^{k+1}\phi\|_n}_{\leqslant\|\phi\|_{(N+1)}} \tag{32.63a}$$

$$\leqslant C\|\chi\|_{(N)}\|\phi\|_{(N+1)}\frac{1}{n}\sum_{k=0}^{N}\sum_{i=0}^{k}\binom{k}{i}2^{i+1} \tag{32.63b}$$

$$= \frac{C'}{n} \tag{32.63c}$$

où C' est une constante qui dépend de χ, de ϕ et de N, mais pas de n. Vu que $\frac{C'}{n}\to 0$ nous avons bien

$$\langle u,\phi\chi_n\rangle = 0, \tag{32.64}$$

ce qu'il fallait démontrer. □

Proposition 32.28 ([144]).
Soit $u\in\mathscr{D}(\mathbb{R})'$ avec $\operatorname{supp}(u)=\{x_0\}$. Alors $u=\sum_{i=0}^{N}a_i\partial^i\delta_{x_0}$ où N est l'ordre de u.

Démonstration. D'abord il faut préciser que l'ordre de u est fini parce que son support est compact (proposition 32.25) ; nous notons N cet ordre.

Soit $\phi\in\mathscr{D}(\mathbb{R})$. Nous considérons $\chi\in\mathscr{D}(\mathbb{R})$ telle que

$$\chi(x) = \begin{cases} 1 & \text{si } x\in\overline{B(x_0,1)}\\ 0 & \text{si } |x-x_0|>2. \end{cases} \tag{32.65}$$

Encore une fois, $1-\chi$ s'annule sur un voisinage autour de x_0, ce qui fait que

$$\operatorname{supp}(u)\cap\operatorname{supp}\big((1-\chi)\phi\big) = \varnothing, \tag{32.66}$$

et donc $\langle u,(1-\chi)\phi\rangle = 0$. Au final,

$$\langle u,\phi\rangle = \langle u,\chi\phi\rangle. \tag{32.67}$$

C'est le moment de poser

$$\psi(x) = \chi(x)\big[\phi(x) - \sum_{k=1}^{N}\frac{1}{k!}(\partial^k\phi)(x_0)(x-x_0)^k\big] \tag{32.68}$$

La fonction ψ ayant un support disjoint de celui de u, nous avons aussi $\langle u,\psi\rangle=0$, ce qui donne

$$\langle u,\phi\rangle = \langle u,\chi\phi\rangle = \langle u,\chi\sum_{k=0}^{N}\frac{1}{k!}(\partial^k\phi)(x_0)(x-x_0)^k\rangle. \tag{32.69}$$

9. Dans [144], la dernière égalité vient avec une inégalité, et je comprends pas pourquoi.

En posant $a_k = \frac{1}{k!}\langle u, x \mapsto \chi(x)(x-x_0)^k\rangle$ nous avons alors

$$\langle u, \phi\rangle = \sum_{k=0}^{N} a_k(\partial^k\phi)(x_0) = \sum_k (-1)^k a_k(\partial^k\delta_{x_0})(\phi). \tag{32.70}$$

□

32.3 Distributions tempérées

L'espace de Schwartz[10] $\mathscr{S}(\Omega)$ est défini dans la définition 28.87 ; sa topologie y est discutée.

Définition 32.29.
*Une **distribution tempérée** est une forme linéaire continue sur $\mathscr{S}(\mathbb{R}^d)$. L'ensemble des distributions tempérées est noté $\mathscr{S}'(\mathbb{R}^d)$. Si T est une telle distribution, nous notons $\langle T, \varphi\rangle$ l'image de φ par T.*

32.30.
Pour rappel, la topologique sur $\mathscr{S}(\mathbb{R}^d)$ est celle de la définition 28.92.

Si f est une fonction sur $\mathbb{R}^d$ telle que $f\varphi \in L^1(\mathbb{R}^d)$ pour tout $\varphi \in \mathscr{S}(\mathbb{R}^d)$, alors nous définissons la distribution $T_f \in \mathscr{S}'(\mathbb{R}^d)$ par

$$\langle T_f, \varphi\rangle = \int_{\mathbb{R}^d} f(x)\varphi(x)dx. \tag{32.71}$$

Cette définition ne fonctionne pas pour toutes les fonctions. Par exemple pour $f(x) = e^{x^2}$, et $\varphi(x) = e^{-x^2} \in \mathscr{S}(\mathbb{R})$ nous avons $f\varphi = 1$ qui n'est pas du tout intégrable sur $\mathbb{R}$.

Lemme-définition 32.31.
L'application

$$\begin{aligned}\delta\colon \operatorname{Fun}(\mathbb{R}^d) &\to \mathbb{C}\\ \varphi &\mapsto \varphi(0).\end{aligned} \tag{32.72}$$

est

(1) une distribution,

(2) une distribution tempérée.

*Cette distribution est nommée **distribution de Dirac**.*

Démonstration. Juste pour rappel, $\operatorname{Fun}(X)$ est l'ensemble de toutes les fonctions sur X. Pour prouver que δ est une distribution, nous devons démontrer que $\delta\colon \mathscr{D}(\mathbb{R}) \to \mathbb{C}$ est continue. Et pour qu'elle soit une distribution tempérée, il faut démontrer que $\delta\colon \mathscr{S}(\mathbb{R}) \to \mathbb{C}$ est continue.

Nous utilisons le théorème 32.14(2) pour démontrer la continuité de δ sur $\mathscr{D}(\mathbb{R})$. Soit un compact $K \subset \mathbb{R}$ et $\varphi \in \mathscr{D}(K)$. En prenant $m = 0$ nous devons avons la majoration

$$|\delta(\varphi)| = \varphi(0) \leqslant \|\varphi\|_{K,\infty} = p_{0,K}(\varphi). \tag{32.73}$$

Pour la continuité de δ sur $\mathscr{S}(\mathbb{R}^d)$, nous utilisons les résultats de 28.93. Soit une suite $\varphi_n \xrightarrow{\mathscr{S}} 0$. En particulier, $p_{0,0}(\varphi_n) = \sup_x |\varphi_n(x)| \to 0$. Donc $\varphi_n(0) \to 0$ comme il le faut. □

Exemple 32.32
La **valeur principale** de la fonction $x \mapsto \frac{1}{x}$ est la distribution

$$\begin{aligned}T\colon \mathscr{S}(\mathbb{R}) &\to \mathbb{R}\\ \varphi &\mapsto \lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \int_{|x|>\epsilon} \frac{\varphi(x)}{x}.\end{aligned} \tag{32.74}$$

10. Attention : ce Schwartz (avec un t) est le Schwartz des distributions dont le prénom est Laurent. À ne pas confondre avec Schwarz (sans t) dont le prénom est Cauchy.

Montrons que cela définit bien une distribution tempérée.

D'abord l'intégrale existe pour tout ϵ, par exemple parce que pour les grands $|x|$ nous avons par exemple $|\varphi(x) \leqslant x^3|$ et donc $\varphi(x)/x \leqslant 1/x^2$ dont l'intégrale converge. Nous devons maintenant regarder la limite.

Nous considérons une suite $\epsilon_n \to 0$ et la suite

$$a_n = \int_{|x|\geqslant\epsilon_n} \frac{\varphi(x)}{x} dx. \tag{32.75}$$

Nous montrons que cette suite converge dans $\mathbb{R}$ en montrant qu'elle est de Cauchy. Pour cela nous travaillons un peu la forme de φ :

$$\varphi(x) = \varphi(0) + \int_0^x \varphi'(t)dt = \varphi(0) + \int_0^1 x\varphi'(x\theta)d\theta. \tag{32.76}$$

Ce qui est dans l'intégrale est borné par $K = \|M_x\varphi'\|_\infty$ qui est parfaitement fini parce que φ est à décroissance rapide. Lorsque nous calculons $|a_m - a_n|$, le terme $\varphi(0)/x$ donne une intégrale nulle parce que le domaine d'intégration $\epsilon_n \leqslant |x| \leqslant \epsilon_n$ est symétrique alors que la fonction $1/x$ est impaire.

$$|a_m - a_n| \leqslant \Big| \int_{\epsilon_m<|x|<\epsilon_n} K \Big| = 2|\epsilon_n - \epsilon_m|K \tag{32.77}$$

Tout cela nous dit que T est bien définie. Nous devons encore étudier sa continuité.

Soit χ une fonction dans $C_c^\infty(\mathbb{R})$ telle valant 1 sur $[-1,1]$, paire et à valeurs dans $[0,1]$. Pour tout $\epsilon > 0$ nous avons $\int_{|x|>\epsilon} \frac{\chi(x)}{x} dx = 0$.

Nous avons aussi $\varphi = \chi\varphi + (1-\chi)\varphi$, et donc

$$\begin{aligned}
\int_{|x|>\epsilon} \frac{\varphi(x)}{x} dx &= \int_{|\epsilon|>0} \chi(x) \frac{\varphi(x) - \varphi(0)}{x} dx + \int_{|\epsilon|>0} \big(1-\chi(x)\big) \frac{\varphi(x)}{x} dx && \text{(32.78a)}\\
&= \int_{|\epsilon|>0} \chi(x) \int_0^1 \underbrace{\varphi'(\theta x)}_{\leqslant\|\varphi'\|_\infty} d\theta + \int_{|x|\geqslant 1} \big(1-\chi(x)\big) \frac{\varphi(x)}{x} dx && \text{(32.78b)}\\
&\leqslant \|\varphi'\|_\infty \int_{|x|\geqslant\epsilon} \chi(x)dx + \|\varphi\|_{L^1} && \text{(32.78c)}\\
&= C\|\varphi'\|_\infty + \|\varphi\|_1. && \text{(32.78d)}
\end{aligned}$$

Cela est valable pour toute fonction $\varphi \in \mathscr{S}(\mathbb{R})$. Mais nous savons que si $\varphi_n \overset{\mathscr{S}(\mathbb{R})}{\to} 0$, alors $\|\varphi_n\|_\infty \to 0$, $\|\varphi_n'\|_\infty \to 0$ et $\|\varphi_n\|_1 \to 0$; donc si $\varphi_n \overset{\mathscr{S}(\mathbb{R})}{\to} 0$, alors

$$T(\varphi_n) = \lim_{\substack{\epsilon\to 0\\ \epsilon>0}} \int_{|x|>\epsilon} \frac{\varphi(x)}{x} \leqslant C\|\varphi_n'\|_\infty + \|\varphi_n\|_1 \to 0. \tag{32.79}$$

△

32.3.1 Topologie

La topologie que nous mettons sur l'espace $\mathscr{S}'(\mathbb{R}^d)$ est le même type que celle que nous mettons sur $\mathscr{D}'(\mathbb{R}^d)$, c'est à dire celle des semi-normes $p_\varphi(T) = |T(\varphi)|$. La définition 32.15 et la proposition 32.16 restent.

Proposition 32.33.
Nous avons $T_n \overset{\mathscr{S}'(\mathbb{R}^d)}{\longrightarrow} T$ si et seulement si pour tout $\varphi \in \mathscr{S}(\mathbb{R}^d)$ nous avons $T_n(\varphi) \to T(\varphi)$.

32.3.2 Distributions associées à des fonctions

Si $f \in L^1_{loc}(\mathbb{R}^d)$ alors nous lui associons une distribution $T_f \in \mathscr{D}'(\mathbb{R}^d)$ définie par la formule

$$T_f(\varphi) = \int_{\mathbb{R}^d} f(x)\varphi(x)dx. \tag{32.80}$$

Proposition 32.34.
L'application ainsi définie

$$\begin{aligned} L^1_{loc}(\mathbb{R}^d) &\to \mathscr{D}'(\mathbb{R}^d) \\ f &\mapsto T_f \end{aligned} \tag{32.81}$$

est injective.

Démonstration. Si $T_f = 0$ alors pour tout $\varphi \in \mathscr{D}$ nous avons $\int_{\mathbb{R}^d} f\varphi = 0$. En vertu de la proposition 28.77 cela implique $f = 0$ presque partout. □

32.3.3 Composition avec une fonction

Proposition 32.35 ([373], page 113 et 32)**.**
Soit $T \in \mathscr{S}'(\Omega)$ et $f \in C^k(A \times \Omega)$ où A est ouvert dans $\mathbb{R}^d$. Nous posons

$$\begin{aligned} F\colon A &\to \mathbb{R} \\ \lambda &\mapsto T\big(f(\lambda, .)\big). \end{aligned} \tag{32.82}$$

Alors $F \in C^k(A)$.

32.3.4 Transformée de Fourier d'une distribution tempérée

Définition 32.36.
*La **transformée de Fourier** de la distribution tempérée $T \in \mathscr{S}'(\mathbb{R}^d)$ est la distribution $\hat{T}$ définie par*

$$\hat{T}(\varphi) = T(\hat{\varphi}) \tag{32.83}$$

pour tout $\varphi \in \mathscr{S}(\mathbb{R}^d)$.

Lemme 32.37.
Si $f \in \mathscr{S}(\mathbb{R}^d)$, nous avons $\hat{T}_f = T_{\hat{f}}$.

Démonstration. En utilisant les définitions,

$$\hat{T}_f(\varphi) = T_f(\hat{\varphi}) = \int_{\mathbb{R}^d} f(x)\hat{\varphi}(x)dx = \int_{\mathbb{R}^d} f(x)\left[\int_{\mathbb{R}^d} \varphi(y)e^{-iyx}dy\right]dx \tag{32.84}$$

où nous avons noté xy le produit scalaire $x \cdot y$. Nous permutons les intégrales en utilisant le théorème de Fubini 16.239 avec la fonction

$$(x, y) \mapsto f(x)\varphi(y)e^{-ixy} \tag{32.85}$$

qui est parfaitement dans $L^1(\mathbb{R}^d \times \mathbb{R}^d)$. Nous écrivons alors

$$\hat{T}_f(\varphi) = \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d} f(x)\varphi(y)e^{-iyx}dx\right]dy = \int_{\mathbb{R}^d} \varphi(y)\hat{f}(y)dy = T_{\hat{f}}(\varphi). \tag{32.86}$$

□

32.3.5 Convolution d'une distribution par une fonction

Nous savons que si $\psi \in \mathscr{S}(\mathbb{R}^d)$ et si $x \in \mathbb{R}^d$ alors la fonction $y \mapsto \psi(x-y)$ est encore une fonction dans $\mathscr{S}(\mathbb{R}^d)$. Donc si $T \in \mathscr{S}'(\mathbb{R}^d)$ nous pouvons considérer la fonction $T * \psi = \psi * T$ définie par

$$(T * \psi)(x) = T\big(y \mapsto \psi(x-y)\big). \tag{32.87}$$

Notons que $T * \psi$ est bien une fonction et non une distribution.

Le but de la définition est d'avoir

$$T_f * \psi = f * \psi. \tag{32.88}$$

En effet

$$(T_f * \psi)(x) = T_f\big(y \mapsto \psi(x-y)\big) = \int_{\mathbb{R}^d} f(y)\psi(x-y)dy = (f * \psi)(x). \tag{32.89}$$

Exemple 32.38
La distribution de Dirac est le neutre pour le produit de convolution. En effet

$$(\delta * \psi)(x) = \delta\big(y \mapsto \psi(x-y)\big) = \psi(x), \tag{32.90}$$

c'est à dire $\delta * \psi = \psi$. △

Proposition 32.39 ([6]).
Si $T \in \mathscr{S}'(\mathbb{R}^d)$ et $\psi \in \mathscr{S}(\mathbb{R}^d)$, alors la distribution associée à la fonction $T * \psi$ est tempérée.

Démonstration. En agissant sur $\varphi \in \mathscr{D}(\mathbb{R}^d)$ nous avons

$$\begin{aligned}
T_{T*\psi}(\varphi) &= \int_{\mathbb{R}^d} T\big(y \mapsto (t_x\psi)(y)\big)\varphi(x)dx && \text{(32.91a)}\\
&= \int_{\mathbb{R}^d} T\big(y \mapsto \varphi(x)\psi(x-y)\big)dx && \text{(32.91b)}\\
&= T\left(y \mapsto \int_{\mathbb{R}^d} \varphi(x)\psi(x-y)dx\right) && \text{(32.91c)}\\
&= T\big(y \mapsto (\varphi * \check{\psi})(y)\big) && \text{(32.91d)}\\
&= T(\varphi * \check{\psi}). && \text{(32.91e)}
\end{aligned}$$

Attention : **Problèmes et choses à faire**
Le passage à la ligne (32.91c) n'est pas justifié. □

32.3.6 Approximation de la distribution de Dirac

Lemme 32.40 ([1]).
Soient des fonctions $j_n \colon \mathbb{R} \to \mathbb{R}^+$ de classe C^∞ telles que

(1) Pour chaque n, la fonction $x \mapsto j_n\big(|x|\big)$ est strictement décroissante et converge ponctuellement vers zéro.

(2) Pour chaque x, la suite $n \mapsto j_n(x)$ est décroissante et converge vers 0.

(3) Pour tout $M > 0$, la suite j_n converge vers zéro uniformément sur $B(0,M)^c$.

(4) Pour tout δ et ϵ, il existe un $N \in \mathbb{N}$ tel que $|\int_{B(0,\delta)} j_n(x)dx - 1| \leqslant \epsilon$.

(5) Pour tout n, nous avons $\int_{\mathbb{R}} j_n = 1$.

Alors si $u \in \mathscr{S}(\mathbb{R})$ nous avons

$$\lim_{n\to\infty} \int_{\mathbb{R}} u(x)j_n(x)dx = u(0). \tag{32.92}$$

Démonstration. Nous posons

$$I_n = \int_{\mathbb{R}} j_n u \tag{32.93a}$$

$$I_{\delta,n} = \int_{B(0,\delta)} j_n u \tag{32.93b}$$

$$Z_{\delta,n} = \int_{B(0,\delta)} u(0) j_n \tag{32.93c}$$

Nous allons progressivement montrer qu'en prenant δ assez petit et n assez grand, les quantités $|I_n - I_{\delta,n}|$, $|I_{\delta,n} - Z_{\delta,n}|$ et $|Z_{\delta,n} - u(0)|$ peuvent être simultanément majorées par ϵ.

Soient $\delta > 0$ et $\epsilon > 0$; vu que $u \in \mathscr{S}(\mathbb{R})$, il existe M tel que $\int_{|x|>M} |u| < \epsilon$. Soit $N_1 \in \mathbb{N}$ tel que pour tout $n > N_1$ nous avons $|j_n(x)| < 1$ dès que $|x| > M$ (hypothèse (3)). Alors

$$\int_{|x|>M} |j_n(x)u(x)| < \epsilon. \tag{32.94}$$

De plus en posant $s = \max\{|u(x)| \text{ tel que } \delta \leqslant |x| \leqslant M\}$ (qui existe parce que u est continue et prise sur un compact) nous pouvons considérer N_2 tel que $j_n(x) < \epsilon/s$ pour tout $|x| > \delta$.

Avec $n > \max\{N_1, N_2\}$ nous avons

$$|\int_{B(0,\delta)} j_n u - \int_{\mathbb{R}} j_n u| = |\int_{B(0,\delta)^c} j_n u| \tag{32.95a}$$

$$\leqslant \int_{\delta\leqslant|x|\leqslant M} |j_n u| + \int_{|x|\geqslant M} |j\nu| \tag{32.95b}$$

$$\leqslant \epsilon(1 + |M - \delta|). \tag{32.95c}$$

En redéfinissant le ϵ nous avons donc montré que pour tout ϵ et δ, il existe un $N \in \mathbb{N}$ tel que

$$|I_{\delta,n} - I_n| \leqslant \epsilon \tag{32.96}$$

dès que $n \geqslant N$.

La fonction u est uniformément continue sur tout $\overline{B(0,\delta)}$, et nous pouvons donc choisir δ tel que $|u(0) - u(x)| \leqslant \epsilon$ pour tout $x \in \overline{B(0,\delta)}$. Pour ce δ, nous avons déjà trouvé un N tel que $|I_{\delta,n} - I_n| \leqslant \epsilon$ dès que $n > N$. Nous avons :

$$|I_{\delta,n} - Z_{\delta,n}| \leqslant \int_{B(0,\delta)} |u(x) - u(0)| j_n(x) dx \tag{32.97a}$$

$$\leqslant \epsilon \int_{B(0,\delta)} j_n \tag{32.97b}$$

$$\leqslant \epsilon. \tag{32.97c}$$

Nous avons donc prouvé que pour tout $\epsilon > 0$, il existe un δ et un N tels que

$$\begin{cases} |I_{\delta,n} - I_n| \leqslant \epsilon & (32.98a) \\ |I_{\delta,n} - Z_{\delta,n}| \leqslant \epsilon & (32.98b) \end{cases}$$

dès que $n \geqslant N$.

Enfin nous avons

$$|z_{\delta,n} - u(0)| = u(0) \left(\int_{B(0,\delta)} j_n - 1 \right), \tag{32.99}$$

et par l'hypothèse (4) nous pouvons choisir n assez grand pour que la parenthèse soit plus petite que ϵ.

Pour ϵ donné, nous avons donc trouvé un δ et un N tels que

$$|I_n - u(0)| \leqslant |I_n - I_{\delta,n}| + |I_{\delta,n} - Z_{\delta,n}| + |Z_{\delta,n} - u(0)| \leqslant 3\epsilon. \tag{32.100}$$

En passant à la limite nous avons bien $I_n \to u(0)$ dans $\mathbb{R}$. □

Il va sans dire que nous connaissons de telles fonctions. Nous en donnons une maintenant.

Exemple 32.41([375])
Nous introduisons la fonction f_ϵ ($\epsilon > 0$) donnée par

$$f_\epsilon(x) = e^{-\epsilon|x|}. \tag{32.101}$$

Nous calculons la transformée de Fourier de f_ϵ en divisant le domaine d'intégration :

$$\hat{f}_\epsilon(k) = \int_{\mathbb{R}} e^{-ikx}e^{-\epsilon|x|}dx = \int_{-\infty}^{0} e^{\epsilon x}e^{-ikx}dx + \int_0^\infty e^{-\epsilon x}\ e^{-ikx}dx \tag{32.102}$$

En décomposant les parties imaginaires et réelles, et avec un peu de changement de variables, nous pouvons utiliser les intégrales (19.463) et (19.464) pour obtenir

$$\hat{f}_\epsilon(k) = \frac{2\epsilon}{k^2 + \epsilon^2}. \tag{32.103}$$

Sachant que $\operatorname{arctg}(x)$ est une primitive de $\frac{1}{x^2+1}$ et avec encore un peu de changement de variables, nous avons [11]

$$\int_{\mathbb{R}} \hat{f}_\epsilon(k)dk = \int_{-\infty}^{\infty} \frac{2\epsilon}{k^2 + \epsilon^2} = 2[\operatorname{arctg}(x/\epsilon)]_{-\infty}^{\infty} = 2\pi. \tag{32.104}$$

Cela montre que si nous introduisons la fonction δ_ϵ donnée par

$$\delta_\epsilon(k) = \frac{1}{\pi}\frac{\epsilon}{\epsilon^2 + k^2}, \tag{32.105}$$

alors nous avons une fonction qui tout en même temps ressemble à $\hat{f}_\epsilon$ et vérifie

$$\int_{\mathbb{R}} \delta_\epsilon(k)dk = 1 \tag{32.106}$$

pour tout ϵ.

Jusqu'ici nous avons montré que

$$\int_{\mathbb{R}} e^{-ikx}e^{-\epsilon|x|}dx = 2\pi\delta_\epsilon(k). \tag{32.107}$$

Pour chaque $\epsilon > 0$ nous avons $\delta_\epsilon \in L^1(\mathbb{R})$. △

Proposition 32.42 ([1])**.**
Soit $g \in \mathscr{S}(\mathbb{R})$. Alors nous avons

$$\int_{\mathbb{R}}\int_{\mathbb{R}} g(x)e^{-ixy}dx\,dy = 2\pi g(0). \tag{32.108}$$

Démonstration. Soit $u \in \mathscr{S}(\mathbb{R})$; nous multiplions l'équation (32.107) par $u(k)$ et nous intégrons par rapport à k :

$$\int_{\mathbb{R}} u(k)\left[\int_{\mathbb{R}} e^{-ikx}e^{-\epsilon|x|}dx\right]dk = 2\pi\int_{\mathbb{R}} u(k)\delta_\epsilon(k)dk. \tag{32.109}$$

Il s'agit de passer à la limite dans l'équation (32.109). Les intégrales à gauche peuvent effectuées séparément parce qu'elles respectent le théorème de Fubini. En effet soit la fonction

$$f(k, x) = u(k)e^{-ikx}e^{-\epsilon|x|} \tag{32.110}$$

qui est dans $L^1(\mathbb{R}\times\mathbb{R})$ en vertu du critère du corollaire 16.238 et du fait que à la fois $k \mapsto |u(k)|$ et $x \mapsto e^{-\epsilon|x|}$ sont dans $L^1(\mathbb{R})$.

11. Et en écrivant correctement l'intégrale sur $\mathbb{R}$ comme une limite, etc.

Nous pouvons donc grouper et dégrouper les intégrales et en particulier les inverser. Si nous effectuons d'abord l'intégrale sur k nous trouvons

$$\int_{\mathbb{R}} u(k) \left[\int_{\mathbb{R}} e^{-ikx} e^{-\epsilon|x|} dx \right] dk = \int_{\mathbb{R}} e^{-\epsilon|x|} \int_{\mathbb{R}} u(k) e^{-ikx} dk\, dx = \int_{\mathbb{R}} e^{-\epsilon|x|} \hat{u}(x) dx. \tag{32.111}$$

La fonction $x \mapsto |e^{-\epsilon|x|}\hat{u}(x)|$ est majorée (uniformément en ϵ) par $x \mapsto \hat{u}(x)$ qui est intégrable parce que la transformée de Fourier d'une fonction de $\mathscr{S}$ est dans $\mathscr{S}$ par la proposition 31.14. Le théorème de la convergence dominée de Lebesque 16.140 nous permet de permuter la limite $\epsilon \to 0$ avec l'intégrale et obtenir

$$\lim_{\epsilon \to 0} \int_{\mathbb{R}} u(k) \int_{\mathbb{R}} e^{-ikx} e^{-\epsilon|x|} dx\, dk = \int_{\mathbb{R}} \hat{u}(x) dx = \int_{\mathbb{R}} \int_{\mathbb{R}} u(k) e^{-ikx} dk\, dx. \tag{32.112}$$

Notons qu'en passant à la limite nous avons perdu le droit de permuter les intégrales.

Nous devons encore prouver que

$$\lim_{\epsilon \to 0} \int_{\mathbb{R}} u(k) \delta_\epsilon(k) dk = u(0). \tag{32.113}$$

Cela n'est rien d'autre que le lemme 32.40 appliqué à la suite de fonctions $j_n = \delta_{1/n}$. □

32.43.
Notons que les intégrales dans (32.108) ne peuvent pas être permutées parce que $\int_{\mathbb{R}} e^{-ixy} dy$ n'existe pas. Il faut avouer que, malgré tous les conseils du type « attention : permuter des intégrales doit être fait avec prudence », ce n'est pas tous les jours que nous trouvons des intégrales qui ne peuvent pas être permutées, autrement que dans des exemples fait exprès.

32.3.7 Peigne de Dirac

Proposition 32.44.
La formule

$$\Delta_a = \sum_{k \in \mathbb{Z}} \delta_{ka} \tag{32.114}$$

définit un élément de $\mathscr{D}'(\mathbb{R})$.

La forme linéaire Δ_a est le **peigne de Dirac** de pas a.

Démonstration. Nous utilisons le critère de continuité séquentielle en zéro du théorème 32.14. Soit une suite $\varphi_n \to 0$ dans $\mathscr{D}(\mathbb{R})$. Par le théorème 32.11 il existe un compact K de $\mathbb{R}$ pour lequel $\varphi_n \in \mathscr{D}(K)$ pour tout n et $\varphi_n \to 0$ dans $\mathscr{D}(K)$. La somme 32.114 est donc finie et nous pouvons la permuter avec une limite :

$$\lim_{n \to \infty} \Delta_a(\varphi_n) = \sum_{k \in \mathbb{Z}} \lim_{n \to \infty} \varphi_n(ka). \tag{32.115}$$

La limite $\varphi_n \to 0$ dans $\mathscr{D}(K)$ signifie que nous avons convergence uniforme de la fonction et de toutes ses dérivées vers 0. En particulier $\|\varphi_n\|_\infty \to 0$; disons que la somme (qui est finie) fasse s termes :

$$\sum_{k \in \mathbb{Z}} \varphi_n(ka) \leqslant s \|\varphi_n\|_\infty. \tag{32.116}$$

Le terme de droite tend vers zéro lorsque n tend vers l'infini. □

Donc Δ_a est bien une distribution au sens de la définition 32.13.

Lemme 32.45 ([345]).
Le peigne de Dirac vérifie la relation

$$\Delta_a = \frac{1}{a}\Delta_1 \circ D_a \tag{32.117}$$

où D_a est l'application $D_a\colon \mathscr{D}(\mathbb{R}) \to \mathscr{D}(\mathbb{R})$,

$$(D_a f)(x) = af(ax). \tag{32.118}$$

Démonstration. Pour $\varphi \in \mathscr{D}(\mathbb{R})$ nous avons

$$\Delta_a(\varphi) = \sum_{k\in\mathbb{Z}} \varphi(ka) = \frac{1}{a}\sum_{k\in\mathbb{Z}} (D_a\varphi)(k) = \frac{1}{a}\Delta_1(D_a\varphi). \tag{32.119}$$

□

Proposition 32.46.
Le peigne de Dirac est une distribution tempérée.

Notez qu'il y a plus de fonctions dans $\mathscr{S}(\mathbb{R})$ que dans $\mathscr{D}(\mathbb{R})$; il est donc plus difficile de rentrer dans $\mathscr{S}'(\mathbb{R})$ que dans $\mathscr{D}'(\mathbb{R})$: il est plus compliqué d'avoir existence de $T(\varphi)$ pour tout $\varphi \in \mathscr{S}(\mathbb{R})$ que pour tout $\varphi \in \mathscr{D}(\mathbb{R})$.

Démonstration. Soit $\varphi \in \mathscr{S}(\mathbb{R})$. Nous avons

$$|\Delta_a(\varphi)| = |\sum_k \varphi(ak)| = \left|\sum_k \frac{(1+a^2k^2)\varphi(ak)}{1+a^2k^2}\right| \leqslant_{x\in\mathbb{R}} \left|(1+x^2)\varphi(x)\right| \sum_k \frac{1}{1+a^2k^2}. \tag{32.120}$$

La somme $\sum_k \frac{1}{1+a^2k^2}$ est une somme convergente, et et supremum est borné par la proposition 28.90 en prenant $Q(x) = 1 + x^2$. En effet sur $\overline{B(0,r)}$ la fonction $x \mapsto (1+x^2)\varphi(x)$ est bornée par ce que c'est une fonction continue sur un compact, et à l'extérieur de $B(0,r)$ cette fonction est alors bornée par 1. □

Si aucune ambigüité n'est à craindre, nous noterons f la distribution T_f.

Exemple 32.47
La transformée de Fourier de la distribution de Dirac est la fonction constante : $\hat{\delta} = 1$. En effet si nous agissons sur une fonction test,

$$\hat{\delta}(\varphi) = \delta(\hat{\varphi}) = \hat{\varphi}(0) = \int_{\mathbb{R}^d} \varphi(x)dx. \tag{32.121}$$

△

32.4 L'espace $C^\infty(\mathbb{R}, \mathscr{D}'(\mathbb{R}^d))$

D'abord parlons un peu de continuité en recopiant la proposition 10.83 dans notre contexte.

Proposition 32.48.
Soient I un intervalle ouvert de $\mathbb{R}$ et $u\colon I \to \mathscr{D}'(\mathbb{R}^d)$ une fonction continue. Alors

(1) Pour tout $\varphi \in \mathscr{D}(\mathbb{R}^d)$, l'application $t \mapsto u_t(\varphi)$ est continue.

(2) Pour tout $\varphi \in \mathscr{D}(\mathbb{R}^d)$, nous avons la limite

$$\lim_{t\to t_0} u_t(\varphi) = u_{t_0}(\varphi). \tag{32.122}$$

(3) Nous avons la limite dans $\mathscr{D}'(\mathbb{R}^d)$

$$\lim_{t\to t_0} u_t = u_{t_0}. \tag{32.123}$$

En ce qui concerne la définition de l'espace $C^\infty(I,\mathscr{D}'(\mathbb{R}^d))$, c'est la définition 10.84. Grâce au point (1) de la proposition 32.48, nous retenons que la propriété fondamentale d'une application $T\in C^k\big(I,\mathscr{D}'(\Omega)\big)$ est que pour tout $\varphi\in(\Omega)$, l'application

$$\begin{aligned} I &\to \mathbb{C}\\ t &\mapsto T_t(\varphi) \end{aligned} \tag{32.124}$$

est de classe C^k.

Proposition 32.49.
Soit $(T_t)\in C^0\big(I,\mathscr{D}'(\Omega)\big)$ *et* $\psi\in\mathscr{D}(I\times\Omega)$. *Alors l'application*

$$t\mapsto T_t\big(\psi(t,.)\big) \tag{32.125}$$

est continue sur I.

Démonstration. La fonction dont nous voulons prouver la continuité est une fonction $\mathbb{R}\to\mathbb{R}$; il est donc loisible de se contenter de la continuité séquentielle. Soient $t_0\in I$ et (t_j) une suite dans I convergeant vers t_0. Nous posons $U_j=T_{t_j}$ et $\psi_j=\psi\big(t_j,.\big)$. Par hypothèse de continuité de (T_t) nous avons $U_j\xrightarrow{\mathscr{D}'(\Omega)}T_{t_0}$. D'autre part le support de ψ étant compact nous avons $\text{supp}(\psi)\subset[c,d]\times K$ où $[c,d]\subset I$ et K est compact dans Ω. Par conséquent nous avons aussi $\text{supp}(\psi_j)\subset K$.

Affin d'alléger les notations notons $\tilde{\psi}(x)=\psi(t_0,x)$. Pour tout multiindice α et pour tout j nous avons

$$p_\alpha(\psi_i-\tilde{\psi})=\Big|\partial^\alpha\psi(t_j,x)-\partial^\alpha\psi(t_0,x)\Big|\leqslant|t_j-t_0|\sup_{\substack{t\in[c,d]\\x\in K}}|\partial_t\partial^\alpha\psi(t,x)|\to 0. \tag{32.126}$$

Nous avons donc la convergence

$$\psi_j\xrightarrow{\mathscr{D}(K)}\psi(t_0,.). \tag{32.127}$$

Étant donné que $U_j\xrightarrow{\mathscr{D}'(\Omega)}T_{t_0}$ et $\psi_j\xrightarrow{\mathscr{D}(\Omega)}\tilde{\psi}$, le point (3) du corollaire 28.8 nous donne la convergence

$$U_j(\psi_j)\to T_{t_0}(\tilde{\psi}) \tag{32.128}$$

dans $\mathbb{C}$. Cela est bien la continuité de la fonction $t\mapsto T_t\big(\psi(t,.)\big)$. □

Proposition 32.50 ([373])**.**
Soit $(T_t)\in C^0\big(I,\mathscr{D}'(\Omega)\big)$. *Nous définissons l'application* $T\colon\mathscr{D}(\Omega)\to\mathbb{C}$ *par la formule*

$$T(\psi)=\int_I T_t\big(\psi(t,.)\big)\,dt \tag{32.129}$$

pour tout $\psi\in\mathscr{D}(I\times\Omega)$. *Alors* $T\in\mathscr{D}'(I\times\Omega)$.

Démonstration. La proposition 32.49 nous indique que la fonction $t\mapsto T_t\big(\psi(t,.)\big)$ est continue. Étant donné qu'elle est seulement non nulle sur un compact, l'intégrale

$$\int_I T_t\big(\psi(t,.)\big)dt \tag{32.130}$$

a un sens et est finie. L'application $T\colon\mathscr{D}(I\times\Omega)\to\mathbb{C}$ ainsi définie est linéaire. Il reste à voir qu'elle est continue. Pour cela nous allons utiliser le théorème 32.14(2) qui nous dit que nous pouvons nous fixer un compact $[c,d]\times K\subset I\times\Omega$ et considérer $\psi\in\mathscr{D}\big([c,d]\times K\big)$.

Soit, pour commencer, donnée une application $\varphi \in \mathscr{D}(K)$. L'application $t \mapsto T_t(\varphi)$ est continue et non nulle sur le et il existe donc $C_\varphi > 0$ tel que

$$|T_t(\varphi)| \leqslant C_\varphi \tag{32.131}$$

pour tout $t \in [c,d]$.

Nous voulons utiliser le théorème de Banach-Steinhaus dans sa version 28.9 sur la famille d'applications paramétrée par $u \in [c,d]$:

$$\begin{aligned} U_u \colon \mathscr{D}\big([c,d] \times K\big) &\to \mathbb{R} \\ \psi &\mapsto T_u\big(\psi(u,.)\big). \end{aligned} \tag{32.132}$$

Commençons par prouver que cela est une application continue pour chaque u. Ce sera le cas si la projection

$$\begin{aligned} \texttt{proj} \colon \mathscr{D}\big([c,d] \times K\big) &\to \mathscr{D}(K) \\ \psi &\mapsto \psi(u,.) \end{aligned} \tag{32.133}$$

est continue. Pour cela nous notons P_{kl} la semi-norme sur $\mathscr{D}\big([c,d] \times K\big)$ donnée par

$$P_{k,l}(\psi) = \sum_{n \leqslant k} \sum_{|\alpha| \leqslant l} \sup_{\substack{t \in [c,d] \\ x \in K}} \big|\partial_t^n \partial^\alpha \psi(t,x)\big|. \tag{32.134}$$

Nous montrons que $\texttt{proj}$ est séquentiellement continue ; étant donné que les topologies sur $\mathscr{D}(K)$ et $\mathscr{D}\big([c,d] \times K\big)$ sont données par des métriques (proposition 32.12), cela suffit pour assurer la continuité grâce à la proposition 10.14. Montrons que si $\psi_n \xrightarrow{\mathscr{D}([c,d]\times K)} 0$, alors $\texttt{proj}(\psi_n) \xrightarrow{\mathscr{D}(K)} 0$. Pour cela nous remarquons que

$$p_j\big(\texttt{proj}(\psi)\big) = \sum_{|\alpha| \leqslant j} \sup_{x \in K} |\partial^\alpha \psi(u,x)| \tag{32.135a}$$

$$\leqslant \sum_{|\alpha| \leqslant j} \sup_{t \in [c,d]} \sup_{x \in K} |\partial^\alpha \psi(t,x)| \tag{32.135b}$$

$$= P_{0,j}(\psi). \tag{32.135c}$$

Par conséquent

$$p_j\big(\texttt{proj}(\psi_n)\big) \leqslant P_{0,j}(\psi_n) \to 0 \tag{32.136}$$

où nous avons utilisé la proposition 10.76. Utilisant cette même proposition à l'envers, nous déduisons que $\texttt{proj}(\psi_n) \xrightarrow{\mathscr{D}(K)} 0$. Les applications U_u sont donc continues ; elles sont également bornées parce que si $\psi \in \mathscr{D}\big([c,d] \times K\big)$ nous avons

$$\sup_{u \in [c,d]} \big|U_u(\psi)\big| = \sup_u \big|T_u\big(\psi(u,.)\big)\big|, \tag{32.137}$$

et la continuité déjà évoquée, sur le compact $[c,d]$, nous dit que cette quantité est finie. Le théorème de Banach-Steinhaus peut maintenant être appliqué et il existe $C > 0$ et $k, l \in \mathbb{N}$ tels que pour tout $\psi \in \mathscr{D}\big([c,d] \times K\big)$,

$$\big|U_u(\psi)\big| \leqslant C P_{k,l}(\psi) = C \sum_{|\alpha| \leqslant k} \sum_{n \leqslant l} \sup_{t,x} \big|\partial_t^n \partial^\alpha \psi(t,x)\big| \leqslant C \sum_{|\alpha|+n \leqslant k+l} \sup_{t,x} \big|\partial_t^n \partial^\alpha \psi(t,x)\big|. \tag{32.138}$$

Quelques remarques

— Nous n'avons pas mis de maximum devant le supremum (alors que la conclusion (28.13) en demande) parce que dans le cas des semi-normes P_{kl}, c'est toujours celle avec k et l le plus grand possible qui sont les plus grandes parce qu'elles sont des sommes emboitées les unes les autres.

— La fusion de deux sommes est bien une majoration parce qu'il y a plus de termes dans la seconde que dans la première.
— La quantité la plus à droite est (à part le C) ce que nous pouvons noter $P_{k+l}(\psi)$: c'est bien une des semi-normes associées à l'espace de dimension $d+1$.

Nous majorons maintenant $T(\psi)$ par

$$\big|T(\psi)\big| \leqslant \int_c^d \big|T_t\big(\psi(t,.)\big)\big|dt = \int_c^d \big|U_t(\psi)\big|dt \leqslant C|d-c|P_{k+l}(\psi). \tag{32.139}$$

Maintenant le théorème 32.14(2) appliqué à l'ouvert $I\times\Omega$ et avec ψ au lieu de φ nous informe que $T\in\mathscr{D}(I\times K)$. □

32.4.1 Dérivation

Quelques propriétés de dérivation des fonctions $I\to\mathscr{D}(\Omega)$ seront directement énoncées et démontrées dans le cas des distributions tempérées. Les résultats 32.57 et 32.58 seront a fortiori valables si nous remplaçons $\mathscr{S}$ par $\mathscr{D}$.

32.5 L'espace $C^\infty(\mathbb{R},\mathscr{S}'(\mathbb{R}^d))$

Dans cette section nous notons I un ouvert de $\mathbb{R}$ et Ω un ouvert de $\mathbb{R}^d$; si ψ est une fonction sur $I\times\Omega$ nous allons noter $\psi_t\colon\Omega\to\mathbb{R}$ la fonction $\psi_t(x)=\psi(t,x)$. C'est une notation plus légère que $\psi(t,.)$.

32.5.1 Propriétés générales

La définition de l'espace $C^\infty(I,\mathscr{S}'(\Omega))$ est encore la définition 10.84 et les propriétés énoncées dans la proposition 32.48 sont encore bonnes ici.

D'abord parlons un peu de continuité en recopiant la proposition 10.83 dans notre contexte.

Proposition 32.51.
Soient I un intervalle ouvert de $\mathbb{R}$ et $u\colon I\to\mathscr{S}'(\mathbb{R}^d)$ une fonction continue. Alors

(1) Pour tout $\varphi\in\mathscr{S}(\mathbb{R}^d)$, l'application $t\mapsto u_t(\varphi)$ est continue.

(2) Pour tout $\varphi\in\mathscr{S}(\mathbb{R}^d)$, nous avons la limite

$$\lim_{t\to t_0} u_t(\varphi) = u_{t_0}(\varphi). \tag{32.140}$$

(3) Nous avons la limite dans $\mathscr{S}'(\mathbb{R}^d)$

$$\lim_{t\to t_0} u_t = u_{t_0}. \tag{32.141}$$

Lemme 32.52.
Nous avons $C^\infty(I,\mathscr{S}'(\Omega))\subset C^\infty(I,\mathscr{D}'(\Omega))$.

Démonstration. Soit $(T_t)\in C^\infty(I,\mathscr{S}'(\Omega))$. Pour chaque t nous avons

$$T_t\in\mathscr{S}'(\Omega)\subset\mathscr{D}'(\Omega). \tag{32.142}$$

Ensuite il suffit de dire que pour tout $\varphi\in\mathscr{D}(\Omega)$ la fonction

$$t\mapsto T_t(\varphi) \tag{32.143}$$

est de classe C^∞ parce que c'est le cas pour toute fonction dans $\mathscr{S}(\Omega)$. La proposition 32.48 (en changeant $\mathscr{D}$ en $\mathscr{S}$) conclut que $(T_t)\in C^\infty(I,\mathscr{D}'(\Omega))$. □

Proposition 32.53.
L'espace $\mathscr{S}(\Omega)$ est complet et métrisable.

Démonstration. En ce qui concerne le métrisable nous reprenons la formule de l'écart (10.109). Dans notre cas pour l'écrire explicitement il faudrait une énumération de $\mathbb{N}^2$ à partir de 1 (et non de zéro). Cette formule donne bien une distance parce que si $d(\varphi_1 - \varphi_2) = 0$ alors en particulier $p_{00}(\varphi_1 - \varphi_2) = \|\varphi_1 - \varphi_2\|_\infty = 0$ et donc $\varphi_1 = \varphi_2$.

Nous montrons maintenant que $\mathscr{S}(\Omega)$ est complet en y considérant une suite de Cauchy (φ_n). Soit $\epsilon > 0$ et $\alpha, \beta \in \mathbb{N}$ ainsi que k, l assez grands pour que $\varphi_k - \varphi_l \in B_{\alpha\beta}(0, \epsilon)$. En particulier pour $\alpha = \beta = 0$ nous avons $\|\varphi_k - \varphi_l\|_\infty \leqslant \epsilon$, ce qui signifie que nous avons une suite vérifiant le critère de Cauchy uniforme 14.313. Elle converge donc uniformément vers une certaine fonction φ que la proposition 14.314 nous assure être continue. Il existe donc $\varphi \in C(\Omega)$ telle que

$$\varphi_k \xrightarrow{unif} \varphi. \tag{32.144}$$

Nous devons montrer que $\varphi \in \mathscr{S}(\Omega)$. Le fait que φ soit de classe C^∞ s'obtient en utilisant les semi-normes $p_{0,\alpha}(\varphi) = \|\partial^\alpha \varphi\|_\infty$ de la même façon que dans la preuve que $\mathscr{D}(\Omega)$ était complet (proposition 32.12). Nous obtenons en particulier que

$$\partial^\alpha \varphi_k \xrightarrow{unif} \partial^\alpha \varphi \tag{32.145}$$

pour tout multiindice α. Montrons encore que φ est à décroissance rapide : nous devons montrer que pour tout α et β nous avons

$$p_{\alpha\beta}(\varphi) = \sup_{x\in\Omega} \left|x^\beta (\partial^\alpha \varphi)(x)\right| < \infty. \tag{32.146}$$

Étant donné que (φ_n) est de Cauchy dans $\mathscr{S}(\Omega)$ nous avons (pour ϵ fixé et k, l assez grands) :

$$\left|x^\beta (\partial^\alpha \varphi_k - \partial^\alpha \varphi_l)(x)\right| \leqslant \epsilon \tag{32.147}$$

pour tout $x \in \Omega$. En considérant l fixé et en prenant la limite $k \to \infty$ et en utilisant la convergence uniforme (32.145) nous trouvons que

$$\left|x^\beta (\partial^\alpha \varphi - \partial^\alpha \varphi_l)(x)\right| \leqslant \epsilon \tag{32.148}$$

Du coup nous pouvons faire la majoration

$$\sup_{x\in\Omega} \left|x^\beta (\partial^{\alpha\varphi})(x)\right| \leqslant \sup_x \left|x^\beta (\partial^\alpha \varphi - \partial^\alpha \varphi_l)(x)\right| + \sup_x \left|(\partial^\alpha \varphi_l)(x)\right| \leqslant \epsilon + p_{\alpha\beta}(\varphi_l) < \infty \tag{32.149}$$

du fait que $p_{\alpha\beta}(\varphi_l) < \infty$ parce que $\varphi_l \in \mathscr{S}(\Omega)$.

Donc $\varphi \in \mathscr{S}(\Omega)$ et ce dernier est alors complet. □

Proposition 32.54.
Soit $(T_t) \in C^0\big(I, \mathscr{S}'(\Omega)\big)$ et $\psi \in \mathscr{S}(I \times \Omega)$. Alors la fonction

$$t \mapsto T_t(\psi_t) \tag{32.150}$$

est continue sur I.

Démonstration. Soient $t_0 \in I$ et une suite convergente vers t_0 : $t_j \to t_0$ dans $\mathbb{R}$. Vu que (T_t) est continue en t, elle est en particulier séquentiellement continue et nous avons

$$T_{t_j} \xrightarrow{\mathscr{S}'(\Omega)} T_{t_0}. \tag{32.151}$$

Montrons que nous avons aussi $\psi_{t_j} \xrightarrow{\mathscr{S}(\Omega)} \psi_{t_0}$. Pour cela nous utilisons les semi-normes [12] $p_{\alpha\beta}$ définies

12. Pas parce que nous en avons envie, mais bien parce qu'elles font partie de la définition de la convergence et de tous ces trucs.

en (28.364) :

$$p_{\alpha\beta}(\psi_{t_j}-\psi_{t_0}) = \sum_{x\in\Omega}\left|x^\beta\big(\partial^\alpha\psi(t_j,x)-\partial^\alpha\psi(t_0,x)\big)\right| \tag{32.152a}$$
$$\leqslant \sup_{x\in\Omega}\left|x^\beta|t_0-t_j|\sup_{t\in[t_0,t_j]}\left|\partial_t\partial^\alpha\psi(t,x)\right|\right| \tag{32.152b}$$
$$\leqslant |t_0-t_j|\sup_{x\in\Omega}\sup_{t\in I}\left|x^\beta\partial_t\partial^\alpha\psi(t,x)\right| \tag{32.152c}$$
$$\leqslant |t_0-t_j|P_{(\alpha t),\beta}(\psi). \tag{32.152d}$$

Pour la première majoration nous avons utilisé le théorème des accroissements finie 14.248. Pour la dernière ligne nous avons noté $P_{\alpha\beta}$ les semi-normes de $\mathscr{S}(I\times\Omega)$ et $(t\alpha)$ est le multiiindice qui commence par la variable t et qui continue par α. Étant donné que $P_{(\alpha t)\beta}(\psi)<\infty$ nous avons bien

$$p_{\alpha\beta}(\psi_{t_j}-\psi_{t_0})\to 0 \tag{32.153}$$

et donc $\psi_{t_j}\xrightarrow{\mathscr{S}(\Omega)}\psi_{t_0}$.

Étant donné que $\mathscr{S}(\Omega)$ est métrisable et complet, le corollaire 28.8 nous dit que

$$T_{t_j}(\psi_{t_j})\to T_{t_0}(\psi_{t_0}), \tag{32.154}$$

ce qui est bien le critère de continuité séquentielle de la fonction (32.150). □

Remarque 32.55.
La proposition 32.50 nous dit, a fortiori, que si $(T_t)\in C^\infty\big(I,\mathscr{S}'(\Omega)\big)$ alors la formule

$$\tilde{T}(\psi)=\int_I T_t(\psi_t) \tag{32.155}$$

donne un élément $\tilde{T}\in\mathscr{D}'(I\times\Omega)$. Au cas où aucune confusion n'est à craindre, nous pourrons noter également T l'élément de $\mathscr{D}'(I\times\Omega)$ déduit de $T\in C^\infty\big(I,\mathscr{S}'(\Omega)\big)$.

Notons que ce T ne sera pas toujours une distribution tempérée comme le montre l'exemple suivant.

Exemple 32.56
En posant $T_t(\varphi)=e^{t^2}\varphi(0)$ avec $I=\mathbb{R}$, l'intégrale

$$T(\psi)=\int_{\mathbb{R}}T_t(\psi_t)=\int_{\mathbb{R}}e^{t^2}\psi(t,0)dt \tag{32.156}$$

ne converge pas pour tout $\psi\in\mathscr{S}(\mathbb{R}\times\Omega)$. En effet par rapport à t, la fonction $\psi(t,0)$ décroît rapidement mais pas spécialement assez rapidement pour compenser e^{t^2}. △

32.5.2 Dérivation

Proposition 32.57 ([373]).
Soit $T\in C^k\big(I,\mathscr{S}'(\Omega)\big)$ et $0\leqslant l\leqslant k$. Pour tout $t_0\in I$ l'application

$$\begin{aligned}T_{t_0}^{(l)}\colon\mathscr{S}(\Omega)&\to\mathbb{C}\\ \varphi&\mapsto\left(\frac{d^l}{dt}T_t(\varphi)\right)(t_0)\end{aligned} \tag{32.157}$$

est bien définie, est une distribution et de plus

$$t\mapsto T_t^{(l)}\in C^{k-l}\big(I,\mathscr{S}'(\Omega)\big). \tag{32.158}$$

Attention que la formule (32.157) est bonne si $\varphi\in\mathscr{S}(\Omega)$. Si par contre $\psi\in\mathscr{S}(I\times\Omega)$ et qu'on veut regarder $u_t^{(1)}(\psi_t)$ alors il faut regarder la proposition 32.58 et utiliser la formule (32.163) dans laquelle se trouve $u_t^{(1)}(\psi_t)$.

Démonstration. Pour $k=0$ nous avons $T_t^{(0)}=T_t$ et c'est bon. Pour le cas $k=1$ et $l=0$ c'est encore $T_t^{(0)}=T_t$ qui fonctionne.

Le premier cas non trivial à traiter est $k=1$ et $l=1$. Nous considérons $t_0\in I$; par définition de la dérivée, pour tout $\varphi\in\mathscr{S}(\Omega)$, nous avons (pour peu que les limites existent) :

$$T_{t_0}^{(1)}(\varphi)=\frac{d}{dt}\Big[T_t(\varphi)\Big]_{t=t_0}=\lim_{j\to\infty}\frac{T_{t_0+\epsilon_j}(\varphi)-T_{t_0}(\varphi)}{\epsilon_j}=\lim_{j\to\infty}U_j(\varphi)\tag{32.159}$$

où

$$U_j=\frac{1}{\epsilon_j}\big(T_{t_0+\epsilon_j}-T_{t_0}\big).\tag{32.160}$$

et (ϵ_j) est une suite de réels tendant vers zéro.

Vu que $(T_t)\in C^k(I,\mathscr{S}'(\Omega))$, l'application $t\mapsto T_t(\varphi)$ est de classe C^k et en particulier l'expression (32.159) a une limite lorsque $j\to\infty$. Donc $T_{t_0}^{(1)}(\varphi)$ est bien définie. Le point (1) du corollaire 28.8 nous dit que $\lim_{j\to\infty}U_j(\varphi)=T_{t_0}^{(1)}(\varphi)$ et $T_{t_0}^{(1)}$ est une distribution (linéaire et continue).

Nous devons encore voir que $t\mapsto T_t^{(1)}$ est une application $C^0\big(I,\mathscr{S}'(\Omega)\big)$. Cela est une conséquence du fait que (T_t) soit de classe C^1, ce qui se traduit par le fait que l'application

$$t\mapsto\frac{d}{dt}\Big(T_t(\varphi)\Big)\tag{32.161}$$

est continue (définition de la dérivée et point (1) de la proposition 32.51 appliquée à la dérivée).

Les cas $k\geqslant 1$ se traitent par récurrence. □

Proposition 32.58 ([373])**.**
Soit $(T_t)\in C^1\big(I,\mathscr{S}'(\Omega)\big)$ et $\psi\in\mathscr{S}(I\times\Omega)$. Alors la fonction

$$t\mapsto T_t\big(\psi(t,.)\big)\tag{32.162}$$

est de classe C^1 sur I et

$$\frac{d}{dt}\Big(T_t\big(\psi(t,.)\big)\Big)=T_t^{(1)}\big(\psi(t,.)\big)+T_t\left(\frac{\partial\psi}{\partial t}(t,.)\right)\tag{32.163}$$

Démonstration. Soient $t_0\in I$ et $\epsilon_j\to 0$ une suite réelle. Le membre de gauche de (32.163), écrit en t_0, donne

$$\spadesuit=\lim_{j\to\infty}\frac{T_{t_0+\epsilon_j}\big(\psi(t_0+\epsilon_j,.)\big)-T_{t_0}\big(\psi(t_0,.)\big)}{\epsilon_j}\tag{32.164}$$

Afin d'alléger les notations nous allons écrire $\psi_t=\psi(t,.)$. Dans le numérateur de (32.164) nous ajoutons et soustrayons la quantité $T_{t_0+\epsilon_j}(\psi_{t_0})$ et nous découpons la limite en deux morceaux :

$$\spadesuit=\lim_{j\to\infty}\frac{T_{t_0+\epsilon_j}(\psi_{t_0+\epsilon_j}-\psi_{t_0})}{\epsilon_j}+\lim_{j\to\infty}\frac{(T_{t_0+\epsilon_j}-T_{t_0})(\psi_{t_0})}{\epsilon_j}\tag{32.165}$$

Le second terme vaut

$$\frac{d}{dt}\Big(T_t(\psi_{t_0})\Big)_{t=t_0}=T_{t_0}^{(1)}(\psi_{t_0})\tag{32.166}$$

par la proposition 32.57. Occupons nous de l'autre morceau de ♠. Nous posons $U_j=T_{t_0+\epsilon_j}$ et

$$\varphi_j=\frac{1}{\epsilon_j}(\psi_{t_0+\epsilon_j}-\psi_{t_0}).\tag{32.167}$$

Nous voulons utiliser le corollaire 28.8(3) pour obtenir

$$\lim_{j\to\infty} U_j(\varphi_j) = T_{t_0}\Big(\frac{\partial\psi}{\partial t}(t_0,.)\Big). \tag{32.168}$$

D'une part (T_t) est de classe C^∞ en t et nous avons donc la convergence $U_j \xrightarrow{\mathscr{S}'(\Omega)} T_{t_0}$. Reste à prouver que

$$\varphi_j \xrightarrow{\mathscr{S}(\Omega)} \frac{\partial\psi}{\partial t}(t_0,.). \tag{32.169}$$

Cela en remarquant bien que la variable de dérivation n'est pas celle par rapport à laquelle nous voulons la convergence Schwartz [13]. Soient α et β des naturels et calculons un peu :

$$p_{\alpha\beta}\big(\varphi_j - \frac{\partial\psi}{\partial t}(t_0,.)\big) = \sup_{x\in\Omega}\Big|x^\beta\partial^\alpha\Big(\frac{1}{\epsilon_j}\big(\psi(t_0+\epsilon_j,x)-\psi(t_0,x)\big) - \frac{\partial\psi}{\partial t}(t_0,x)\Big)\Big| \tag{32.170}$$

Il est à présent l'heure d'utiliser un développement de Taylor avec le reste de la proposition 14.288 :

$$\psi(t_0+\epsilon_j,x) = \psi(t_0,x) + \epsilon_j\frac{\partial\psi}{\partial t}(t_0,x) + \frac{\epsilon_j^2}{2}\frac{\partial^2\psi}{\partial t^2}(\bar{t},x) \tag{32.171}$$

pour un certain $\bar{t}\in[t_0,t_0+\epsilon_j]$. En mettant ça dans le calcul (32.170) nous restons avec

$$p_{\alpha\beta}\big(\varphi_j - \frac{\partial\psi}{\partial t}(t_0,.)\big) = \sup_{x\in\Omega}\Big|x^\beta\partial^\alpha\Big(\epsilon_j\frac{\partial^2\psi}{\partial t^2}(\bar{t},x)\Big)\Big| \leqslant \epsilon_j P_{\alpha,2;\beta,0}(\psi) \tag{32.172}$$

où $P_{\alpha,k;\beta,l}$ sont les semi-normes de $\mathscr{S}(I\times\Omega)$ avec la notation plus ou moins évidente de prendre α dérivations sur x, k sur t puis de multiplier par $x^\beta t^l$. Au final nous avons bien

$$\lim_{j\to\infty} p_{\alpha\beta}\big(\varphi_j - \frac{\partial\psi}{\partial t}(t_0,.)\big) = 0 \tag{32.173}$$

et donc la convergence $\varphi_j \xrightarrow{\mathscr{S}(\Omega)} \frac{\partial\psi}{\partial t}(t_0,.)$. □

Lemme 32.59.
Soit $(T_t)\in C^1\big(I,\mathscr{S}'(\Omega)\big)$ alors si $\mathcal{F}$ dénote la transformée de Fourier nous avons

$$\mathcal{F}\big(T_t^{(1)}\big) = (\mathcal{F}T)_t^{(1)} \tag{32.174}$$

où $(\mathcal{F}T)$ est la famille de distributions $(\mathcal{F}T)_t = \mathcal{F}T_t$.

Démonstration. Pour la preuve il suffit de tester l'égalité sur une fonction $\varphi\in\mathscr{S}(\Omega)$:

$$(\mathcal{F}T_t^{(1)})(\varphi) = T_t^{(1)}(\mathcal{F}\varphi) = \frac{d}{dt}\Big(T_t(\mathcal{F}\varphi)\Big) = \frac{d}{dt}\Big((\mathcal{F}T_t)(\varphi)\Big) = (\mathcal{F}T)_t^{(1)}(\varphi). \tag{32.175}$$

□

32.6 Une équation de distribution

Nous allons étudier l'équation

$$(x-x_0)^\alpha u = 0 \tag{32.176}$$

pour $u\in\mathscr{D}'(\mathbb{R})$ et $\alpha\in\mathbb{N}$ est donné fixé. Notons tout de suite que (32.176) est un petit abus de notation pour dire qu'en vertu de la définition 32.19 du produit d'une distribution par une fonction, pour tout $\phi\in\mathscr{D}(\mathbb{R})$, nous avons $u\Big(x\mapsto(x-x_0)\phi(x)\Big) = 0$.

13. Je ne sais pas si je me suis bien fait comprendre là.

Lemme 32.60 ([144]).
Soit $\alpha \in \mathbb{N}$. Une solution à l'équation

$$(x - x_0)^\alpha u = 0 \tag{32.177}$$

est une distribution à support dans $\{x_0\}$ et d'ordre fini.

Démonstration. Nous commençons par prouver que u est une solution de (32.177) si et seulement si [14] $\langle u, \phi\rangle = 0$ pour tout $\phi \in \mathscr{D}$ telle que

$$\phi(x_0) = \ldots = \partial^{\alpha-1}\phi(x_0) = 0. \tag{32.178}$$

Condition nécéssaire Supposons que u soit une solution. Alors le corollaire 19.38 du théorème de Hadamard donne $\psi \in \mathscr{D}(\mathbb{R})$ telle que $\phi(x) = (x-x_0)^\alpha\psi(x)$. Dans ce cas, si u est solution de (32.177), alors

$$0 = \langle (x - x_0)^\alpha u, \psi\rangle = \langle u, (x - x_0)^\alpha\psi(x)\rangle = \langle u, \phi\rangle. \tag{32.179}$$

Nous avons vu que si u est solution, alors $\langle u, \phi\rangle = 0$ pour tout ϕ satisfaisant la condition (32.178).

Condition suffisante Supposons maintenant l'inverse : u est une distribution s'annulant sur toute fonction $\phi \in \mathscr{D}'$ satisfaisant (32.178). Nous allons alors prouver que u est une solution. Soit donc $\psi \in \mathscr{D}$ et calculons

$$\langle (x - x_0)u, \psi\rangle = \langle u, (x - x_0)\psi\rangle = 0 \tag{32.180}$$

parce que la fonction $(x - x_0)\psi(x)$ vérifie la condition (32.178).

Nous passons maintenant au cœur de la preuve : nous supposons que u est une solution. Si le support de ϕ est contenu dans $\mathbb{R}\backslash\{x_0\}$ alors ϕ est nulle dans un voisinage de x_0 (et donc $\partial^k\phi = 0$ pour tout k) et $\langle u, \phi\rangle = 0$. Autrement dit, pour tout $\phi \in \mathscr{D}\big(\mathbb{R}\backslash\{x_0\}\big)$ nous avons $\langle u, \phi\rangle = 0$, ce qui signifie que $\text{supp}(u) \cap \big(\mathbb{R}\backslash\{x_0\}\big) = \varnothing$ ou encore que $\text{supp}(u) = \{x_0\}$.

Maintenant que u a un support compact, la proposition 32.25 nous indique qu'elle est d'ordre fini. □

Théorème 32.61 ([144]).
Soit $\alpha \in \mathbb{N}$ et l'équation

$$(x - x_0)^\alpha u = 0 \tag{32.181}$$

pour $u \in \mathscr{D}'(\mathbb{R})$. Les solutions sont les combinaisons linéaires des dérivées de δ_{x_0} jusqu'à la α^eexclue.

Démonstration. D'abord montrons que les $\partial^i\delta_{x_0}$ sont des solutions. Avec les définition 32.19 et 32.20 des dérivées de distributions et de leur produits avec des fonctions [15],

$$(x - x_0)^\alpha\partial^i\delta_{x_0}(\phi) = \delta_{x_0}\Big(\partial^i\big((x - x_0)^\alpha\phi(x)\big)\Big) \tag{32.182}$$

Si $i < \alpha$ alors dans chaque terme de Leibnitz, il y aura un facteur $(x - x_0)$, et la prise de δ_{x_0} annulera. Si par contre $i \geqslant \alpha$ alors il y aura le terme

$$\binom{i}{\alpha}\partial^\alpha\big((x - x_0)^\alpha\big)\partial^{i-\alpha} = \binom{i}{\alpha}\alpha!(\partial^{i-\alpha}\phi)(x_0) \tag{32.183}$$

qui est le seul terme contenant $(\partial^{i-\alpha}\phi)(x_0)$. Il suffit alors de choisir $\phi \in \mathscr{D}(\mathbb{R})$ de sorte que

$$(\partial^k\phi)(x_0) = \begin{cases} 0 & \text{si } k \neq i - \alpha \\ 1 & \text{si } k = i - \alpha \end{cases} \tag{32.184}$$

14. En réalité nous n'aurons besoin que de la condition nécessaire, en particulier pour le théorème 32.61.

15. Comme souvent, dans l'expression suivante, il y a un abus de notation parce que x est une variable muette : il faudrait écrire « $x \mapsto$ » au début de la grade parenthèse.

et alors on est certain que le tout n'est pas nul, et donc que $(x-x_0)^\alpha(\partial^i\delta_{x_0})\neq 0$.

Jusqu'ici nous avons prouvé que $\partial^i\delta_{x_0}$ est solution si et seulement si $0\leqslant i<\alpha$.

Il faut encore prouver que les solutions sont toutes des combinaisons linéaires de dérivées de delta de Dirac centrées en x_0. Pour cela nous invoquons d'abord le lemme 32.60 qui nous assure que u est d'ordre fini et de support $\{x_0\}$. Ensuite la proposition 32.28 nous indique que u doit alors être une combinaisons linéaire de dérivées de Dirac. □

Chapitre 33

Espaces de Sobolev, équations elliptiques

33.1 Espaces de Sobolev

Rappel : la définition de la dérivée faible est 32.1.

33.1.1 Sur un intervalle de $\mathbb{R}$

Sauf mention du contraire dans cette section I est un intervalle borné ouvert $I =]a, b[$ de $\mathbb{R}$.

Définition 33.1.
*Soit $I =]a, b[$ un ouvert borné de $\mathbb{R}$. L'**espace de Sobolev** $H^1(I)$ est l'ensemble*

$$H^1(I) = \left\{ u \in L^2(I) \text{ tel que } \exists g \in L^2(I) \text{ tel que } \forall \varphi \in C_c^\infty(I), \int_I u\varphi' = -\int_I g\varphi \right\}. \tag{33.1}$$

L'unique élément g de $L^2(I)$ vérifiant $\int_I u\varphi' = -\int_I g\varphi$ est noté u' est est nommé **dérivée** ; nous verrons dans les prochaines pages pourquoi.

L'espace H^1 accepte le produit scalaire suivant :

$$\langle u, v \rangle = \int_I uv + \int_I u'v', \tag{33.2}$$

et nous notons $\|.\|_{H^1}$ la norme correspondante qui n'est autre que

$$\|u\|_{H^1} = \langle u, u \rangle = \|u\|_{L^2}^2 + \|u'\|_{L^2}. \tag{33.3}$$

Nous introduisons l'espace $L^1_{loc}(I)$ des fonctions étant L^1 sur tout compact de I.

Corollaire 33.2.
Si $u \in H^1(I)$ et si $u' = 0$ alors il existe une constant C telle que $u = C$ presque partout.

Démonstration. L'hypothèse $u' = 0$ signifie que pour tout fonction $\varphi \in C_c^\infty(I)$,

$$\int_I u\varphi' = \int_I u'\varphi = 0. \tag{33.4}$$

La proposition 28.77 nous dit alors qu'il existe une constante C telle que $u = C$ presque partout. □

Lemme 33.3.
Tout élément de $H^1(I)$ admet un unique représentant continu.

Nous verrons dans le corollaire 33.5 que ce représentant pourra être prolongé par continuité sur $\bar{I}$.

Démonstration. Soient $y_0 \in I$ et $u \in H^1(I)$. Nous considérons la fonction

$$\bar{u}(x) = \int_{y_0}^{x} u'(t)dt. \tag{33.5}$$

Notons que par définition, $u' \in L^2$ donc l'intégrale ne pose pas de problèmes. Montrons que $\bar{u}$ est continue sur $\bar{I}$. Pour cela nous considérons $x \in \bar{I}$ et h tel que $x + h \in \bar{I}$. Alors

$$\big|\bar{u}(x+h) - \bar{u}(x)\big| = \Big|\int_{x}^{x+h} u'\Big| \leqslant \int_{x}^{x+h} |u'|. \tag{33.6}$$

Mais la fonction $|u'|$ est dans $L^1_{loc}(I)$ par le lemme 28.25 ; elle est en particulier intégrable sur un ouvert contenant x et par conséquent la dernière intégrale tend vers zéro lorsque h tend vers 0.

Nous prouvons à présent que $\bar{u}$ est dans $H^1(I)$ et que sa dérivée est égale à u' ; pour cela nous allons montrer que pour tout $\varphi \in C^\infty_c(I)$,

$$\int_I \bar{u}\varphi' = -\int_I u'\varphi. \tag{33.7}$$

Nous avons

$$\int_I \bar{u}\varphi' = \int_I \left(\int_{y_0}^{x} u'(t)dt\right)\varphi'(x)dx = \int_a^{y_0} \left(\int_{y_0}^{x} u'(t)dt\right)\varphi'(x)dx + \int_{y_0}^{b} \left(\int_{y_0}^{x} u'(t)dt\right)\varphi'(x)dx. \tag{33.8}$$

Pour faire plus court, nous notons $f(t,x) = u'(t)\varphi'(x)$. La première intégrale vaut

$$\begin{aligned}
\int_a^{y_0} \left(\int_{y_0}^{x} u'(t)\varphi'(x)\right) &= \int_a^{y_0} \left(\int_{y_0}^{a} f(t,x)\mathbb{1}_{t<x}(t,x)dt\right)dx && \text{(33.9a)}\\
&= \int_{y_0}^{a}\int_a^{y_0} f(t,x)\mathbb{1}_{t>x}dxdt && \text{(33.9b)}\\
&= \int_{y_0}^{a}\int_a^{t} f(t,x)dxdt && \text{(33.9c)}\\
&= -\int_a^{y_0}\int_a^{t} u'(t)\varphi'(x)dx\,dt && \text{(33.9d)}
\end{aligned}$$

La permutation d'intégrales pour obtenir (33.9b) est due au théorème de Fubini 16.239(3). Par le même petit jeu, la seconde intégrale vaut

$$\int_{y_0}^{b}\int_t^{b} u'(t)\varphi'(x)dx\,dt. \tag{33.10}$$

En refaisant la somme,

$$\begin{aligned}
\int_I \bar{u}\varphi' &= -\int_a^{y_0} u'(t)\left(\int_a^t \varphi'(x)dx\right)dt + \int_{y_0}^{b} u'(t)\left(\int_t^b \varphi'(x)dx\right)dt && \text{(33.11a)}\\
&= -\int_a^{y_0} u'(t)\big(\varphi(t)-\varphi(a)\big)dt + \int_{y_0}^{b} u'(t)\big(\varphi(b)-\varphi(t)\big) && \text{(33.11b)}\\
&= -\int_a^b u'\varphi && \text{(33.11c)}\\
&= -\int_I u'\varphi. && \text{(33.11d)}
\end{aligned}$$

Notons que $\varphi(a) = \varphi(b) = 0$ parce que φ est à support compact dans $]a,b[$. Nous avons donc prouvé que $\bar{u}$ est dans $H^1(I)$ et que $\bar{u}' = u'$. Par le corollaire 33.2, nous avons une constante C telle que $\bar{u} = u + C$ presque partout, c'est à dire $u = \bar{u} + C$ dans $H^1(I)$.

En résumé, $\tilde{u}\tilde{u} = \bar{u} + C$ est un représentant continu de u dans $L^2(I)$.

L'unicité du représentant continu est simplement le fait que deux fonctions continues égales presque partout sont égales (proposition 22.147).

□

Proposition 33.4.
Si $u \in H^1(I)$, alors

$$u(x) - u(y) = \int_y^x u' \tag{33.12}$$

pour tout $x, y \in I$.

Démonstration. Pour fixer les idées, nous supposons $x < y$. Nous considérons une suite $\varphi_n \in C_c^\infty(I)$ convergeant uniformément sur I vers $\mathbb{1}_{[x,y]}$. Nous exigeons de plus que

— φ_n' est positive sur $[a, x + \frac{1}{n}]$
— φ_n' est négative sur $[y - \frac{1}{n}, b]$
— $\varphi_n = 1$ sur $[x + \frac{1}{n}, y - \frac{1}{n}]$.
— $\varphi_n = 0$ sur $[a, x - 1/n]$ et sur $[y + 1/n, b]$.

Pour chaque n, nous découpons l'intégrale comme

$$-\int_I u'\varphi_n = \int_I u\varphi_n' = \int_a^{a-1/n} u\varphi_n' + \int_{x-1/n}^{x+1/n} u\varphi_n' + \int_{x+1/n}^{y-1/n} u\varphi_n' + \int_{y-1/n}^{y+1/n} u\varphi_n' + \int_{y+1/n}^{b} u\varphi_n'. \tag{33.13}$$

Par construction de φ_n, de ces 5 morceaux, il n'en reste que deux de non nulles :

$$\int_I u\varphi' = \underbrace{\int_{x-1/n}^{x+1/n} u(t)\varphi_n'(t)dt}_{A} + \underbrace{\int_{y-1/n}^{y+1/n} u(t)\varphi_n'(t)dt}_{B} \tag{33.14}$$

Soit $\epsilon > 0$ et n suffisamment grand pour avoir $u(t) \in B\big(u(x), \epsilon\big)$ pour tout $t \in B(x, \frac{1}{n})$ et (en même temps) $u(t) \in B\big(u(y), \epsilon\big)$ pour tout $t \in B(y, \frac{1}{n})$. C'est la continuité de u qui permet de trouver un tel n. Pour cette valeur de n, en tenant compte des hypothèses sur la positivité de φ_n' nous avons

$$\int_{x-1/n}^{x+1/n} \big(u(x) - \epsilon\big)\varphi_n'(t)dt \leqslant \int_{x-1/n}^{x+1/n} u(t)\varphi_n'(t)dt \leqslant \int_{x-1/n}^{x+1/n} \big(u(x) + \epsilon\big)\varphi_n'(t)dt, \tag{33.15}$$

mais par hypothèse sur φ_n nous trouvons

$$\int_{x-1/n}^{x+1/n} \varphi_n'(t)dt = \varphi_n(x + \frac{1}{n}) - \varphi(x + \frac{1}{n}) = 1. \tag{33.16}$$

donc

$$u(x) - \epsilon \leqslant \int_{x-1/n}^{x+1/n} u(t)\varphi_n'(t)dt \leqslant u(x) + \epsilon. \tag{33.17}$$

Pour encadrer la seconde, il faut être plus prudent avec les signes parce que φ_n' y est négative. En posant $\psi_n = -\varphi_n$ nous avons

$$-B = \int_{y-1/n}^{y+1/n} u(t)\psi_n(t)dt, \tag{33.18}$$

et donc

$$u(y) - \epsilon \leqslant -B \leqslant u(y) + \epsilon \tag{33.19}$$

ou encore

$$-\epsilon - u(y) \leqslant B \leqslant \epsilon - u(y). \tag{33.20}$$

En additionnant avec (33.17) nous voyons que pour tout $\epsilon > 0$ il existe un $N(\epsilon)$ tel que nous ayons

$$u(x) - u(y) - 2\epsilon \leqslant \int_I u'\varphi_n \leqslant u(x) - u(y) + 2\epsilon \tag{33.21}$$

pour tout $n \geqslant N$. Nous voulons évidemment prendre la limite $\epsilon \to 0$, c'est à dire $n \to \infty$. Étant donné que $\varphi_n(t) < 1$ pour tout t et pour tout n, la fonction $t \mapsto u'(t)\varphi_n(t)$ est dominée par u',

qui est dans $L^1(I)$ par le lemme 28.25. Le théorème de la convergence dominée nous permet donc d'affirmer que

$$\lim_{n\to\infty}\int_I u'\varphi_n = \int_I u'\mathbb{1}_{[x,y]} = \int_x^y u', \tag{33.22}$$

et donc les inégalités (33.21) donnent le résultat, grâce au signe dans (33.13). □

Corollaire 33.5.
Si $[u] \in H^1(I)$, le représentant continu $u \in C^0(I)$ peut être prolongé par continuité en $u \in C^0(\bar{I})$.

Démonstration. Soit (x_n) une suite strictement croissante dans $]a,b[$ convergeant vers b. Nous voulons montrer que la suite $\big(u(x_n)\big)$ est de Cauchy dans $\mathbb{R}$, ce qui nous permettra de définir

$$u(b) = \lim_{n\to\infty} u(x_n). \tag{33.23}$$

qui sera évidemment continue. Cette construction ne dépendra pas du choix de la suite (x_n) parce que deux fonctions continues sur $\bar{I}$ et égales sur I sont égales sur $\bar{I}$.

En notant u' la dérivée de u dans H^1, nous avons par construction du représentant continu : $u(x) = \int_{y_0}^x u'(t)dt$. Et donc

$$\big|u(x_n) - u(x_{n+p})\big| = \left|\int_{y_0}^{x_n} u' - \int_{y_0}^{x_{n+p}} u'\right| = \left|\int_{x_n}^{x_{n+p}} u'\right|. \tag{33.24}$$

Vu que la suite (x_n) est de Cauchy et que u' est intégrable (même sur $\bar{I}$), la limite $n \to \infty$ de cela est zéro, quelle que soit la valeur de p. Donc $\big(u(x_n)\big)$ est ce Cauchy dans $\mathbb{R}$ et est donc convergente. □

Proposition 33.6 ([4, 1])**.**
Quelques propriétés de l'espace de Sobolev $H^1(I)$ où $I =]a,b[$ est un ouvert borné de $\mathbb{R}$.

(1) $H^1(I)$ est un espace de Hilbert.
(2) $H^1(I)$ s'injecte de façon compacte dans $C^0(\bar{I})$.
(3) $H^1(I)$ s'injecte de façon continue dans $L^2(I)$.

Démonstration. Nous prouvons point par point.

(1) Le seul critère à vérifier est la complétude. Pour cela nous considérons une suite de Cauchy (u_n) dans $H^1(I)$. Si $\epsilon > 0$, alors il existe $N > 0$ tel que pour tout $p \geqslant 0$ nous ayons $\|u_{n+p} - u_n\|_{H^1}^2 \leqslant \epsilon$, c'est à dire

$$\|u_{n+p} - u_n\|_{L^2}^2 + \|u'_{n+p} - u'_n\|_{L^2}^2 + \tag{33.25}$$

En particulier les suites (u_n) et (u'_n) sont de Cauchy dans L^2 qui est complet par le théorème de Fischer-Riesz 28.31. Nous notons donc

$$u_n \xrightarrow{L^2} u \tag{33.26a}$$
$$u'_n \xrightarrow{L^2} v. \tag{33.26b}$$

Nous allons maintenant montrer quelques limites.

$\underline{u_n\varphi \xrightarrow{L^2} u\varphi}$ Si M est une constante qui majore φ alors $\|u_n\varphi - u\varphi\|_2 \leqslant M\|u_n - u\|_2 \to 0$.

$\underline{u'_n\varphi \xrightarrow{L^2} v\varphi}$ C'est la même chose avec $\|u'_n\varphi - v\varphi\|_2 \leqslant M\|u'_n - v\|_2 \to 0$.

$\underline{u \in H^1(I) \text{ avec } u' = v}$ Attendu le corollaire 28.26 qui permet de permuter intégrale et limite dans $L^2(I)$ et les limites que nous venons de prouver,

$$\int_I u\varphi' = \lim_{n\to\infty}\int_I u_n\varphi' = -\lim_{n\to\infty}\int_I u'_n\varphi = -\int_I v\varphi. \tag{33.27}$$

Cela signifie que v est la dérivée faible de u : $u' = v$.

$\underline{u_n \xrightarrow{H^1} u}$ Nous pouvons alors prouver que $u_n \to u$ dans $H^1(I)$:

$$\|u_n - u\|^2_{H^1(I)} = \|u_n - u\|^2_{L^2} + \|u_n' - u'\|^2_{L^2}. \tag{33.28}$$

Mais nous savons déjà que $u_n \to u$ dans L^2 (d'ailleurs c'est la définition de u) et que $u' = v$ alors que par définition de v, nous avons $u_n' \to v$ dans L^2.

Tout cela donne que $u_n \to u$ dans $H^1(I)$ et donc que $H^1(I)$ est un espace complet.

(2) L'application que nous allons prouver être compacte entre $H^1(I)$ et $C^0(\bar{I})$ est

$$\begin{aligned} \psi\colon H^1(I) &\to C^0(\bar{I}) \\ [u] &\mapsto \tilde{u} \end{aligned} \tag{33.29}$$

où $[u]$ désigne une classe de fonction dans $H^1(I)$ et $\tilde{u}$ est son représentant continu prolongé par continuité à $\bar{I}$[1], qui existe par le lemme 33.3 et le corollaire 33.5. Cette application est une injection par l'unicité du représentant continu. Nous allons prouver que c'est une application compacte en utilisant le critère (2) de la proposition 28.4. Pour cela nous allons commencer par utiliser le théorème d'Ascoli sur l'ensemble $\tilde{\mathcal{B}}$ des représentants continus des éléments de $\mathcal{B}$, prolongés par continuité sur $\bar{I}$; c'est à dire $\tilde{B} \subset C^0(\bar{I})$.

Soit $u \in \tilde{\mathcal{B}}$; par la proposition 33.4, nous avons

$$\begin{aligned} \big|u(x) - u(y)\big| &= \Big| \int_y^x u'(t)dt \Big| && \text{(33.30a)} \\ &= \left| \int_I \mathbb{1}_{[x,y]}(t) u'(t) dt \right| && \text{(33.30b)} \\ &\leqslant \|\mathbb{1}_{[x,y]}\|_{L^2} \|u'\|_{L^2} && \text{(33.30c)} \\ &\leqslant \sqrt{|x-y|} \|u'\|_{H^1} && \text{(33.30d)} \\ &\leqslant \sqrt{|x-y|}. && \text{(33.30e)} \end{aligned}$$

où nous insistons sur le fait que la continuité n'impliquant pas la dérivabilité, le u' ici est la dérivé au sens de H^1, et non la dérivée usuelle. Quoi qu'il en soit, l'ensemble $\tilde{\mathcal{B}}$ est équicontinu[2]. Nous montrons à présent qu'il est également borné pour la norme uniforme. Soit $u \in \tilde{\mathcal{B}}$; vu la construction du représentant continu au lemme 33.3, nous avons

$$\begin{aligned} \big|u(x)\big| &= \left| \frac{1}{b-a} \int_a^b u(x) dy \right| && \text{(33.31a)} \\ &= \left| \frac{1}{b-a} \int_a^b \left(\int_y^x u'(t)dt - u(y) \right) dy \right| && \text{(33.31b)} \\ &= \left| \frac{1}{b-a} \int_a^b \int_y^x u'(t) dt dy - \frac{1}{b-a} \int_a^b u(y) dy \right| && \text{(33.31c)} \\ &\leqslant \frac{1}{b-a} \int_a^b \int_a^b |u'(t)| dt\, dy + \frac{1}{b-a} \int_a^b |u(y)| dy. && \text{(33.31d)} \end{aligned}$$

À ce niveau, il faut remarquer que dans la première intégrale, le passage de la valeur absolue à l'intérieur de l'intégrale en même temps que l'élargissement des bornes n'a rien d'innocent. Si $x < y$, les bornes ne sont pas « dans le bon ordre » et nous ne pouvons pas faire la majoration usuelle en entrant simplement la valeur absolue. Ici nous tenons compte de cela en élargissant les bornes, et en les mettant dans le bon ordre. Le passage exact est le suivant : si $x, y \in]a, b[$, nous avons

$$\left| \int_y^x f(t)dt \right| \leqslant \left| \int_y^x |f(t)|dt \right| \leqslant \left| \int_a^b |f(t)|dt \right| = \int_a^b |f(t)|dt. \tag{33.32}$$

1. Encore que par soucis d'économie d'encre nous n'allons pas écrire toujours les tildes et noter u le représentant continu prolongé à $\bar{I}$ par le corollaire 33.5.
2. Définition 10.69.

Notons en particulier que dans le cas du passage vers l'équation (33.31d), le nombre x est fixé alors que y est une variable d'intégration. Donc l'ordre des deux est certainement de temps en temps le « mauvais ».

Quoi qu'il en soit, la première intégrale se réduit à une multiplication par $b-a$ et le calcul continue :

$$|u(x)| \leqslant \int_I |u'(t)|dt + \frac{1}{b-a}\int_I |u| \tag{33.33a}$$
$$\leqslant \sqrt{b-a}\|u'\|_{L^2} + \frac{1}{\sqrt{b-a}}\|u\|_{L^2} \tag{33.33b}$$
$$\leqslant \left(\sqrt{b-a} + \frac{1}{\sqrt{b-a}}\right)\left(\|u'\|_{L^2} + \|u\|_{L^2}\right) \tag{33.33c}$$
$$\leqslant \left(\sqrt{b-a} + \frac{1}{\sqrt{b-a}}\right)\|u\|_{H^1} \tag{33.33d}$$
$$= \sqrt{b-a} + \frac{1}{\sqrt{b-a}}. \tag{33.33e}$$

Donc $\tilde{\mathcal{B}}$ est borné pour la norme L^∞. Et c'est même borné par un nombre facilement calculable connaissant I. En particulier l'ensemble

$$\{u(x) \text{ tel que } u \in H^1\} \tag{33.34}$$

est pour, tout x, contenu dans la boule de rayon $\sqrt{a-b} + \frac{1}{\sqrt{a-b}}$ et donc est relativement compact dans $\mathbb{R}$. Par conséquent le théorème d'Ascoli 28.6 nous dit que l'ensemble $\tilde{B}$ est relativement compact dans $C^0(I)$.

Par conséquent nous avons montré que l'image par ψ de la boule unité fermée $\mathcal{B}$ de $H^1(I)$ est relativement compacte dans $C^0(\bar{I})$, ce qui signifie que ψ est une application compacte.

(3) Les éléments de $H^1(I)$ sont des éléments de $L^2(I)$; donc l'identité est une injection. Nous devons seulement étudier la continuité. Si (u_n) est une suite dans H^1 convergeant dans H^1 vers u, alors

$$\|u_n - u\|_{L^2} \leqslant \|u_n - u\|_{L^2} + \|u_n' - u'\|_{L^2} = \|u_n - u\|_{H^1} \to 0. \tag{33.35}$$

Donc la suite des images (par l'identité) converge dans L^2. L'identité est donc continue.

□

33.1.2 Sur un ouvert de $\mathbb{R}^n$

Soit Ω, un ouvert de $\mathbb{R}^n$ et $v \in L^2(\Omega)$ (définition 28.57). Les fonctions considérées sont à valeurs réelles.

33.1.2.1 Définition

Définition 33.7 (Espace de Sobolev $H^1(\Omega)$)**.**
Soit Ω une partie de $\mathbb{R}^n$. L'espace de ***Sobolev*** *$H^1(\Omega)$ est :*

$$H^1(\Omega) = \{v \in L^2(\Omega) \text{ tel que } \forall i = 1, \dots, n, \partial_i v \in L^2(\Omega)\}. \tag{33.36}$$

Nous munissons cet espace d'un produit scalaire

$$(u, v)_{H^1} = \langle u, v\rangle_{L^2} + \langle \nabla u, \nabla v\rangle_{L^2}, \tag{33.37}$$

où $\nabla u = \sum_i \partial_i u \in L^2$.

L'existence des intégrales dans le produit scalaire est assurée par le fait que u, v, ∇u et ∇v sont dans $L^2(\Omega)$. La définition du produit scalaire dans L^2 est la définition 28.176 (mais sans la conjugaison complexe).

Pour la même raison, $(u,u)_{H^1}=0$ demande que chacun des deux termes est séparément nul, et nous avons $u=0$ dans L^2, et donc aussi dans H^1.

Théorème 33.8 ([376]).
L'espace $H^1(\Omega)$ est un espace de Hilbert[3].

Démonstration. Nous devons nous assurer que l'espace H^1 est complet. Pour cela nous considérons une suite de Cauchy (u_n) dans H^1. Soit $\epsilon>0$; il existe $N>0$ tel que si $n,m>N$ alors $\|u_n-u_m\|_{H^1}<\epsilon$. Dans ce cas nous avons en particulier

$$\|u_m-u_n\|_{H^1}^2=(u,u)_{H^1}=\langle u,u\rangle+\langle\nabla u,\nabla u\rangle=\|u\|_{L^2}^2+\|\nabla u\|_{L^2}^2, \tag{33.38}$$

et en particulier les suites (u_n) et (∇u_n) sont de Cauchy dans L^2. Vu que L^2, lui, est complet (théorème 28.30), il existe $u\in L^2$ et $v_i\in L^2$ tels que

$$u_n\xrightarrow{L^2}u \tag{33.39a}$$
$$\partial_i u_n\xrightarrow{L^2}v_i. \tag{33.39b}$$

Nous savons que l'injection $i\colon L^2\to\mathscr{D}'$ est continue par la proposition 32.18. Nous avons donc aussi les limites

$$T_{u_n}\xrightarrow{\mathscr{D}'}T_u \tag{33.40a}$$
$$T_{\partial_i u_n}\xrightarrow{\mathscr{D}'}T_{v_i}. \tag{33.40b}$$

La dérivée étant une opération continue sur $\mathscr{D}'$ nous avons de plus

$$\partial_i(T_{u_n})\xrightarrow{\mathscr{D}'}\partial_i(T_u) \tag{33.41}$$

En utilisant le lemme 32.21 nous avons alors

$$T_{\partial_i u_n}=\partial_i(T_{u_n})\xrightarrow{\mathscr{D}}\partial_i(T_u)=T_{\partial_i u}. \tag{33.42}$$

En comparant avec (33.40b) et par l'unicité de la limite, nous avons $T_{v_i}=T_{\partial_i u}$. Cela implique $v_i=\partial_i u$.

Vu que $v_i\in L^2$ nous avons aussi $\partial_i u\in L^2$. Par conséquent $u\in H^1(\Omega)$ parce que ses dérivées sont dans L^2.

Nous devons maintenant prouver que $u_n\xrightarrow{H^1}u$. Nous avons

$$\|u_n-u\|_{H^1}=\|u_n-u\|_{L^2}+\|\nabla u_n-\nabla u\|_{L^2} \tag{33.43}$$

Le premier terme tend vers zéro parce que $u_n\xrightarrow{L^2}u$ et le second parce que $\partial_i u_n\xrightarrow{L^2}\partial_i u$. □

33.1.3 Espace de Sobolev fractionnaire

Définition 33.9.
*Pour $m\in\mathbb{N}$ et un ouvert Ω de $\mathbb{R}^d$ nous définissons l'**espace de Sobolev***

$$H^m(\Omega)=\{u\in L^2(\Omega)\text{ tel que }\partial^\alpha u\in L^2(\Omega)\,\forall|\alpha|\leqslant m\}. \tag{33.44}$$

Nous définissons également un produit scalaire sur H^m par

$$(u,v)_{H^m}=\sum_{|\alpha|\leqslant m}\langle\partial^\alpha u,\partial^\alpha v\rangle_{L^2}. \tag{33.45}$$

3. Définition 27.2.

En particulier la topologie est celle de la norme dérivée du produit scalaire :

$$\|u\|_{H^m(\Omega)} = \sum_{|\alpha|\leqslant m} \|\partial^\alpha u\|_{L^2(\Omega)}. \tag{33.46}$$

Le lemme suivant montre que la proposition 31.13 fonctionne encore avec L^2 au lieu de $\mathscr{S}$.

Lemme 33.10 (Lemme de transfert[377], thème 64)**.**
Soit $f \in H^m(\mathbb{R}^d)$. Alors pour tout multiindice α avec $|\alpha| \leqslant m$ nous avons

$$\mathcal{F}(\partial^\alpha f) = \big[\xi \mapsto i^{|\alpha|}\xi^\alpha \hat{f}(\xi)\big]. \tag{33.47}$$

Lemme 33.11.
Il existe des constantes c_1 et c_2 telles que pour tout $x \in \mathbb{R}^d$,

$$c_1(1+\|x\|^2)^m \leqslant \sum_{|\alpha|\leqslant m} (x^\alpha)^2 \leqslant c_2(1+\|x\|^2)^m. \tag{33.48}$$

Lemme 33.12.
Soit $u \in L^2(\mathbb{R}^d)$. Nous avons $u \in H^m(\mathbb{R}^d)$ si et seulement si l'application

$$\xi \mapsto \big(1+|\xi|^2\big)^{k/2}\hat{u} \tag{33.49}$$

est dans $L^2(\mathbb{R}^d)$ pour tout $k \leqslant m$. Ici $|\xi|$ est la norme euclidienne de ξ dans $\mathbb{R}^d$.

Démonstration. Vu le lemme 33.10, il suffit de montrer que

$$\big(1+|\xi|^2\big)^{k/2}\hat{u} \tag{33.50}$$

est dans L^2 pour tout $k \leqslant m$ si et seulement si

$$\xi^\alpha \hat{u} \tag{33.51}$$

l'est pour tout α avec $|\alpha| \leqslant m$.

L'expression (33.50) est une somme d'expressions du type (33.51). Donc l'implication dans un sens est montrée. Pour l'autre sens, nous savons que

$$\xi^\alpha = \xi_1^{\alpha_1}\dots\xi_n^{\alpha_n}, \tag{33.52}$$

et donc

$$|\xi^\alpha| \leqslant |\xi_1|^{\alpha_1}\dots|\xi_n|^{\alpha_n}. \tag{33.53}$$

Or $|\xi|^{|\alpha|} = |\xi|^{\sum_i \alpha_i} = |\xi|^{\alpha_1}\dots|\xi|^{\alpha_n}$ et $|\xi| \geqslant |\xi_i|$ pour tout i, donc

$$|\xi^\alpha| \leqslant |\xi|^{|\alpha|}. \tag{33.54}$$

D'autre part pour tout $x \in \mathbb{R}^+$ et tout k positif nous avons

$$(1+x^2)^{k/2} \geqslant x^k \tag{33.55}$$

qui est facile à vérifier en prenant le carré des deux membres.

En remettant tout ensemble,

$$|\xi^\alpha\hat{u}| \leqslant |\xi^\alpha||\hat{u}| \leqslant |\xi|^{|\alpha|}|\hat{u}| \leqslant \big(1+|\xi|^2\big)^{|\alpha|/2}|\hat{u}|. \tag{33.56}$$

Donc si le membre de droite est de carré intégrable, celui de gauche l'est également. □

Définition 33.13 (Espace de Sobolev H^s[378]).
Pour $s > 0$ nous définissons l'espace de Sobolev $H^s(\mathbb{R}^d)$ par

$$H^s(\mathbb{R}^d) = \{u \in L^2(\mathbb{R}^d) \text{ tel que } \big(1 + \|\xi\|^2\big)^{s/2}\hat{u} \in L^2(\mathbb{R}^d)\}. \tag{33.57}$$

Nous y mettons le produit scalaire

$$(u, v)_{H^s} = \int_{\mathbb{R}^d} \hat{u}(\xi)\overline{\hat{v}(\xi)}(1 + \|\xi\|^2)^s d\xi. \tag{33.58}$$

33.14.
Vu que $\mathscr{D}(\mathbb{R}^d)$ est dense dans $L^2(\mathbb{R}^d)$ (théorème 28.34), on pourrait croire à la densité a fortiori dans $H^s(\mathbb{R}^d)$. Mais attention : $\mathscr{D}(\mathbb{R}^d)$ est dense dans L^2 pour la norme L^2. Nous n'avons encore rien dit pour la norme $H^s(\mathbb{R}^d)$.

Proposition 33.15 ([379]).
La partie $\mathscr{S}(\mathbb{R}^d)$ est dense dans $H^s(\mathbb{R}^d)$.

Démonstration. Soit $u \in H^s(\mathbb{R}^d)$. Par définition l'application

$$\xi \mapsto (1 + \|\xi\|^2)^{s/2}\hat{u} \tag{33.59}$$

est dans $L^2(\mathbb{R}^d)$. Elle peut donc être approximée au sens L^2 par des fonctions dans $\mathscr{D}(\mathbb{R}^d)$ (théorème 28.34(5)), c'est à dire qu'il existe des fonctions $\phi_n \in \mathscr{D}(\mathbb{R}^d)$ telles que

$$\phi_n \xrightarrow{L^2(\mathbb{R}^d)} (1 + \|\xi\|^2)^{s/2}\hat{u}. \tag{33.60}$$

Nous posons

$$\psi_n = \frac{\phi_n}{(1 + \xi^2)^{s/2}} \tag{33.61}$$

Cela est encore une fonction de $\mathscr{D}(\mathbb{R}^d)$, et donc de $\mathscr{S}(\mathbb{R}^d)$. Vu que la transformée de Fourier est une bijection de $\mathscr{D}(\mathbb{R}^d)$ (proposition 31.14), nous pouvons considérer une suite $\varphi_n \in \mathscr{D}(\mathbb{R}^d)$ telle que $\hat{\varphi}_n = \psi_n$, et nous allons montrer que $\varphi_n \xrightarrow{H^s(\mathbb{R}^d)} u$.

Nous avons :

$$\begin{aligned}
\|\varphi_n - u\|_{H^s}^2 &= \int_{\mathbb{R}^d} |\hat{\varphi}_n - \hat{u}|^2(1 + \xi^2)^s d\xi && \text{(33.62a)}\\
&= \int_{\mathbb{R}^d} \Big|\frac{\phi_n(\xi)}{(1 + \xi^2)^{s/2}} - \hat{u}(\xi)\Big|^2(1 + \xi^2)^s d\xi && \text{(33.62b)}\\
&= \int_{\mathbb{R}^d} |\phi_n(\xi) - \hat{u}(\xi)(1 + \xi^2)^{s/2}|^2 d\xi && \text{(33.62c)}\\
&= \|\phi_n - (1 + \xi^2)^{s/2}\hat{u}\|_{L^2}^2. && \text{(33.62d)}
\end{aligned}$$

Par définition de la suite ϕ_n nous avons donc bien

$$\|\varphi_n - u\|_{H^s}^2 = \|\phi_n - (1 + \xi^2)^{s/2}\hat{u}\|_{L^2}^2 \to 0. \tag{33.63}$$

Notons que même si ϕ_n est dans $\mathscr{D}(\mathbb{R}^d)$, nous n'avons pas prouvé la convergence $\phi_n \xrightarrow{H^s} u$, mais bien $\varphi_n \xrightarrow{H^s} u$. Or les fonctions φ_n sont dans $\mathscr{S}(\mathbb{R}^d)$, et rien n'assure qu'elles soient à support compact. Nous avons donc bien prouvé la densité de $\mathscr{S}$ et non celle de $\mathscr{D}$. □

Remarque 33.16.
Pour qui a tout compris, cela peut sembler une évidence, mais nous précisions que nous parlons de densité de $\mathscr{S}(\mathbb{R}^d)$ dans $H^s(\mathbb{R}^d)$, à aucun moment la topologie de $\mathscr{S}(\mathbb{R}^d)$ n'entre en compte.

Un peu moins évident : ce que nous avons réellement montré est la densité de $\iota\big(\mathscr{D}(\mathbb{R}^d)\big)$ dans $H^s(\mathbb{R}^d)$ où ι est l'application « prise de classe ». Nous n'avons pas insisté là-dessus, mais il faut

dire que dans la preuve de la proposition 33.15, u est un représentant d'un élément choisi dans $H^s(\mathbb{R}^d)$.

Nous avons ensuite prouvé la convergence $\|\varphi_n - u\|_{H^s(\mathbb{R}^d)} \to 0$ qui est une convergence d'une suite dans $\mathbb{R}$, et dans laquelle l'opération $\|.\|_{H^s}$ est définie sur un espace de fonctions et n'est pas une norme (c'est pour que cela devienne une norme que l'on prend les classes).

Nous en avons déduit la convergence $\varphi_n \xrightarrow{H^s(\mathbb{R}^d)} u$ où maintenant φ_n et u sont des classes dans $H^s(\mathbb{R}^d)$.

Proposition 33.17.
La partie $\mathscr{D}(\mathbb{R}^d)$ est dense dans $\big(H^s(\mathbb{R}^d), \|.\|_{H^s(\mathbb{R}^d)}\big)$.

Démonstration. Nous savons déjà que $\mathscr{D}(\mathbb{R}^d)$ est dense dans $H^s(\mathbb{R}^d)$ par la proposition 33.15. Nous devons seulement prouver que $\mathscr{D}(\mathbb{R}^d)$ est dense dans $\big(\mathscr{S}(\mathbb{R}^d), \|.\|_{H^2(\mathbb{R}^d)}\big)$. Pour cela nous utilisons la densité de $\mathscr{D}(\mathbb{R}^d)$ dans $\mathscr{S}(\mathbb{R}^d)$ de la proposition 28.97. Soit donc $f \in \mathscr{D}(\mathbb{R}^d)$ et une suite f_k dans $\mathscr{D}(\mathbb{R}^d)$ telle que

$$f_k \xrightarrow{\mathscr{S}(\mathbb{R}^d)} f. \tag{33.64}$$

Vu que la transformée de Fourier est continue sur $\mathscr{S}(\mathbb{R}^d)$ (proposition 31.14) nous avons aussi

$$\hat{f}_k \xrightarrow{\mathscr{S}(\mathbb{R}^d)} \hat{f}, \tag{33.65}$$

et en particulier pour tout polynôme P nous avons la convergence uniforme

$$P\hat{f}_k \xrightarrow{unif} P\hat{f}. \tag{33.66}$$

D'autre part la fonction $\xi \mapsto |\hat{f}_k(\xi) - \hat{f}(\xi)|^2(1+\xi^2)^s$ est Schwartz et en tout point décroissante en k. Soientt $\epsilon > 0$ et $r > 0$ choisis de telle sorte à avoir

$$\int_{\|\xi\|>r} |\hat{f}_k(\xi) - \hat{f}(\xi)|(1+\xi^2)^s d\xi < \epsilon. \tag{33.67}$$

pour tout k. La convergence uniforme (33.66) permet de considérer k_0 tel que pour tout $k > k_0$,

$$|\hat{f}_k - \hat{f}|(1+\xi^2)^s < \frac{\epsilon}{\mathrm{Vol}\big(B(0,r)\big)} \tag{33.68}$$

dans $B(0,r)$. Avec tout cela, dès que $k > k_0$ nous avons

$$\|f_k - f\|_{H^s(\mathbb{R}^d)} = \int_{\mathbb{R}} |\hat{f}_k - \hat{f}|(1+\xi^2)^s d\xi = \int_{B(0,r)} \ldots + \int_{\|\xi\|>r} \ldots \leqslant 2\epsilon. \tag{33.69}$$

Donc nous avons bien $\|f_k - f\|_{H^s(\mathbb{R}^d)} \to 0$ et convergence de f_k vers f dans $H^s(\mathbb{R}^d)$. □

33.2 Trace

Définition 33.18 ([378]).
*Nous définissons la **trace** d'une fonction par*

$$\begin{aligned} \gamma_0 \colon \mathscr{D}(\mathbb{R}^d) &\to \mathscr{D}(\mathbb{R}^{d-1}) \\ (\gamma_0 v)(x_1, \ldots, x_{d-1}) &= v(x_1, \ldots, x_{d-1}, 0). \end{aligned} \tag{33.70}$$

Théorème 33.19 ([380, 378]).
Si $s > \frac{1}{2}$, alors γ_0 accepte une unique extension en opérateur linéaire borné

$$\gamma_0 \colon H^s(\mathbb{R}^d) \to H^{s-\frac{1}{2}}(\mathbb{R}^{d-1}). \tag{33.71}$$

Démonstration. Nous subdivisons la preuve en plusieurs pas.

Une inégalité pour $\varphi \in \mathscr{D}(\mathbb{R}^d)$ Nous commençons par considérer $v \in \mathscr{D}(\mathbb{R}^d)$ (fonction C^∞ à support compact). Nous allons alors prouver que

$$\|\gamma_0\varphi\|_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})} \leqslant K\|\varphi\|_{H^s(\mathbb{R}^d)} \tag{33.72}$$

pour une certaine constante K (qui ne dépend en particulier pas de φ).
Nous avons

$$\begin{aligned}
\|\gamma_0\varphi\|^2_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})} &= (\gamma_0\varphi, \gamma_0\varphi)_{H^{s-\frac{1}{2}}} && (33.73a)\\
&= \int_{\mathbb{R}^{d-1}} |\widehat{\gamma_0\varphi(\xi)}|^2(1+\|\xi\|^2)^{s-\frac{1}{2}}d\xi && (33.73b)\\
&= \int_{\mathbb{R}^{d-1}} \Big|\int_{\mathbb{R}^{d-1}} (\gamma_0\varphi)(x)e^{-i\xi x}dx\Big|^2(1+\|\xi\|^2)^{s-\frac{1}{2}}d\xi && (33.73c)\\
& && (33.73d)
\end{aligned}$$

Nous appliquons la trace en appliquant la formule du corollaire 31.23,

$$(\gamma_0\varphi)(x) = \varphi(x,0) = \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-iky}\varphi(x,y)dy\,dk \tag{33.74}$$

En remplaçant dans (33.73c) nous avons

$$\|\gamma_0\varphi\|^2_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})} = \frac{1}{2\pi}\int_{\mathbb{R}^{d-1}} \Big|\int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-iky}e^{-i\xi x}\varphi(x,y)dy\,dk\,dx\Big|^2(1+\|\xi\|^2)^{s-\frac{1}{2}}d\xi. \tag{33.75}$$

Nous voudrions permuter les intégrales en k et en x. Pour cela nous étudions la fonction $u\colon \mathbb{R}\times\mathbb{R}^{d-1}\to\mathbb{C}$ donnée par

$$u(k,x) = e^{-ikx}\int_{\mathbb{R}} e^{-i\xi x}\varphi(x,y)dy \tag{33.76}$$

Effectuer l'intégrale par rapport à y revient à calculer la transformée de Fourier partielle dont nous parlons dans la proposition 31.15 [4]. Elle est donc une fonction Schwartz de k et de x (conjointement et non seulement séparément) et est donc dans $L^1(\mathbb{R}\times\mathbb{R}^{d-1})$. Les intégrales sut k et sur x peuvent donc être réunies et permutées par le théorème de Fubini 16.239 (n'oubliez tout de même pas de vous convaincre que la condition (2) est remplie).
Nous avons donc

$$\|\gamma_0\varphi\|^2_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})} = \frac{1}{2\pi}\int_{\mathbb{R}^{d-1}} \Big|\int_{\mathbb{R}}\int_{\mathbb{R}^{d-1}}\int_{\mathbb{R}} e^{-iky}e^{-i\xi x}\varphi(x,y)dy\,dx\,dk\Big|^2(1+\|\xi\|^2)^{s-\frac{1}{2}}d\xi. \tag{33.77}$$

Étant donné que φ est à support compact, les intégrales sur x et sur y peuvent se réunir en utilisant encore le théorème de Fubini ; ces intégrales donnent :

$$\int_{\mathbb{R}^{d-1}\times\mathbb{R}} e^{-iky}e^{-i\xi x}\varphi(x,y)dx\otimes dy = \int_{\mathbb{R}^{d-1}\times\mathbb{R}} e^{-i(\xi,k)\,\cdot\,(x,y)}\varphi(x,y)dx\otimes dy = \hat{\varphi}(\xi,k). \tag{33.78}$$

Nous restons avec

$$\|\gamma_0\varphi\|^2_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})} = \frac{1}{2\pi}\int_{\mathbb{R}^{d-1}} \Big|\int_{\mathbb{R}} \hat{\varphi}(\xi,k)dk\Big|^2(1+\|\xi\|^2)^{s-\frac{1}{2}}d\xi. \tag{33.79}$$

Nous allons maintenant traiter la partie du milieu :

$$\clubsuit = \Big|\int_{\mathbb{R}} \hat{\varphi}(\xi,k)dk\Big| = \Big|\int_{\mathbb{R}} \hat{\varphi}(\xi,k)(1+\xi^2+k^2)^{s/2}\frac{1}{(1+\xi^2+k^2)^{s/2}}dk\Big| = |\langle f_1, f_2\rangle_{L^2(\mathbb{R}^d)}| \tag{33.80}$$

4. Dont une relecture de la preuve ne serait vraiment pas de trop, ainsi que la preuve de 28.91.

Ici ξ est vu comme une constante et les fonctions f_1 et f_2 sont

$$f_1 : k \to \hat{\varphi}(\xi,k)(1+\xi^2+k^2)^{s/2} \tag{33.81a}$$

$$f_2 : k \to \frac{1}{(1+\xi^2+k^2)^{s/2}} \tag{33.81b}$$

Nous pouvons utiliser l'inégalité de Cauchy-Schwarz 11.6 :

$$\clubsuit \leqslant \left(\int_{\mathbb{R}} |\hat{\varphi}(\xi,k)|^2 (1+\xi^2+k^2)^s dk \right)^{1/2} \left(\int_{\mathbb{R}} \frac{1}{(1+\xi^2+k^2)^s} dk \right)^{1/2} \tag{33.82}$$

Nous notons $g(\xi)$ ce qui se trouve dans la seconde parenthèse (après intégration sur k). Avec cela nous continuons :

$$\|\gamma_0\varphi\|^2_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})} \leqslant \frac{1}{2\pi} \int_{\mathbb{R}^{d-1}} \int_{\mathbb{R}} |g(\xi)||\hat{\varphi}(\xi,k)|^2 (1+\xi^2+k^2)^s (1+\|\xi\|^2)^{s-\frac{1}{2}} dk\, d\xi. \tag{33.83}$$

Vu que $\hat{\varphi}$ est Schwartz, la fonction qui est à l'intérieur des deux intégrales est dans $L^1(\mathbb{R}^{d-1}\times\mathbb{R})$ et nous pouvons réunir les deux intégrales :

$$\|\gamma_0\varphi\|^2_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})} \leqslant \frac{1}{2\pi} \int_{\mathbb{R}\times\mathbb{R}^{d-1}} |g(\xi)||\hat{\varphi}(\xi,k)|^2 (1+\xi^2+k^2)^s (1+\|\xi\|^2)^{s-\frac{1}{2}} dk \otimes d\xi. \tag{33.84}$$

À ce point nous démontrons qu'en réalité la combinaison $g(\xi)(1+\xi^2)^{s-\frac{1}{2}}$ ne dépend pas de ξ. En effet

$$g(\xi)(1+\xi^2)^{s-\frac{1}{2}} = (1+\xi^2)^{s-\frac{1}{2}} \int_{\mathbb{R}} \frac{1}{(1+\xi^2+k^2)} dk \tag{33.85a}$$

$$= \frac{1}{(1+\xi^2)^{1/2}} \int_{\mathbb{R}} \left(\frac{1+\xi^2}{1+\xi^2+k^2} \right)^s dk \tag{33.85b}$$

$$= \frac{1}{(1+\xi^2)^{1/2}} \int \left(\frac{1}{1+\frac{k^2}{1+\xi^2}} \right)^s dk. \tag{33.85c}$$

Nous effectuons le changement de variables $t = \frac{k}{\sqrt{1+\xi^2}}$, $dk = (1+\xi^2)^{1/2}dt$, et le tout vaut

$$\int_{\mathbb{R}} \left(\frac{1}{1+t^2} \right)^s dt, \tag{33.86}$$

qui est effectivement indépendant de ξ. Nous nommons cela K (auquel nous ajoutons le $\frac{1}{2\pi}$) :

$$\|\gamma_0\varphi\|^2_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})} \leqslant K \int_{\mathbb{R}\times\mathbb{R}^{d-1}} |\hat{\varphi}(\xi,k)|^2 (1+\xi^2+k^2)^s dk \otimes d\xi = K\|\varphi\|^2_{H^s(\mathbb{R}^d)}. \tag{33.87}$$

Nous avons donc prouvé pour tout $\varphi \in \mathscr{D}(\mathbb{R}^d)$ (avec redéfinition du K) :

$$\|\gamma_0\varphi\|_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})} \leqslant K\|\varphi\|_{H^s(\mathbb{R}^d)}. \tag{33.88}$$

À propos de classes Il serait tentant de conclure en disant que $\mathscr{D}(\mathbb{R}^d)$ est dense dans $H^s(\mathbb{R}^d)$. Hélas, techniquement, l'ensemble $\mathscr{D}(\mathbb{R}^d)$ n'est même pas un sous-ensemble de $H^s(\mathbb{R}^d)$ parce que ce dernier est un ensemble de *classes* de fonctions. Ce petit détail a ici son importance parce que γ_0 n'est pas une application qui descend aux classes. En effet, $\mathbb{R}^{d-1}$ étant de mesure nulle dans $\mathbb{R}^d$, deux fonctions de la même classe peuvent différer en *tous* les points de $\mathbb{R}^{d-1}$ en même temps.

Si nous notons ι l'application qui consiste à prendre la classe ce qui est dense dans $H^s(\mathbb{R}^d)$, c'est $\iota(\mathscr{D}(\mathbb{R}^d))$. Or chaque classe contient au maximum une seule fonction continue (qui sera même de classe C^∞ à support compact pour les éléments de $\iota(\mathscr{D})$).

L'application γ_0 considérée est l'application composée entre le γ_0 classique et le choix du représentant continu dans la classe. La formule (33.88) que nous venons de prouver est valide pour l'application γ_0 vue comme

$$\gamma_0 : \iota\big(\mathscr{D}(\mathbb{R}^d)\big) \to \mathscr{D}(\mathbb{R}^{d-1}). \tag{33.89}$$

Densité et conclusion Ce que la majoration (33.88) prouve est la continuité de l'application

$$\gamma_0 : \big(\mathscr{D}(\mathbb{R}^d), \|.\|_{H^s(\mathbb{R}^d)}\big) \to \big(H^{s-\frac{1}{2}}(\mathbb{R}^{d-1}), \|.\|_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})}\big). \tag{33.90}$$

Mais la proposition 33.17 nous donne la densité de la partie $\mathscr{D}(\mathbb{R}^d)$ dans $H^s(\mathbb{R}^d)$. La proposition 19.121 nous donne alors une extension

$$\gamma_0 : \big(H^s(\mathbb{R}^d), \|.\|_{H^s(\mathbb{R}^d)}\big) \to \big(H^{s-\frac{1}{2}}(\mathbb{R}^{d-1}), \|.\|_{H^{s-\frac{1}{2}}(\mathbb{R}^{d-1})}\big). \tag{33.91}$$

□

Remarque 33.20.
L'extension n'est pas évidente parce que les éléments de $H^s(\mathbb{R}^d)$ sont en général des classes de fonctions dont les valeurs sur le bord ne sont pas du tout fixées du fait que le bord soit de mesure nulle.

33.3 Théorème de plongement

L'objet des théorèmes de plongement de Sobolev est de montrer que si $s > \frac{d}{2} + k$ alors les éléments de $H^s(\mathbb{R}^d)$ possèdent des représentants de classe C^k. Avant de démontrer le théorème, pour alléger, nous allons donner deux lemmes.

Lemme 33.21.
Soit (u_j) une suite dans $\mathscr{S}(\mathbb{R}^d)$ telle que

$$u_j \xrightarrow{H^s(\mathbb{R}^d)} u \tag{33.92}$$

avec $s > 0$. Alors nous avons aussi la convergence

$$u_j \xrightarrow{L^2(\mathbb{R}^d)} u. \tag{33.93}$$

Démonstration. Vu que $s > 0$ nous avons $(1 + k^2)^s > 1$ (ici nous écrivons k^2 pour $\|k\|^2$). Par conséquent

$$(u, v)_{H^s(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \hat{u}\bar{\hat{v}}(1 + k^2)^s dk \geqslant \int_{\mathbb{R}^d} \hat{u}\bar{\hat{v}} dk = \langle \hat{u}, \hat{v} \rangle_{L^2(\mathbb{R}^d)}. \tag{33.94}$$

Nous avons alors

$$\begin{aligned}
\|u_j - u\|_{L^2} &= \frac{1}{(2\pi)^d} \|\hat{u}_j - \hat{u}\|_{L^2} && \text{(33.95a)}\\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\hat{u}_j - \hat{u}|^2 && \text{(33.95b)}\\
&\leqslant \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\hat{u}_j - \hat{u}|^2 (1 + k^2)^s dk && \text{(33.95c)}\\
&= \frac{1}{(2\pi)^d} \|u_j - u\|_{H^s(\mathbb{R}^d)}. && \text{(33.95d)}
\end{aligned}$$

□

Lemme 33.22.
Soient des fonctions $u_j \in \mathscr{S}(\mathbb{R}^d)$ telles que

$$u_j \xrightarrow{\left(C_0^0(\mathbb{R}^d),\|.\|_\infty\right)} v. \tag{33.96}$$

Alors nous avons la convergence

$$\int_{\mathbb{R}^d} u_j\varphi \to \int_{\mathbb{R}^d} v\varphi \tag{33.97}$$

pour tout $\varphi \in \mathscr{S}(\mathbb{R}^d)$.

Démonstration. La suite (u_j) est équibornée. En effet il existe une queue de suite pour laquelle $\|u_j - v\|_\infty < \epsilon$; cette queue de suite est alors équibornée par $\|v\|_\infty + \epsilon$. Le début de la suite est un nombre fini de fonctions, toutes bornées. Le maximum des bornes donne alors une borne.

Soit donc $M > 0$ tel que $|u_j(x)| < M$ pour tout $x \in \mathbb{R}^d$ et pour tout $j \in \mathbb{N}$. Nous avons alors $|u_j\varphi| < M|\varphi|$ pour tout j et les fonctions $|u_j\varphi|$ sont majorées par la fonction $M|\varphi|$ qui est intégrable. Nous pouvons donc utiliser le théorème de la convergence dominée de Lebesgue 16.140 nous donne

$$\lim_{j\to\infty} \int_{\mathbb{R}^d} u_j\varphi = \int_{\mathbb{R}^d} v\varphi. \tag{33.98}$$

□

Nous pouvons écrire la conclusion du lemme 33.22 sous la forme

$$\langle u_j, \varphi\rangle_{L^2(\mathbb{R}^d)} \to \langle v, \varphi\rangle_{L^2(\mathbb{R}^d)} \tag{33.99}$$

pour tout $\varphi \in \mathscr{S}(\mathbb{R}^r)$ (et non pour tout $\varphi \in L^2(\mathbb{R}^d)$.

Théorème 33.23 (Théorème de Sobolev avec $k = 0$[381])**.**
Soit $s > \frac{d}{2}$ et $u \in H^s(\mathbb{R}^d)$. Alors u possède un représentant dans $C_0^0(\mathbb{R}^d)$ (les fonctions continues et qui s'annulent à l'infini). Nous écrivons cela $H^s(\mathbb{R}^d) \subset C_0^0(\mathbb{R}^d)$.

Démonstration. Nous commençons par supposer que $u \in H^s(\mathbb{R}^d) \cap \mathscr{S}(\mathbb{R}^d)$, et dans ce cas nous notons u le représentant dans $\mathscr{S}(\mathbb{R}^d)$. Nous allons prouver l'inégalité

$$\|u\|_\infty \leqslant c\|u\|_{H^s(\mathbb{R}^d)}. \tag{33.100}$$

La formule d'inversion de Fourier 31.22 appliquée à u_j donne

$$u(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{ikx}\hat{u}(k)dk, \tag{33.101}$$

nous avons alors

$$\begin{aligned}
(2\pi)^d|u(x)| &\leqslant \int_{\mathbb{R}^d} |\hat{u}(k)|dk && \text{(33.102a)}\\
&= \int_{\mathbb{R}^d} (1+k)^{s/2}|\hat{u}(k)|(1+k^2)^{-s+2}dk && \text{(33.102b)}\\
&= \int_{\mathbb{R}^d} \underbrace{\left(|\hat{u}(k)|^2(1+k^2)^s\right)^{1/2}}_{f} \underbrace{\left((1+k^2)^{-s}\right)^{1/2}}_{g} dk && \text{(33.102c)}\\
&= \langle f, g\rangle_{L^2(\mathbb{R}^d)}. && \text{(33.102d)}
\end{aligned}$$

Ici il convient nous arrêter un instant pour nous convaincre que f et g sont réellement des éléments de L^2. En ce qui concerne f, c'est facile : $\hat{u}$ est une fonction Schwartz. En ce qui concerne g il faut l'intégrabilité de $|g|^2$, c'est à dire de $k \mapsto (1+k^2)^{-s}$. Cela a lieu si et seulement si $2s > n$

et donc a lieu dans les hypothèses du théorème. Nous utilisons le théorème de Cauchy-Schwarz [5] pour continuer :

$$(2\pi)^d|u(x)| \leqslant \|f\|_{L^2}\|g\|_{L^2} \tag{33.103a}$$
$$= c\left(\int_{\mathbb{R}^d} |\hat{u}(k)|^2(1+k^2)^s dk\right)^{1/2} \tag{33.103b}$$
$$= c\|u\|_{H^s(\mathbb{R}^d)}. \tag{33.103c}$$

Donc en introduisant le facteur $(2\pi)^d$ dans la constante c nous avons

$$\|u\|_\infty \leqslant c\|u\|_{H^s(\mathbb{R}^d)}. \tag{33.104}$$

Cela est tout ce que nous voulions faire avec $u \in \mathscr{S}(\mathbb{R}^d)$.

Nous considérons maintenant $u \in H^s(\mathbb{R}^d)$. Vu que la densité des fonctions Schwartz dans H^s (proposition 33.15) nous considérons une suite (u_j) dans $\mathscr{S}(\mathbb{R}^d)$ telle que

$$u_j \xrightarrow{H^s(\mathbb{R}^d)} u \tag{33.105}$$

Ici u est une classe, mais nous identifions u_j avec sa classe (parce qu'il ne faut pas exagérer non plus). La suite (u_j) est de Cauchy dans H^s, donc si $\epsilon > 0$ est donné, il existe N tel que si $n, m > N$, $\|u_m - u_n\| \leqslant \epsilon$. Nous avons alors aussi

$$\|u_m - u_n\|_\infty \leqslant c\epsilon, \tag{33.106}$$

ce qui signifie que (u_j) est également une suite de Cauchy dans $\big(C^0(\mathbb{R}^d), \|.\|_\infty\big)$ qui est un espace complet par la proposition 14.316.

Il existe donc une fonction $v \in C_0^0(\mathbb{R}^d)$ telle que

$$u_j \xrightarrow{\left(C_0^0(\mathbb{R}^d),\|.\|_\infty\right)} v. \tag{33.107}$$

La question est de savoir si nous pouvons déduire que v est un représentant de u.

Par le lemme 33.21 nous avons également la convergence

$$u_j \xrightarrow{L^2(\mathbb{R}^d)} u. \tag{33.108}$$

Pour tout $\varphi \in \mathscr{S}(\mathbb{R}^d)$ nous avons alors

$$\langle u_j, \varphi\rangle_{L^2} \to \langle u, \varphi\rangle_{L^2}. \tag{33.109}$$

Mais en même temps, la convergence (33.107) couplée au lemme 33.22 donne également

$$\langle u_j, \varphi\rangle_{L^2} \to \langle v, \varphi\rangle_{L^2}. \tag{33.110}$$

Par unicité de la limite (dans $\mathbb{R}$) nous avons

$$\langle v, \varphi\rangle_{L^2} = \langle u, \varphi\rangle_{L^2} \tag{33.111}$$

pour tout $\varphi \in \mathscr{S}(\mathbb{R}^d)$. La proposition 28.1 appliquée à $u - v$ montre alors que $u - v = 0$ presque partout, c'est à dire que v est bien un représentant de u.

Le représentant v de u est non seulement continu (comme limite uniforme de fonctions continues), mais également bornée, comme limite uniforme de fonctions Schwartz. □

Proposition 33.24 ([381]).
Si $u \in H^s(\mathbb{R}^d)$ ($s \in \mathbb{R}$) alors

5. Formule 11.4.

(1) $\partial^\alpha u \in H^{s-|\alpha|}(\mathbb{R}^d)$,

(2) *l'application*

$$\partial^\alpha : H^s(\mathbb{R}^d) \to H^{s-|\alpha|}(\mathbb{R}^d) \tag{33.112}$$

est continue.

Note : ici ∂ est l'opération de dérivée faible.

Démonstration. Nous allons seulement prouver que $\partial_j : H^s(\mathbb{R}^d) \to H^{s-1}(\mathbb{R}^d)$ est bien définie[6] et continue. Par composition, la thèse suivra.

Soit $u \in H^s(\mathbb{R}^d)$ par le lemme 33.10 nous avons

$$\widehat{\partial_j u} = i\xi_j \hat{u}. \tag{33.113}$$

D'autre part, la fonction

$$\begin{aligned} f : \mathbb{R}^n &\to \mathbb{R} \\ x &\mapsto \frac{x_i}{1 + \|x\|^2} \end{aligned} \tag{33.114}$$

est bornée (et même indépendamment de i) par une constante K. Donc nous avons pour tout[7] s :

$$k_i(1 + \|k\|^2)^{-s} < K(1 + \|k\|^2)^{-s+1}. \tag{33.115}$$

Avec cela nous pouvons calculer un peu : si $u \in H^s(\mathbb{R}^d)$, nous avons

$$\|\partial_j u\|_{H^{s-1}(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |\widehat{\partial_j u}|(1 + k^2)^{s-1} dk \tag{33.116a}$$

$$= \int_{\mathbb{R}^d} k_j |\hat{u}|(1 + k^2)^{s-1} dk \tag{33.116b}$$

$$\leqslant \int_{\mathbb{R}^d} K|\hat{u}|(1 + k^2)^{s} dk \tag{33.116c}$$

$$= K\|u\|_{H^s(\mathbb{R}^d)}. \tag{33.116d}$$

Nous avons donc que $\|\partial_j u\|_{H^{s-1}(\mathbb{R}^d)}$ est fini lorsque $u \in H^s(\mathbb{R}^d)$.

La majoration $\|\partial_j u\| \leqslant K\|u\|$ donne la majoration suivante pour la norme de l'opérateur ∂_j :

$$\|\partial_j\| = \sup_{\|u\|_{H^s}=1} \|\partial_j u\|_{H^{s-1}} \leqslant K. \tag{33.117}$$

Le fait d'être borné implique d'être continu par la proposition 12.21. □

Théorème 33.25 (Théorème de plongement de Sobolev [381])**.**
Soient $k \in \mathbb{N}$ et $m > \frac{d}{2} + k$. Alors

$$H^s(\mathbb{R}^d) \subset C_0^k(\mathbb{R}^d). \tag{33.118}$$

Remarques :

— L'espace $C_0^k(\mathbb{R}^d)$ est l'ensemble des fonctions de classe C^k qui s'annulent à l'infini.

— L'inclusion (33.118) signifie que tout élément dans H^s possède un représentant dans $C_0^k(\mathbb{R}^d)$.

Démonstration. Pour $k = 0$, c'est le théorème 33.23. Si $|\alpha| < k$ nous savons que $\partial^\alpha u \in H^{s-k} \subset C_0^0(\mathbb{R}^d)$. Cela signifie que les dérivées faibles sont continues, mais pas qu'il existe un représentant qui est réellement k fois continument dérivable.

Soit $u \in H^s(\mathbb{R}^d)$ et une suite (u_j) dans $\mathscr{S}(\mathbb{R}^d)$ telle que

$$u_j \xrightarrow{H^s(\mathbb{R}^d)} u. \tag{33.119}$$

6. Au sens où l'espace d'arrivée est bien celui-là.
7. Question : dans [381], il faut dépendre cette constante de s. Je ne comprends pas pourquoi.

Vu que l'espace topologique $\left(C_0^k(\mathbb{R}^d), \|.\|_\infty\right)$ est complet il existe $v \in C_0^k$ tel que

$$u_j \xrightarrow{C_0^k} v. \tag{33.120}$$

Il reste à montrer que v est un représentant de u. Cela se fait comme plus haut en montrant que $u_j \xrightarrow{L^2} u$. □

Chapitre 34

Équations différentielles ordinaires

Une équation différentielle ordinaire est la recherche de toutes les fonctions définie sur une partie de $\mathbb{R}$ satisfaisant à une certaine égalité, faisant intervenir les dérivées de la fonction recherchée.

Dans la suite, I désignera un intervalle de $\mathbb{R}$. Une fonction sera **dérivable sur** I si elle est dérivable au sens usuel sur l'intérieur de I, et si elle est dérivable à droite (resp. à gauche) sur l'éventuel bord gauche (resp. droit) de I.

Définition 34.1.
Une ***équation différentielle ordinaire d'ordre*** n ***sur*** I *est la recherche d'une fonction* $y : I \to \mathbb{R}$ *dérivable* n *fois, satisfaisant à une équation du type*

$$F(t, y(t), y'(t), \dots, y^{n\prime}(t)) = 0 \quad \textit{pour tout } t \in I \tag{34.1}$$

où I *est un intervalle de* $\mathbb{R}$ *et* $F : (I \times D) \subset (\mathbb{R} \times \mathbb{R}^{n+1}) \to \mathbb{R}$ *est une fonction donnée.*

Remarque 34.2.
L'équation différentielle (34.1) sera raccourcie sous la forme

$$F(t, y, y', \dots, y^{n\prime}) = 0 \tag{34.2}$$

où la dépendance en t est sous-entendue.

Exemple 34.3
Soit $f : I \to \mathbb{R}$ une fonction continue fixée. L'équation différentielle

$$y' = f(t) \tag{34.3}$$

se ramène à la recherche des primitives de f sur l'intervalle I. △

Le lemme suivant sert de temps en temps.

Lemme 34.4 (Lemme de Grönwall)**.**
Soient ϕ *et* ψ *deux fonctions telles que pour tout* $t \in [t_0, t_1]$, $\phi(t) \geqslant 0$, $\psi(t) \geqslant 0$ *et*

$$\phi(t) \leqslant K + L \int_{t_0}^{t} \psi(s)\phi(s)ds \tag{34.4}$$

où K *et* L *sont des constantes positives. Alors*

$$\phi(t) \leqslant K \exp\big(L \int_{t_0}^{t} \psi\big). \tag{34.5}$$

Lemme 34.5 (Lemme de Grönwall[382])**.**
Si $u, a, b \in C^0\big([0, T], \mathbb{R}^+\big)$ *sont telles que*

$$u(t) \leqslant b(t) + \int_0^t a(s)u(s)ds \tag{34.6}$$

pour tout $t \in [0,T]$ alors pour tout $t \in [0,T]$ nous avons aussi

$$u(t) \leqslant b(t) + \int_0^t b(s)a(s)e^{\int_s^t a(u)du}ds. \tag{34.7}$$

34.1 Équation homogène, solution particulière

Voici un petit morceau d'algèbre linéaire. Soient des espaces vectoriels V et W ainsi qu'une application linéaire $D\colon V \to W$. Nous voulons résoudre $D(u) = v$, c'est à dire déterminer l'ensemble

$$D^{-1}(v) = \{u \in V \text{ tel que } Du = v\}. \tag{34.8}$$

Lemme 34.6.
Soient des espaces vectoriels V et W ainsi qu'une application linéaire $D\colon V \to W$. Si $u_P \in V$ satisfait à $Du_P = v$ alors

$$D^{-1}(v) = \ker(D) + u_P. \tag{34.9}$$

Démonstration. Si $u \in \ker(D) + u_p$ alors $u = k + u_p$ avec $Dk = 0$, ce qui donne tout de suite $Du = Dk + Du_P = v$. Donc $u \in D^{-1}(v)$.

Dans l'autre sens, si $u \in D^{-1}(v)$ alors nous pouvons écrire $u = (u - u_P) + u_P$. Vu que $u - u_P \in \ker(D)$ nous avons bien $u \in \ker(D) + u_P$. □

Ce petit lemme explique pourquoi la résolution d'équation différentielles passe par le principe « générale de l'homogène plus particulière de la non-homogène ». Cela marche autant pour les équations différentielles ordinaires que pour celles aux dérivées partielles.

Exemple 34.7
Considérons l'équation différentielle ordinaire

$$y' - y = 4. \tag{34.10}$$

L'opérateur dont nous parlons est par exemple

$$D\colon C^\infty(\mathbb{R}) \to C^\infty(\mathbb{R}) y \mapsto y' - y. \tag{34.11}$$

Nous devons résoudre $Dy = 4$ où « 4 » est l'élément fonction constante égale à 4 dans $C^\infty(\mathbb{R})$. L'ensemble $\ker(D)$ sont les éléments $y \in C^\infty(\mathbb{R})$ tels que $y' = y$:

$$\ker(D) = \{t \mapsto Ke^t \text{ tel que } K \in \mathbb{R}\}. \tag{34.12}$$

Nous devons trouver un élément quelconque y_P de $D^{-1}(4)$. Facile : $y_P(t) = -4$.

Au final,

$$D^{-1}(4) = \{t \mapsto Ke^t - 4 \text{ tel que } K \in \mathbb{R}\}. \tag{34.13}$$

△

Dans cet exemple nous avons pris $V = W = C^\infty(\mathbb{R})$. Mais souvent nous sommes amenés à considérer des espaces plus subtils, parce qu'il existe simplement pas de solutions dans C^∞, ou alors parce que beaucoup de solutions n'y sont pas.

34.2 Que faire avec $f(z)dz = g(t)dt$?

Dans de nombreux exercices d'équations différentielles, nous tombons sur $u' = f(t)$, et nous faisons formellement

$$\frac{du}{dt} = f(t) \quad \Rightarrow \quad du = f(t)dt, \tag{34.14}$$

et ensuite, il y a la formule un peu magique

$$u - u_0 = \int_{t_0}^{t} f(t)dt. \tag{34.15}$$

Voyons ce qu'il en est. Tout d'abord, il faut comprendre ce que signifie la formule

$$f(z)dz = g(t)dt. \tag{34.16}$$

Il s'agit d'une égalité entre deux formes différentielles sur $\mathbb{R}$ où z est une fonction de t. Étant donné que z est une fonction de t, il faut voir dz comme la différentielle de cette fonction. La différentielle d'une fonction à une variable est donné par la dérivée :

$$dz_t = z'(t)dt \tag{34.17}$$

Écrire l'équation (34.16) pour chaque t revient donc à écrire

$$f\big(z(t)\big)z'(t)dt = g(t)dt \tag{34.18}$$

Cela est une égalité entre deux formes différentielles. Nous avons donc égalité entre les intégrales des formes sur un chemin. Prenons un chemin tout simple de t_0 vers t :

$$\int_{t_0}^{t} f\big(z(t)\big)z'(t)dt = \int_{t_0}^{t} g(t)dt. \tag{34.19}$$

Dans le premier membre, nous faisons un changement de variable $\xi = z(t)$, $d\xi = z'(t)dt$, et nous obtenons

$$\int_{z_0}^{z(t)} f(\xi)d\xi = \int_{t_0}^{t} g(t)dt. \tag{34.20}$$

où nous avons remplacé la constante $z(t_0)$ par z_0 dans la borne d'intégration. Si F est une primitive de f et G une primitive de g, nous avons

$$F(z) - F(z_0) = G(t) - G(t_0). \tag{34.21}$$

Si aucun problème de Cauchy n'est donné, les constantes $F(z_0)$ et $G(t_0)$ sont mises en une seule et nous écrivons la solution

$$F\big(z(t)\big) = G(t) + C, \tag{34.22}$$

qui est une équation implicite pour $z(t)$.

Nous trouvons assez souvent le cas simple

$$f(z)dz = dt. \tag{34.23}$$

En remplaçant $g(t) = 1$ dans (34.20), nous trouvons la fameuse

$$t - t_0 = \int_{z_0}^{z} f(z)dz, \tag{34.24}$$

dans laquelle il y a un abus de notation terrible entre le z de la borne (que les étudiants oublient souvent) et la variable d'intégration z !!

Le passage de (34.23) à (34.24) sera très souvent utilisé dans le cours de mécanique par exemple.

34.3 Équations linéaires du premier ordre

Une **équation différentielle linéaire** est une équation de la forme

$$y' + u(t)y = v(t). \tag{34.25}$$

Exemple 34.8
Tant qu'il n'y a pas de second membre, c'est facile. Prenons l'exemple suivant :

$$y' + 2ty = 0. \tag{34.26}$$

Nous mettons tous les t d'un côté et tous les y et y' de l'autre :

$$\frac{y'}{y} = -2t, \tag{34.27}$$

et puis on intègre sans oublier la constante d'intégration :

$$\ln(y) = -t^2 + C, \tag{34.28}$$

et donc $y(t) = Ke^{-t^2}$. △

Exemple 34.9

Lorsqu'il y a un second membre, il y a une astuce. Prenons par exemple

$$y' + 2ty = 4t. \tag{34.29}$$

L'astuce est de commencer par résoudre l'équation sans le second membre (l'équation homogène associée). Nous notons y_H la solution. Ici, la réponse est

$$y_H(t) = Ke^{-t^2}. \tag{34.30}$$

Ensuite le truc est d'essayer de trouver la solution de l'équation (34.29) sous la forme

$$y(t) = K(t)e^{t^2}. \tag{34.31}$$

L'idée est de prendre la même que la solution de l'équation homogène (sans second membre), mais en disant que K est une fonction. Afin de trouver la fonction K qui donne la solution, il suffit de remettre l'essai (34.31) dans l'équation (34.29) :

$$\underbrace{K'e^{-t^2} - 2tKe^{-t^2}}_{y'(t)} + \underbrace{2tKe^{-t^2}}_{2ty(t)} = 4t \tag{34.32}$$

Les deux termes avec K se simplifient et il reste

$$K'(t) = 4te^{t^2}, \tag{34.33}$$

ce qui signifie $K(t) = 2e^{t^2+C}$. Nous avons donc déterminé la fonction qui fait fonctionner l'essai, et la solution à l'équation est

$$y(t) = \big(2e^{t^2} + C\big)e^{-t^2} = 2 + Ce^{-t^2}. \tag{34.34}$$

△

La technique pour résoudre cette équation est de commencer par résoudre l'équation homogène associée. Si $U(t)$ est une primitive de $u(t)$, nous avons

$$\begin{aligned} y'_H(t) + u(t)y_H(t) &= 0 \\ \frac{y'_H}{y_H} &= -u(t) \\ \ln(y_H) &= -U(t) + C \\ y_H(t) &= e^{-U(t)+C} = Ke^{-U(t)} \end{aligned} \tag{34.35}$$

où $K = e^C$.

Cela fournit la solution générale de l'équation homogène. Il existe un truc génial qui permet d'en tirer la solution générale du système non homogène. Lorsque nous avons trouvé $y_H(t) = Ke^{-U(t)}$, le symbole K désigne une constante. La méthode de **variation des constantes** consiste à essayer la solution

$$y(t) = K(t)e^{-U(t)}, \tag{34.36}$$

c'est à dire à dire que la constante est en réalité une fonction. Afin de trouver quelle fonction $K(t)$ fait en sorte que l'essai (34.36) soit une solution, nous la remplaçons dans l'équation de départ $y' + uy = v$. Maintenant,

$$y'(t) = K'(t)e^{-U(t)} - K(t)u(t)e^{-U(t)}. \tag{34.37}$$

En remettant dans l'équation,

$$y' + uy = K'e^{-U} - Kue^{-U} + uKe^{-U} = K'e^{-U} = v. \tag{34.38}$$

Notez que les termes en K se sont miraculeusement simplifiés. Cela est directement dû au fait que e^{-U} est solution de l'équation homogène. Nous restons avec l'équation

$$K' = \frac{v}{e^{-U}} \tag{34.39}$$

pour $K(t)$. La solution générale du problème non homogène est donc finalement donnée par

$$y(t) = \big(W(t) + C\big)e^{-U(t)} \tag{34.40}$$

si $W(t)$ est une primitive de $v(t)e^{U(t)}$.

Tout ceci est un peu heuristique. La proposition suivante dit dans quels cas ça fonctionne.

Proposition 34.10.
Soient u et v continues sur I et U, une primitive de u sur I et W une primitive de ve^{-U} sur I. Une fonction $y\colon I \to \mathbb{R}$ est solution de $y' + u(t)y = v(t)$ si et seulement s'il existe une constante $C \in \mathbb{R}$ telle que

$$y(t) = \big(W(t) + C\big)e^{U(t)} \tag{34.41}$$

pour tout $t \in I$.

34.3.1 Pourquoi la variation des constantes fonctionne toujours ?

Prenons une équation non homogène

$$z'(t) = f(t)z(t) + g(t), \tag{34.42}$$

et supposons avoir une solution de l'homogène associée sous la forme $z_H(t) = Ch(t)$. Le coup de la variation des constates consiste à essayer une solution pour l'équation non homogène sous la forme [1]

$$z(t) = K(t)h(t). \tag{34.43}$$

1. Je ne sais plus qui a eu l'idée de changer le nom de la constante de C vers K au moment de la transformer en fonction, mais c'est une bonne idée.

Nous injectons cette solution dans l'équation de départ en utilisant le fait que $z'(t) = K'(t)h(t) + K(t)h'(t)$:

$$K'(t)h(t) + K(t)h'(t) = f(t)K(t)h(t) + g(t). \tag{34.44}$$

Le terme $K(t)h'(t)$ se récrit en utilisant la propriété de définition de h, c'est à dire que $h'(t) = f(t)h(t)$. Nous voyons que les termes ne contenant pas de K' se simplifient ; il reste

$$K'h = g. \tag{34.45}$$

Cette équation a comme solution

$$K = \int \frac{f}{h} + C. \tag{34.46}$$

J'insiste sur la constante d'intégration ! En réalité, celles et ceux qui auront compris l'équation (34.24) sauront que K est donné par

$$K(t) = \int_{\xi_0}^{t} \frac{f(\xi)}{g(\xi)} d\xi \tag{34.47}$$

où ξ_0 joue le rôle de la constante d'intégration.

Quoi qu'il en soit, la solution générale de l'équation non homogène est

$$z(t) = K(t)h(t) = \left(\int \frac{g}{h} + C \right) h. \tag{34.48}$$

Cette solution comprend deux termes : Ch qui est solution de l'homogène, et $\left(\int \frac{g}{h}\right) h$ qui est une particulière de l'équation non homogène.

Quelques conclusions :

(1) Si vous avez encore du K (et pas que du K') dans votre équation qui donne K, c'est que vous n'être pas dans le cadre d'une équation de type (34.42). Le plus souvent, c'est que vous avez fait une faute de calcul quelque part.

(2) La méthode des variations des constantes n'est pas en contradiction avec le principe de « SGEH+SPENH ». En effet, la SGEP et la SPENH sont toutes deux dans la solution (34.48).

(3) La variation des constantes peut être vue comme une façon cool de trouver une solution particulière de l'équation non homogène.

(4) La simplification ne se fait que après avoir remplacé Kh' par Kfh, c'est à dire après avoir utilisé le fait que z_H est solution de l'homogène. Sinon, la simplification n'est pas du tout évidente a priori. Il se peut même que, visuellement, les termes Kh' et Kfh ne se ressemblent pas du tout. Un exemple de cela arrivera par exemple dans l'exemple 34.13, pour arriver à l'équation (34.64).

34.4 Équations à variables séparées

Une **équation à variables séparées** est une équation de la forme

$$y' = u(t)f(y) \tag{34.49}$$

où $u\colon I \to \mathbb{R}$ et $f\colon J \to \mathbb{R}$ sont deux fonctions continues données. Les propositions 34.11 et 34.12 résolvent ce cas, mais avant de voir cela, nous allons donner quelques indications « pratiques ».

34.4.1 La méthode rapide

On peut évidemment mettre tous les y et y' d'un côté :

$$\frac{y'}{f(y)} = u(x). \tag{34.50}$$

Une fois que cela est fait, on écrit $y' = \frac{dy}{dx}$, et on envoie le dx du côté des x :

$$\frac{dy}{f(y)} = u(x)dx. \tag{34.51}$$

Maintenant il suffit de prendre l'intégrale des deux côtés : comme la position des dx et dy l'indiquent, il faut intégrer par rapport à y d'un côté et par rapport à dx de l'autre côté.

L'intégrale à gauche est facile : c'est $\ln(y)$. À droite, par contre, ça dépend tout à fait de u.

34.4.2 La méthode plus propre

$$y'(t) = u(t)f\big(y(t)\big). \tag{34.52}$$

Nous considérons U, une primitive de u sut I et G, une primitive de $1/f$ sur J. Si $I' \subseteq I$ et $y\colon I' \to J$, alors y est solution de (34.49) si et seulement s'il existe une constante C telle que

$$G\big(y(t)\big) = U(t) + C. \tag{34.53}$$

La recherche des solutions de l'équation différentielle se ramène donc à la recherche de primitives et de solutions d'une équation algébrique (il faut isoler $y(t)$ dans (34.53)). Réciproquement toute solution régulière de cette dernière relation est solution de l'équation différentielle.

Remarque : lorsque nous cherchons U et G, nous ne cherchons que *une* primitive. Il ne faut pas considérer des constantes d'intégration à ce niveau.

34.4.3 Les théorèmes

Proposition 34.11.
Nous considérons l'équation (34.49) avec $u(t)$ continue sur I et f continue sur J avec $f(\eta) \neq 0$ pour tout $\eta \in J$. Soit U, une primitive de u sur I, et G, une primitive de $1/f$ sur J.

Si $y\colon Y' \to J$ est une fonction sur un intervalle $I' \subset I$, alors y est solution de l'équation (34.49) si et seulement s'il existe $C \in \mathbb{R}$ tel que

$$G\big(y(t)\big) = U(t) + C. \tag{34.54}$$

Cette proposition dit que toutes les solutions qui ne s'annulent jamais sur un intervalle ont la forme $G\big(y(t)\big) = U(t) + C$ et peuvent donc être trouvées en calculant des primitives.

La formule (34.54) peut être obtenue de la façon heuristique suivante, en écrivant $y' = dy/dt$, et en passant le dt à droite. Nous trouvons successivement

$$\begin{aligned} y' &= u(t)f(y) \\ dy &= u(t)f(y)dt \\ \frac{dy}{f(y)} &= u(t)dt \\ \int \frac{dy}{f(y)} &= \int u(t)dt \\ G(y) &= U(t) + C. \end{aligned} \tag{34.55}$$

Proposition 34.12.
Soient u continue sur I et f continue sur J, et $f(\eta) \neq 0$ sur J. Soient $t_0 \in I$ et $y_0 \in J$. Alors il existe $I' \subset I$ avec $t_0 \in I'$ et $f \in C^1(I' \to J)$ tels que

(1) y est solution de (34.49) sur I' et vérifie $y(t_0) = y_0$,

(2) si z est une solution de (34.49) sur $I'' \subset I'$ avec $t_0 \in I''$ et $z(t_0) = y_0$, alors $I'' \subset I'$ et $z(t) = y(t)$ pour tout $t \in I''$.

Exemple 34.13

Résoudre l'équation différentielle

$$y - \cos(t)y' = \cos(t)\big(1 - \sin(t)\big)y^2. \tag{34.56}$$

La fonction $y = 0$ est solution. En posant $z = 1/y$, nous trouvons l'équation

$$z + \cos(t)z' = \cos(t)\big(1 - \sin(t)\big) \tag{34.57}$$

à laquelle z doit satisfaire. L'équation homogène est

$$z'_H = -\frac{z_H}{\cos(t)}. \tag{34.58}$$

Ceci est une équation à variables séparées que nous résolvons en suivant les méthodes données plus haut : nous posons

$$\begin{aligned} u(t) &= \frac{1}{\cos(t)}, \\ f(z) &= -z, \\ U(t) &= \ln\left[\tan\left(\frac{\pi}{4} + \frac{t}{2}\right)\right] \quad \text{(voir formulaire)}, \\ G(z) &= \ln\left(\frac{1}{z}\right). \end{aligned} \tag{34.59}$$

La solution z_H est donnée par l'équation

$$\ln\left(\frac{1}{z}\right) = \ln\left[K\tan\left(\frac{\pi}{4} + \frac{t}{2}\right)\right], \tag{34.60}$$

c'est à dire

$$z_H(t) = \frac{K}{\tan\left(\frac{\pi}{4} + \frac{t}{2}\right)}. \tag{34.61}$$

Nous appliquons maintenant la méthode de variation des constantes sur cette solution afin de trouver la solution générale de l'équation (34.57). En utilisant la règle de Leibnitz, $z' = K'z_H + Kz'_H$, nous trouvons

$$\frac{K}{\tan\left(\frac{\pi}{4} + \frac{t}{2}\right)} + \cos(t)\left(\frac{K'}{\tan\left(\frac{\pi}{4} + \frac{t}{2}\right)} - \frac{K}{2\sin^2\left(\frac{\pi}{4} + \frac{t}{2}\right)}\right) = \cos(t)\big(1 - \sin(t)\big). \tag{34.62}$$

Malgré leurs apparences, les deux termes en K se simplifient. En effet, en vertu de l'équation $z'_H = \frac{-z_H}{\cos(t)}$, nous avons

$$\frac{-K}{2\sin^2\left(\frac{\pi}{4} + \frac{t}{2}\right)} = \frac{-K}{\cos(t)\tan\left(\frac{\pi}{4} + \frac{t}{2}\right)}. \tag{34.63}$$

Le travail de voir quel est le lien entre $\sin^2\left(\frac{\pi}{4} + \frac{t}{2}\right)$, $\tan\left(\frac{\pi}{4} + \frac{t}{2}\right)$ et $\cos(t)$ est en réalité fait dans votre formulaire au moment où vous l'avez utilisé pour intégrer u pour obtenir le $U(t)$ de (34.59).

Après cette simplification durement méritée, nous trouvons l'équation suivante pour $K(t)$:

$$\frac{K'}{\tan\left(\frac{\pi}{4} + \frac{t}{2}\right)} = 1 - \sin(t). \tag{34.64}$$

Résoudre cela revient à trouver la primitive de

$$\big(1 - \sin(t)\big)\tan\left(\frac{\pi}{4} + \frac{t}{2}\right), \tag{34.65}$$

ce qui est relativement compliqué. La réponse est

$$\begin{aligned} K(t) = \ln\left(\sin\left(\frac{2\,x+\pi}{4}\right)+1\right) + \ln\left(\sin\left(\frac{2\,x+\pi}{4}\right)-1\right) \\ + 2\,\ln\sec\left(\frac{2\,x+\pi}{4}\right) + 2\,\sin^2\left(\frac{2\,x+\pi}{4}\right) \end{aligned} \tag{34.66}$$

Nous pouvons un peu simplifier en utilisant le fait que $\ln(a+b)+\ln(a-b)=\ln(a^2-b^2)$:

$$K(t) = \ln\left(-\cos^2\left(\frac{2\,x+\pi}{4}\right)\right) + 2\,\ln\sec\left(\frac{2\,x+\pi}{4}\right) + 2\,\sin^2\left(\frac{2\,x+\pi}{4}\right). \tag{34.67}$$

Il me semble toutefois qu'il faudrait prendre des valeurs absolues pour les logarithmes.

△

34.5 Équations linéaires d'ordre supérieur

34.5.1 Équations et systèmes linéaire à coefficients constants

Nous regardons l'équation

$$y^{(n)} + a_1 y^{(n-1)} + \cdots + a_{n-1}y' + a_n y = v(t) \tag{34.68}$$

où les coefficients a_k sont maintenant des constantes. Il faut commencer par résoudre le polynôme caractéristique

$$r^n + a_1 r^{n-1} + \cdots + a_n = 0. \tag{34.69}$$

Si $\lambda_1,\ldots,\lambda_k$ sont les solutions avec multiplicité $\mu_1,\ldots,\mu_k$, alors le **système fondamental** de solutions linéairement indépendantes est l'ensemble suivant de solutions à l'équation homogène :

$$\begin{aligned} &e^{\lambda_1 t}, te^{\lambda_1 t},\ldots,t^{\mu_1-1}e^{\lambda_1 t} \\ &\vdots \\ &e^{\lambda_k t}, te^{\lambda_k t},\ldots,t^{\mu_k-1}e^{\lambda_k t}. \end{aligned} \tag{34.70}$$

Nous notons y_i ces solutions. La solution générale de l'équation homogène est donc donnée par

$$y_H = \sum_i c_i y_i. \tag{34.71}$$

Afin de trouver la solution générale de l'équation non homogène, nous appliquons la méthode de variation des constantes, en imposant les $n-1$ conditions

$$\sum_{i=1}^{n} c_i'(t) y_i^{(l)}(t) = 0 \tag{34.72}$$

avec $l=0,\ldots,n-2$. Ces condition plus l'équation de départ (34.68) forment un système de n équations différentielles pour les n fonctions inconnues $c_i(t)$.

Cette condition peut paraître mystérieuse. Il est cependant encore possible de travailler sans poser la condition (34.72) en suivant la recette, en calculant des déterminants de Wronskien. Des exemples sont donnés dans les exercices sur le second ordre.

34.5.2 Si les coefficients ne sont pas constants ?

Une équation différentielle linéaire d'ordre n sur I est une équation de la forme

$$y^{(n)} + u_1(t)y^{(n-1)} + \cdots + u_{n-1}(t)y' + u_n(t)y = v(t) \tag{34.73}$$

où v et u_k sont des fonctions continues fixées de I vers $\mathbb{R}$.

Pour résoudre cette équation, il faut commencer par résoudre l'équation homogène correspondante (c'est à dire celle que l'on obtient en posant $v(t) = 0$). Ensuite, nous trouvons la solution de l'équation (34.73) en appliquant la méthode de la **variation des constantes**.

Donnons un exemple du pourquoi la méthode de variations des constantes est efficace. Soit l'équation

$$u' + f(t)u = g(t), \tag{34.74}$$

et disons que u_H est une solution de l'équation homogène. La méthode de variations des constantes consiste à poser $u(t) = K(t)u_H(t)$, et donc $u'(t) = K'u_H + Ku'_H$. En remettant dans l'équation de départ,

$$K'u_H + Ku'_H + fKu_H = g. \tag{34.75}$$

La somme $Ku'_H + fKu_H$ est nulle, par définition de u_H. Par conséquent, il ne reste que

$$K' = \frac{g(t)}{u_H}. \tag{34.76}$$

Lorsqu'on utilise la méthode de variation des constantes, nous trouvons toujours une simplification « miraculeuse ».

Dans l'immédiat, nous ne considérons que le cas où les u_i sont des constantes. Le cas où les u_i deviennent des fonctions de t sera vu plus tard.

34.6 Système d'équations linéaires

34.6.1 La magie de l'exponentielle...

Prenons l'équation différentielle très simple

$$y' = ay. \tag{34.77}$$

La solution est $y(t) = Ae^{at}$. Et si on a la donnée que Cauchy $y(t_0) = y_0$, alors

$$y(t) = Ae^{at}e^{-at_0}e^{at_0} = e^{a(t-t_0)}y(t_0). \tag{34.78}$$

Donc on a le facteur multiplicatif $e^{a(t-t_0)}$ qui sert à faire passer de $y(0)$ à $y(t)$. C'est un peu un opérateur d'évolution. Ce qui fait la magie de l'exponentielle, c'est son développement en série

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots \tag{34.79}$$

qui est tel que chaque terme est la dérivée du terme suivant.

Maintenant, si on a un système

$$\bar{y}' = A\bar{y}, \tag{34.80}$$

il n'est pas du tout étonnant d'avoir comme solution $\bar{y}(t) = e^{At}$ où l'exponentielle de la matrice est définie exactement par la série (34.79). C'est un peu longuet, mais dans le cours, c'est effectivement ce qui est prouvé. La matrice résolvante $R(t, t_0)\colon \bar{y}_0 \to \bar{y}(t; t_0, y_0)$ est donné par

$$R(t, t_0) = e^{(t-t_0)A}, \tag{34.81}$$

exactement comme dans l'équation (34.78).

34.6.2 . . . mais la difficulté

Maintenant, il est suffisant de calculer des exponentielles de matrices pour résoudre des systèmes. Hélas, il est en général très difficile de calculer des exponentielles. Tu peux essayer de prouver les deux suivantes :

$$\begin{aligned} A = \begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix} &\rightsquigarrow e^A = \begin{pmatrix} \cos(a) & \sin(a) \\ -\sin(a) & \cos(a) \end{pmatrix} \\ S = \begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix} &\rightsquigarrow e^S = \begin{pmatrix} \cosh(a) & \sinh(a) \\ \sinh(a) & \cosh(a) \end{pmatrix}. \end{aligned} \tag{34.82}$$

La première, tu vas la revoir si tu fais de la géométrie différentielle ou de la mécanique quantique : l'algèbre de Lie du groupe des matrices orthogonales de déterminant 1 est l'algèbre des matrices antisymétriques.

La seconde se retrouve en relativité parce que e^S est la matrice qui préserve x^2-y^2, tout comme e^A préserve x^2+y^2. Quelques mots sur l'uutilisation des fonctions hyperboliques en relativité dans 43.4.8.1.

34.6.3 La recette

Afin d'éviter de devoir calculer explicitement des exponentielles de matrices, nous faisons appel à toutes sortes de trucs, dont la forme de Jordan. Le résultat final est la méthode suivante. Soit le système homogène

$$\bar{y}' = A\bar{y}. \tag{34.83}$$

1. D'abord, nous calculons les valeurs propres de A.
2. Ensuite les vecteurs propres.
3. Une bonne valeur propre, c'est une valeur propre dont l'espace propre a une dimension égale à sa multiplicité. C'est à dire que si λ est de multiplicité m, alors on a, dans les bons cas, m vecteur propres linéairement indépendants.

   Dans ce cas, si $v_1, \dots, v_m$ sont les vecteurs, alors on a les solutions linéairement indépendantes suivantes :

   $$\begin{pmatrix} \vdots \\ v_1 \\ \vdots \end{pmatrix} e^{\lambda t}, \dots, \begin{pmatrix} \vdots \\ v_m \\ \vdots \end{pmatrix} e^{\lambda t}. \tag{34.84}$$

   Pour chaque bonne valeur propre, ça nous fait un tel paquet de solutions linéairement indépendantes.
4. Si λ n'est pas une bonne valeur propre, alors les choses se compliquent. Mettons que λ ait k vecteurs propres en moins que sa multiplicité. Dans ce cas, il faut chercher des solutions sous la forme

   $$\begin{pmatrix} a_1^{(k)}t^k + \dots + a_1^{(0)} \\ \vdots \\ a_1^{(k)}t^k + \dots + a_n^{(0)} \end{pmatrix} e^{\lambda t}. \tag{34.85}$$

   C'est à dire qu'on prend comme coefficient de $e^{\lambda t}$, un vecteur de polynômes de degré k. Il faut mettre cela dans l'équation de départ pour voir quelles sont les contraintes sur les constantes $a_i^{(j)}$ introduites.
5. Nous avons un cas particulier du cas précédent. Si λ est une valeur propre de multiplicité m qui n'a que un seul vecteur propre v, alors il faut chercher des polynômes de degré $m-1$, et on peut directement fixer le coefficient de t^{m-1}, ce sera l'unique vecteur propre :

   $$\left[\begin{pmatrix} \vdots \\ v \\ \vdots \end{pmatrix} + \begin{pmatrix} a_1^{(m-2)} \\ \vdots \\ a_n^{(m-2)} \end{pmatrix} t^{m-2} + \dots \right] e^{\lambda t}. \tag{34.86}$$

Cela économise quelques calculs par rapport à poser brutalement (34.85).

34.6.4 Système d'équations linéaires avec matrice constante

Nous considérons l'équation différentielle

$$y'(t) = Ay(t) \tag{34.87}$$

pour la fonction $y\colon \mathbb{R} \to \mathbb{R}^n$ et A est une matrice ne dépendant pas de t. Nous supposons que A est diagonalisable pour les vecteurs propres v_i et les valeurs propres λ_i correspondantes.

La matrice

$$R(t) = \begin{bmatrix} e^{\lambda_1 t}v_1 & \dots & e^{\lambda_n t}v_n \end{bmatrix} \tag{34.88}$$

est la **matrice résolvante** du système. Alors la solution du système (34.87) pour la condition initiale $y(0) = y_0$ est

$$y(t) = R(t)y_0. \tag{34.89}$$

En effet

$$AR(t) = \begin{bmatrix} A\begin{pmatrix}\uparrow\\ e^{\lambda_1 t}v_1\\ \downarrow\end{pmatrix} & \dots & A\begin{pmatrix}\uparrow\\ e^{\lambda_n t}v_n\\ \downarrow\end{pmatrix} \end{bmatrix} = R'(t). \tag{34.90}$$

Par conséquent $y'(t) = R'(t)y_0 = AR(t)y_0 = Ay(t)$.

34.6.5 Système d'équations linéaires avec matrice non constante

Théorème 34.14 ([383]).
Soient I un intervalle de $\mathbb{R}$ et $M\colon \mathbb{R} \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ une fonction. Si les composantes M_{ij} sont des fonctions continues sur I alors :

(1) pour tout $t_0 \in I$ et pour tout $y_0 \in R^n$ le système

$$y'(t) = M(t)y(t) \tag{34.91}$$

admet une unique solution maximale définie sur I telle que $y(t_0) = y_0$;

(2) l'ensemble des solutions de l'équation (34.91) sur I est un espace vectoriel de dimension n.

34.7 Réduction de l'ordre

Afin de diminuer l'ordre d'une équation dans laquelle le paramètre n'apparaît pas, il y a deux changements de variables très utiles. Le premier, le plus simple, est simplement de poser $z(t) = y'(t)$, ce qui donne $z'(t) = y''(t)$. Le second, *qui n'est pas le même*, est $z\big(y(t)\big) = y'(t)$, qui entraîne $y''(t) = z'\big(y(t)\big)z(t)$. Dans ce second cas, il faut également changer de variable, et utiliser $y(t)$ comme variable au lieu de t.

Si ça ne marche pas, il faut suivre la procédure ci-après.

Nous supposons avoir une équation différentielle d'ordre p dans laquelle $y^{(p)}$ est isolée des autres dérivées :

$$y^{(p)}(t) = f\big(t, y(t), y'(t), \dots, y^{(p-1)}(t)\big) \tag{34.92}$$

où f est une fonction $f\colon \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}$, et la fonction cherchée est $y\colon \mathbb{R} \to \mathbb{R}$.

La méthode proposée ici consiste à transformer cette équation d'ordre p en un système d'équations d'ordre 1. Pour cela nous posons

$$\begin{aligned} F\colon \mathbb{R} \times \mathbb{R}^p &\to \mathbb{R}^p \\ (t, x) &\mapsto \begin{pmatrix} x_2 \\ \vdots \\ x_p \\ f(t, x_1, \dots, x_p) \end{pmatrix}. \end{aligned} \tag{34.93}$$

Nous considérons alors l'équation différentielle

$$Y'(t) = F\big(t, Y(t)\big). \tag{34.94}$$

pour $Y : \mathbb{R} \to \mathbb{R}^p$. La fonction $y = Y_1$ résout l'équation (34.92) si et seulement si la fonction Y résout l'équation (34.94).

De plus si l'équation (34.92) est donnée avec les conditions initiales $y^{(k)} = a_k$ $(k = 0, \dots, p-1)$ alors l'équation (34.94) vient avec les conditions initiales

$$Y(t_0) = \begin{pmatrix} a_0 \\ \vdots \\ a_{p-1} \end{pmatrix}, \tag{34.95}$$

c'est à dire $Y(t_0) = A_0$ avec $A_0 \in \mathbb{R}^p$.

Le théorème de Cauchy-Lipschitz 19.49 nous donne existence et unicité locale de la solution au système 34.94. Lorsque le système est linéaire, c'est à dire sous la forme $Y'(t) = M(t)Y(t)$, alors il y a mieux : le théorème 34.14.

Exemple 34.15

Nous considérons l'équation différentielle

$$\begin{cases} -u''(t) - u(t) = 1 & \text{(34.96a)} \\ u(0) = a_0 \in \mathbb{R} & \text{(34.96b)} \\ u'(0) = a_1 \in \mathbb{R}, & \text{(34.96c)} \end{cases}$$

et nous voulons montrer que ce système accepte une unique solution. Vu que l'équation différentielle se présente sous la forme $u'' = f(t, u)$ avec $f(t, u) = -1 - u$ nous posons

$$\begin{aligned} F : \mathbb{R} \times \mathbb{R}^2 &\to \mathbb{R}^2 \\ (t, x) &\mapsto \begin{pmatrix} x_2 \\ -1 - x_1 \end{pmatrix}, \end{aligned} \tag{34.97}$$

et nous considérons l'équation différentielle

$$Y' = F(t, Y) \tag{34.98}$$

pour la fonction $Y : \mathbb{R} \to \mathbb{R}^2$. La fonction F est Lipschitz (et même globalement) par rapport à Y. En effet,

$$\|F(x) - F(y)\| = (x_2 - y_2)^2 + (x_1 - y_1)^2 = \|x - y\|^2. \tag{34.99}$$

Le théorème de Cauchy-Lipschitz 19.49 s'applique et en posant $Y_0 = (a_0, a_1)$, il existe une unique solution à l'équation $Y' = F(t, Y)$ vérifiant $Y(0) = Y_0$. Nous notons $t \mapsto Y(t)$ cette solution.

En quoi cela nous aide ? Nous posons $u(t) = Y_1(t)$. Alors

$$u'(t) = Y_1'(t) = F_1'(t, Y) = Y_2(t). \tag{34.100}$$

En dérivant encore,

$$u''(t) = Y_2'(t) = F_2'(t, Y) = -1 - Y_1(t) = -1 - u(t), \tag{34.101}$$

ce qu'il fallait. La fonction $t \mapsto Y_1(t)$ est solution de notre équation de départ. Quid des conditions initiales ? Vu que $u = Y_1$ et $u' = Y_2$ nous avons

$$u(0) = Y_1(0) = a \tag{34.102}$$

et

$$u'(0) = Y_2(0) = b. \tag{34.103}$$

Toutes les prescriptions sont respectées.

Si vous voulez vraiment résoudre cette équation, il faudra plus de travail. D'abord résoudre l'équation homogène associée, c'est à dire l'équation caractéristique $r^2+1=0$, ce qui va donner

$$u_H(t) = Ae^{it} + Be^{-it}, \tag{34.104}$$

et ensuite faire le coup de la variation des constantes pour déterminer la solution générale du problème non homogène. △

Exemple 34.16
Nous reprenons l'équation différentielle

$$-u''(t) - u(t) = 1. \tag{34.105}$$

Nous avons déjà vu dans l'exemple 34.15 que cette équation avait une solution unique pour toute condition initiale. Cette fois nous voulons étudier les solutions lorsque nous imposons les conditions aux limites $u(0)=u(\pi)=0$. Nous allons voir qu'il n'y a pas de telles solutions.

Pour ce faire, soit une solution u. D'abord u'' existant, la fonction u est de classe au moins C^1. Mais $u''=-1-u$, donc u'' est également C^1, ce qui donne la régularité C^3 pour u. En continuant ainsi nous trouvons que u est de classe C^∞.

Le truc est de considérer a fonction $v(t)=\sin(t)$ qui vérifie l'équation différentielle

$$\begin{cases} -v''-v=0 & \text{(34.106a)}\\ v(0)=0 & \text{(34.106b)}\\ v'(0)=0. & \text{(34.106c)} \end{cases}$$

Nous calculons le produit scalaire sur $L^2\big(]0,\pi[\big)$ de (34.105) avec v :

$$\langle u'', v\rangle + \langle u, v\rangle = -\langle v, 1\rangle. \tag{34.107}$$

Le calcul de $\langle v,1\rangle$ est simplement l'intégrale de $\sin(t)$ pour t allant de 0 à π, c'est à dire $\langle v,1\rangle=2$. Vu que u et v sont toutes deux des fonctions qui s'annulent en 0 et en π nous pouvons faire des intégrations par partie les yeux fermés et exprimer $\langle u'', v\rangle$ sans dérivées sur u :

$$\langle u'', v\rangle = -\langle u', v'\rangle = \langle u, v''\rangle = -\langle u, v\rangle \tag{34.108}$$

où la dernière égalité n'est autre que le fait que $v=\sin$, donc $v''=-v$. Le membre de gauche de (34.107) vaut donc zéro alors que celui de droite vaut -2.

Nous concluons que le problème au limites posé n'admet pas de solutions. △

34.8 Autour de Cauchy-Lipschitz

Dans cette section nous étudions les équations différentielles du type

$$\begin{cases} y'(t) = f\big(t, y(t)\big) & \text{(34.109a)}\\ y(t_0) = y_0 & \text{(34.109b)} \end{cases}$$

34.8.1 Fuite des compacts et explosion en temps fini

Théorème 34.17 (Fuite des compacts[384, 242])**.**
Nous considérons l'équation différentielle

$$\begin{cases} y'(t) = f\big(t, y(t)\big) & \text{(34.110a)}\\ y(t_0) = y_0 & \text{(34.110b)} \end{cases}$$

où $f\colon I\times\Omega\to\mathbb{R}^n$ est continue et Ω ouvert dans $\mathbb{R}^n$. Soit la solution maximale $y_M\colon J_M=]t_{min},t_{max}[\to\Omega$. Si $t_{max}<\sup(I)$ alors $y_M(t)$ sort de tout compact de Ω lorsque $t\to t_{max}$.

Démonstration. Soit K un compact de Ω et nous considérons une suite (t_m) dans $]t_{min}, t_{max}[$ telle que $t_m \to t_{max}$. Si nous supposons que $y_M(t)$ ne sort pas de K alors nous avons $y_M(t_m) \in K$, c'est à dire une suite dans un compact. Quitte à passer à une sous-suite, nous supposons qu'elle est convergente. Soit $x_1 \in K$ la limite $\lim_{m\to\infty} y_M(t_m) = x_1$.

Vu que $t_{max} \in I$, la condition initiale $y(t_{max}) = x_1$ est valide et le théorème de Cauchy-Lipschitz 19.49 nous donne une unique solution maximale y_P définie sur un ouvert J_P autour de t_{max}.

Nous allons maintenant construire une solution au problème initial qui contredit la maximalité de y_M. Attention : il n'est pas évident a priori que $y_P(t) = y_M(t)$ sur l'intersection des domaines. Si c'était évident, la proposition serait démontrée.

Soit $\tilde{J} = J_M \cup J_P \cap]t_{min}, +\infty[$ et la fonction

$$\tilde{y}(t) = \begin{cases} y_M(t) & \text{si } t < t_{max} \\ y_P(t) & \text{si } t \geqslant t_{max}. \end{cases} \tag{34.111}$$

La fonction $\tilde{y}$ est continue par construction parce que

$$\lim_{t\to t_{max}} y_M(t) = x_1 = y_P(t_{max}). \tag{34.112}$$

Nous vérifions à présent que $\tilde{y}$ est une solution : $\tilde{y}'(t_{max}) = f\big(t_{max}, y(t_{max})\big)$:

$$\lim_{\epsilon\to 0} \frac{\tilde{y}(t_{max}-\epsilon) - \tilde{y}(t_{max})}{\epsilon} = \lim_{\epsilon\to 0} \frac{y_M(t_{max}-\epsilon) - y_P(t_{max})}{\epsilon} \tag{34.113a}$$

$$= \lim_{\epsilon\to 0} \frac{y_M(t_{max}-\epsilon) - y_P(t_{max}-\epsilon) + y_P(t_{max}-\epsilon) - y_P(t_{max})}{\epsilon} \tag{34.113b}$$

$$= \lim_{\epsilon\to 0} \frac{y_P(t_{max}-\epsilon) - y_P(t_{max})}{\epsilon} \tag{34.113c}$$

$$= y'_P(t_{max}). \tag{34.113d}$$

Donc $\tilde{y}$ est solution pour la condition initiale $\tilde{y}(t_{max}) = x_1$ et coïncide avec y_P en t_{max} et avec y_M avant t_{max}. Donc en réalité y_P, y_M et $\tilde{y}$ sont identiques et cela contredit la maximalité de y_M. □

Corollaire 34.18 (Explosion en temps fini)**.**
Soit (y_m, J) la solution maximale du problème de Cauchy (19.166) :

$$\begin{cases} y' = f(t,y) & \text{(34.114a)} \\ y(t_0) = y_0 & \text{(34.114b)} \end{cases}$$

avec $f\colon U = I \times \Omega \to \mathbb{R}^n$ où I est ouvert dans $\mathbb{R}$ et Ω ouvert dans $\mathbb{R}^n$. Nous supposons que f est continue sur U et localement Lipschitz par rapport à y.

Si la solution maximale est définie sur $J =]t_{min}, t_{max}[$ alors nous avons l'alternative suivante :

(1) Soit $t_{max} = \sup(I)$,

(2) soit $t_{max} < \sup(I)$ et $\lim_{t\to t_{max}} \|y(t)\| = \infty$.

Le résultat tient aussi mutatis mutandis *pour* $t_{\min}$.

Remarque 34.19.
Attention : ceci n'est pas une simple paraphrase de la fuite des compacts. L'information supplémentaire que ce corollaire donne est que la solution sort de tout compact *pour ne plus y retourner.*

Démonstration. L'hypothèse $t_{max} < \sup(I)$ signifie que la solution finit d'exister avant que les hypothèses sur f cessent d'être vraies. C'est à dire que la solution maximale est moindre que ce que nous aurions pu espérer.

Soit un compact K.Supposons que que pour tout $t_0 < t_{max}$ il existe $t \in]t, t_{max}[$ tel que $y_M(t) \in K$. Alors cela crée une suite t_k dans J telle que $y_M(t_k)$ est dans K. Comme dans le théorème de la fuite des compacts nous concluons l'impossibilité de la chose.

Donc pour tout compact K de Ω, il existe $T < t_{max}$ tel que $y_M(t) \in \Omega \backslash K$ pour tout $t \in [T, t_{max}[$. En prenant des boules fermées de plus en plus grandes en guise de compacts nous concluons que

$$\lim_{t \to t_{max}} \|y_M(t)\| = \infty. \tag{34.115}$$

□

34.20.
Notons que si $t_{max} < \infty$, si nous sommes dans l'alternative 34.18(2) et si la solution maximale y est de classe C^1 (ce qui est le cas lorsqu'on utilise Cauchy-Lipschitz 19.49) alors la dérivée de y est également non bornée dans un voisinage de t_{max}.

Mais si f est globalement bornée, alors dans l'équation $y' = f(t, y)$, la dérivée y' sera globalement bornée. Dans ce cas, la solution ne peut pas exploser en temps fini et existe donc globalement.

Problèmes et choses à faire

Êtes-vous d'accord avec 34.20 ?

Exemple 34.21
Soit l'équation différentielle

$$\begin{cases} y' = y(y-1)\sin(yt) & \text{(34.116a)} \\ y(0) = \frac{1}{2}. & \text{(34.116b)} \end{cases}$$

La fonction $f(t, y) = y(y-1)\sin(yt)$ ayant une dérivée bornée partout, est localement Lipschitz et le théorème de Cauchy-Lipschitz 19.49 s'applique. Pour toute condition initiale, une solution maximale unique existe.

Si nous oublions la condition initiale, il est facile de trouver des solutions constantes : $y' = 0$ avec $y(t) = k$ donne l'équation

$$0 = k(k-1)\sin(kt). \tag{34.117}$$

Les solutions $y_1(t) = 0$ et $y_2(t) = 1$ sont des solutions existant pour tout t.

Le graphe de la solution correspondante à la condition initiale $y(0) = \frac{1}{2}$ ne pouvant pas croiser les graphes de y_1 et y_2, elle est obligée d'exister pour tout t parce qu'elle ne peut pas exploser en temps fini. △

34.8.2 Écart entre deux conditions initiales

Proposition 34.22 ([382, 385]).
Soit une fonction $f : I \times \Omega \to \mathbb{R}^n$ continue et globalement Lipschitz en sa seconde variable (Ω est un ouvert de $\mathbb{R}^n$). Soient deux solutions $y_1 : I_1 \to \mathbb{R}^n$ et $y_2 : I_2 \to \mathbb{R}^n$ aux problèmes

$$\begin{cases} y_i'(t) = f\big(t, y_i(t)\big) & \text{(34.118a)} \\ y_i(t_0) = a_i & \text{(34.118b)} \end{cases}$$

Alors pour tout $t \in I_1 \cap I_2$ nous pouvons estimer l'écart entre y_1 et y_2 par la formule

$$\|y_1(t) - y_2(t)\| \leqslant e^{L|t-t_0|}\|a_1 - a_2\| \tag{34.119}$$

où L est la constante de Lipschitz de f.

Démonstration. Nous avons d'abord les majorations suivantes, qui semblent juste jouer avec les notations, mais qui utilisent le fait (contenu dans le théorème de Cauchy-Lipschitz) que y_i soit de

classe C^1 :

$$\|y_1(t)-y_2(t)\| = \|\int_{t_0}^{t} \big(y_1'(s)-y_1'(s)\big)\| \tag{34.120a}$$
$$\leqslant L\int_{t_0}^{t} \|f\big(s,y_1(s)\big)-f\big(s,y_2(s)\big)\|ds \tag{34.120b}$$
$$= L\int_{t_0}^{t} \|y_1(s)-y_2(s)\|ds \tag{34.120c}$$

C'est à ce moment que nous utilisons le lemme de Grönwall. Vu que

$$\|y_1(t)-y_2(t)\| \leqslant L\int_{t_0}^{t} \|y_1(s)-y_2(s)\|ds, \tag{34.121}$$

nous sommes dans les hypothèses de Grönwall 34.5 en posant

$$u(t) = \|y_1(t)-y_2(t)\| \tag{34.122a}$$
$$b(t) = \|y_1(0)-y_2(0)\| \tag{34.122b}$$
$$a(t) = L. \tag{34.122c}$$

Nous avons la majoration

$$\|y_1(t)-y_2(t)\| \leqslant \|y_1(0)-y_2(0)\| + L\int_0^t \|y_1(0)-y_2(0)\|e^{L(t-s)}ds. \tag{34.123}$$

Le calcul de l'intégrale intérieure donne

$$\int_0^t e^{L(t-s)}ds = -\frac{1}{L}(e^{-Lt}-1). \tag{34.124}$$

Avec ça, nous avons

$$\|y_1(t)-y_2(t)\| \leqslant e^{Lt}\|y_1(0)-y_2(0)\|. \tag{34.125}$$

□

34.23.
Notons que la proposition 34.22 est plutôt une mauvaise nouvelle parce que les solutions restent seulement linéairement proches l'une de l'autre lorsqu'on rapproche les conditions initiales, mais elle divergent exponentiellement vite avec le temps. Donc deux trajectoires arbitrairement proches au départ finissent assez vite par être bien séparées.

Cette proposition est cependant cruciale parce qu'elle explique que pour des petits t, les solutions ne s'écartent pas beaucoup, c'est à dire que pour t fixé, l'application qui à une donnée initiale fait correspondre la solution en t est continue. C'est le premier pas pour parler de régularité du flot.

34.8.3 Flot d'un champ de vecteurs

Nous reprenons l'équation différentielle du théorème de Cauchy-Lipschitz 19.49. En ce qui concerne les notations, I est un intervalle ouvert de $\mathbb{R}$ contenant 0 et l'application $f\colon I\times\mathbb{R}^n\to\mathbb{R}^n$ est continue et localement Lipschitz en sa seconde variable. Pour $a\in\mathbb{R}^n$, nous notons (J_a,y_a) la solution maximale (donc $y_a\colon J_a\to\mathbb{R}^n$) du problème

$$\begin{cases} y_a(t) = f\big(t,y_a(t)\big) & \text{(34.126a)}\\ y_a(0) = a. & \text{(34.126b)}\end{cases}$$

Nous noterons aussi de temps en temps $\varphi(t,a)=y_a(t)$.

Nous savons que $t \mapsto y_a(t)$ est de classe C^1, et cela est directement dans le théorème de Cauchy-Lipschitz. Une question d'une toute autre difficulté est la régularité de $a \mapsto y_a(t)$ pour t fixé, et encore pire : celle de $(t, a) \mapsto y_a(t)$.

Il se fait que l'application $(t, a) \mapsto y_a(t)$ a la même régularité que celle de f, mais cela va être un peu long à prouver. En ce qui concerne la régularité C^1, ce sera le théorème 34.28 dont la démonstration, comme vous pouvez le voir sera copieuse et demandera des propositions intermédiaires pas simples.

Définition 34.24.
Si t est fixé, l'application

$$\begin{aligned} \varphi_t \colon \mathbb{R}^n &\to \mathbb{R}^n \\ x &\mapsto \varphi(t, x) = y_x(t) \end{aligned} \tag{34.127}$$

*est le **flot** du problème de Cauchy (34.126).*

L'application $t \mapsto \varphi_t$ est ce qui est appelé le groupe à un paramètre de flot, pour des raisons qui arriverons plus tard[2].

Le but est d'étudier les propriétés du flot : est-il continu, une difféomorphisme, existe, pour quels t ? Où se cache le champ de vecteurs du titre dans l'équation différentielle ?

Nous posons

$$\mathcal{D} = \bigcup_{x \in \Omega} \big(J_x \times \{x\}\big). \tag{34.128}$$

Comme tout produit d'espaces métrique, l'ensemble $\mathcal{D}$ est muni d'une métrique via la définition 10.65.

Proposition 34.25 ([382])**.**
Soit un intervalle I ouvert de $\mathbb{R}$ contenant 0 et Ω un ouvert connexe de $\mathbb{R}^n$. Soit une application $f \colon I \times \mathbb{R}^n \to \mathbb{R}^n$ continue et localement Lipschitz en sa seconde variable. Pour $a \in \mathbb{R}^n$, nous notons (J_a, y_a) la solution maximale (donc $y_a \colon J_a \to \mathbb{R}^n$) du problème

$$\begin{cases} y_a(t) = f\big(t, y_a(t)\big) & \text{(34.129a)} \\ y_a(0) = a. & \text{(34.129b)} \end{cases}$$

Nous posons

$$\mathcal{D} = \bigcup_{x \in \Omega} \big(J_x \times \{x\}\big) \tag{34.130}$$

Nous définissons la fonction φ par $\varphi(t, x) = y_x(t)$ là où ça existe.

L'ensemble $\mathcal{D}$ est ouvert. L'application $\varphi \colon \mathcal{D} \to \Omega$ est localement Lipschitz.

Démonstration. Soit $(s, a) \in \mathcal{D}$ et (J_a, y_a) la solution maximale passant par a en $t = 0$. Par définition de $\mathcal{D}$ nous avons $s \in J_a$. Nous considérons J, un compact inclus dans J_a et contenant 0 et s en son intérieur. Nous posons

$$K = J \times y_a(J). \tag{34.131}$$

Vu que y_a est continue, cela est un compact. Chaque point de K possède un voisinage ouvert sur lequel f est Lipschitz[3] ; nous considérons un sous recouvrement fini et le maximum des constantes de Lipschitz. Cela nous crée un voisinage V de K dans $I \times \Omega$ dans lequel f est Lipschitz.

Vu que V est ouvert et K est compact avec $K \subset V$, nous pouvons trouver un ouvert V' et un compact K' tels que

$$K \subset V' \subset K' \subset V. \tag{34.132}$$

Sur ce V', la fonction f est de plus bornée parce que continue sur le compact K'. Nous renommons V' en V. Sur V nous avons :

— $\|f\|_{\infty, V} \leqslant M$,

2. ou pas...
3. Cela est à peu près la définition d'être localement Lipschitz : 14.253, voir aussi 14.254.

— f est Lipschitz en sa seconde variable, de constante de Lipschitz L.

En tant qu'espace produit, nous avons une distance sur $I \times \Omega$ donnée en 10.65 :

$$d\big((t,y),(t',y')\big) = \max\big\{|t-t'|, \|y-y'\|\big\}. \tag{34.133}$$

Nous posons

$$V_\epsilon(K) = \{z \in I \times \Omega \text{ tel que } d(z,K) < \epsilon\} \tag{34.134a}$$
$$W_\epsilon = \{(t,y) \in J \times \Omega \text{ tel que } \|y - y_a(t)\| < \epsilon\}. \tag{34.134b}$$

$\overline{W_\epsilon} \subset \overline{V_\epsilon(K)}$ Soit $(t,y) \in W_\epsilon$. Nous avons :

$$d\big((t,y),K\big) = \inf_{(t',y')\in K} d\big((t,y),(t',y')\big) \tag{34.135a}$$
$$= \inf_{(t',y')\in K} \max\{|t-t'|, \|y-y'\|\}. \tag{34.135b}$$

Mais demander $(t,y) \in W_\epsilon$ signifie que $t \in J$ et $\|y - y_a(t)\| \leqslant \epsilon$. Dans K nous avons l'élément $\big(t, y_a(t)\big)$ qui vérifie

$$d\big((t,y),(t,y_a(t))\big) = \|y - y_a(t)\| \leqslant \epsilon. \tag{34.136}$$

Donc l'infimum de (34.135b) est majoré par ϵ. Nous avons prouvé que $W_\epsilon \subset V_\epsilon(K)$ et donc même inclusion pour les fermetures.

Il existe $\epsilon > 0$ tel que $\overline{V_\epsilon(K)} \subset V$ Supposons que $\overline{V_\epsilon(K)}$ ne soit inclus dans V pour aucun ϵ. Alors nous considérons

$$z_n \in \overline{V_{1/n}(K)} \backslash V. \tag{34.137}$$

Nous avons par définition $d(z_n,K) \leqslant \frac{1}{n}$. Vu que K est compact, il comprend (au moins) un élément réalisant la distance : soit $z'_n \in K$ tel que

$$d(z_n, z'_n) = d(z_n, K). \tag{34.138}$$

Nous avons $d(z_n,z'_n) \to 0$, de telle sorte que les valeurs d'adhérence de (z_n) et (z'_n) sont les mêmes. Et comme (z'_n) est une suite dans un compact, elle a des valeurs d'adhérence. Soit z_∞ l'une d'elles. Vu que c'est une valeur d'adhérence d'une suite contenue dans le compact K, elle est également dans K : $z_\infty \in K$. Mais en même temps, z_n est hors de l'ouvert V, et donc dans le fermé V^c. Les valeurs d'adhérences restent dans le fermé, c'est à dire $z_\infty \notin V$. Vu que $K \subset V$, il y a contradiction.

Donc il existe $\epsilon > 0$ tel que $\overline{V_\epsilon(K)} \subset V$.

Il existe ϵ tel que $\overline{W_\epsilon} \subset V$ Il suffit de prendre le ϵ dont nous venons de parler pour avoir

$$\overline{\epsilon} \subset \overline{V_\epsilon(K)} \subset V. \tag{34.139}$$

Soit le ϵ en question, et $T > 0$ tel que $J \subset [-T,T]$. Nous posons $r = \epsilon e^{-LT}$. Soit $b \in \overline{B(a,r)}$ et

$$X = \{\tau \in J_+ \text{ tel que }]0,\tau] \subset J_b \text{ et } \big(t, y_b(t)\big) \in \overline{W_\epsilon}\, \forall\, t \in [0,\tau]\}. \tag{34.140}$$

Nous allons prouver que $X = J_+$ en prouvant qu'il est ouvert, fermé et non vide dans $J_+ = J \cap]0,\infty[$. Nous parlons bien de la topologie de J_+, celle induite[4] de $\mathbb{R}$. Vu que $0 \in J$, l'ensemble J_+ est ouvert à gauche, mais comme il est compact, il ne va certainement pas jusqu'à $+\infty$, de telle sorte qu'il est fermé à gauche. Les ouverts de J_+ sont les ensembles de la forme $\mathcal{O} \cap J_+$ où $\mathcal{O}$ est ouvert de $\mathbb{R}$. Il y en a de la forme $]0,m]$.

X est fermé C'est parce que $\overline{W_\epsilon}$ et J_b sont fermés.

4. Définition 8.10.

X est ouvert Soit $\tau \in X$. Si $\tau = \sup J_+$ alors $X = J_+$ est un ouvert de J_+. Supposons donc que $0 < \tau < \sup J_+$. Dans ce cas nous avons

$$\big(\tau, y_b(\tau)\big) \in \overline{W_\epsilon} \subset V, \tag{34.141}$$

et nous pouvons résoudre localement le problème de Cauchy

$$\begin{cases} y'(t) = f\big(t, y(t)\big) & \text{(34.142a)} \\ y(\tau) = \varphi(\tau, b) = y_b(\tau). & \text{(34.142b)} \end{cases}$$

Ce y existe jusqu'à $\tau + \eta$ (pour au moins un petit η), et par l'unicité de la solution, $y = y_b$ sur $[\tau, \tau + \eta[$. Ceci pour dire que le flot $\varphi(., b)$ existe au moins jusqu'à $\tau + \eta$.

Grâce à la proposition 34.22 nous pouvons évaluer

$$\|\varphi(\tau, b) - \varphi(\tau, a)\| = \|y_a(\tau) - y_b(\tau)\| \leqslant e^{L\tau}\|b - a\|. \tag{34.143}$$

Comme nous avions choisi $r = \epsilon e^{-LT}$ et $b \in \overline{B(a, r)}$ nous avons aussi $\|b - a\| \leqslant \epsilon e^{-LT}$ et donc

$$\|\varphi(\tau, b) - \varphi(\tau, a)\| \leqslant \epsilon e^{L(\tau - T)} < \epsilon \tag{34.144}$$

parce que nous avions $\tau < \sup J_+ \leqslant T$, ce qui garantit que $e^{L(\tau-T)} < 1$.

Est-ce que ceci nous garantit que $\tau + \eta \in X$? Il faudrait $\big(\tau + \eta, y_b(\tau + \eta)\big) \in \overline{W_\epsilon}$, c'est à dire $\|y_b(\tau + \eta) - y_a(\tau + \eta)\| \leqslant \epsilon$. L'ensemble J_+ étant fermé dans l'ouvert J_a, ce dernier déborde certainement. Prenons donc η assez petit pour que y_a existe jusqu'en $\tau + \eta$.

Vu que y_a et y_b sont continues, et qu'en τ elles sont distantes de moins de ϵ, en $\tau + \eta$, elles restent distantes de moins de ϵ (quitte à prendre encore η plus petit).

Ceci nous permet de conclure que X est ouvert.

X est non vide La solution y_b au problème

$$\begin{cases} y_b'(t) = f\big(t, y_b(t)\big) & \text{(34.145a)} \\ y_b(0) = b & \text{(34.145b)} \end{cases}$$

existe au moins localement et vérifie $\|y_b(0) - y_a(0)\| = \|b - a\| \leqslant \epsilon e^{-LT} < \epsilon$. Par continuité nous avons

$$\|y_b(t) - y_a(t)\| < \epsilon \tag{34.146}$$

pour tout t dans un voisinage de 0. Donc X est non vide.

Conclusion pour X La partie X est ouverte, fermée et non vide dans J_+ qui est connexe. Donc $X = J_+$ par la proposition 8.36(3).

La conclusion $X = J_+$ nous enseigne que pour tout $t \in J_+$ nous avons $]0, t] \in J_b$ et $\big(t, y_b(t)\big) \in \overline{W_\epsilon}$. Nous pouvons faire la même chose pour J_- et au final nous avons que pour tout $\tau \in J$ nous avons d'abord $\tau \in J_b$, ce qui prouve $J \subset J_b$. De plus pour tout $t \in J$ nous avons aussi

$$\big(t, y_b(t)\big) \in \overline{W_\epsilon} \subset V. \tag{34.147}$$

Nous en concluons que

$$J \times \overline{B(a, r)} \subset V. \tag{34.148}$$

Nous savons de plus que pour tout $b \in \overline{B(a, r)}$, $J \subset J_b$. Cela signifie que

$$J \times \overline{B(a, r)} \subset \mathcal{D}. \tag{34.149}$$

Mais $J \times \overline{B(a, r)}$ est un voisinage de (s, a) qui était au début de la preuve un point générique choisi dans $\mathcal{D}$. Donc $\mathcal{D}$ est ouvert parce qu'il contient un voisinage de chacun de ses points.

Il nous reste à voir que $\varphi\colon \mathcal{D} \to \Omega$ est localement Lipschitz. Soit donc le point générique (s, a) dans $\mathcal{D}$ et l'ensemble V qui avait été construit plus haut. Nous allons montrer que φ est Lipschitz sur $J \times \overline{B(a, r)} \subset V$. D'abord sur V, l'application f est Lipschitz, donc

$$\|\varphi(t, b_1) - \varphi(t, b_2)\| \leqslant e^{Lt}\|b_1 - b_2\| \leqslant e^{LT}\|b_1 - b_2\| \tag{34.150}$$

pour tout $t \in J$ et $b_1, b_2 \in \overline{B(a,r)}$.

Ensuite, f est bornée, majorée par M sur V, donc

$$\begin{aligned}
\|\varphi(t_1,b)-\varphi(t_2,b)\| &= |\int_{[t_1,t_2]} y'_(s)ds| && (34.151a)\\
&= |\int_{[t_1,t_2]} f\big(s,y_b(s)\big)| && (34.151b)\\
&\leqslant \int_{[t_1,t_2]} |f\big(s,y_b(s)\big)|ds && (34.151c)\\
&\leqslant M|t_1-t_2|. && (34.151d)
\end{aligned}$$

Et enfin nous prouvons que φ est localement Lipschitz. En posant $k = \max\{e^{LT}, M\}$ nous avons

$$\begin{aligned}
\|\varphi(t_1,b_1)-\varphi(t_2,b_2)\| &\leqslant \|\varphi(t_1,b_1)-\varphi(t_1,b_2)\| + \|\varphi(t_1,b_2)-\varphi(t_2,b_2)\| && (34.152a)\\
&\leqslant e^{LT}\|b_1-b_2\| + M|t_1-t_2| && (34.152b)\\
&\leqslant k\big(\|b_1-b_2\| + |t_1-t_2|\big) && (34.152c)\\
&\leqslant 2k\max\{\|b_1-b_2\|, |t_1-t_2|\} && (34.152d)\\
&= 2kd\big((b_1,t_2),(b_2,t_2)\big). && (34.152e)
\end{aligned}$$

Le flot φ est donc Lipschitz de constante $2k$. □

Exemple 34.26

Problèmes et choses à faire

Cet exemple doit être lu attentivement. Il me semble prouver que le flot n'est pas dérivable en la condition initiale sans que f le soit. Le document [386] semble dire le contraire. Je ne suis pas assez sûr de mon coup pour contredire.

Il n'y a pas de raisons de penser que $a \mapsto y_a(t)$ soit mieux que continue en sans hypothèses supplémentaires sur f. Pour illustrer cela nous considérons l'équation différentielle

$$\begin{cases} \dfrac{\partial X}{\partial s} = f\big(X(s),s\big) & (34.153a)\\ X(t) = x & (34.153b) \end{cases}$$

où t et x sont des paramètres fixés. Nous allons étudier la dérivabilité de X en x lorsque

$$f(x,t) = |x|. \tag{34.154}$$

Cela est un exemple typique de fonction autant Lipschitz que l'on veut sans être dérivable. L'équation différentielle est

$$\frac{\partial X}{\partial s}(s) = |X(s)|. \tag{34.155}$$

Si $x > 0$ alors $X(s) > 0$ dans un voisinage de $s = t$ et nous avons $X(s) = Ke^s$. La constante K se fixe par la condition initiale $X(t) = x$:

$$X(s) = xe^{s-t}. \tag{34.156}$$

Et cette solution tient en réalité pour tout s parce que $X(s)$ est alors toujours positif.

Si au contraire $x < 0$ nous avons la solution

$$X(s) = xe^{t-s}. \tag{34.157}$$

Au final,

$$X(s;x,t) = \begin{cases} xe^{t-s} & \text{si } x<0\\ 0 & \text{si } x=0\\ xe^{s-t} & \text{si } x>0 \end{cases} \tag{34.158}$$

L'application $(s,x,t)\mapsto X(s;x,t)$ est continue. En ce qui concerne la dérivée partielle $\partial_x X$ en $x=0$ nous avons :

$$\frac{\partial X}{\partial x}(s,0,t)=\lim_{\epsilon\to 0}\frac{X(s,\epsilon,t)-X(s,0,t)}{\epsilon}=\lim_{\epsilon\to 0}\frac{X(s,\epsilon,t)}{\epsilon}. \tag{34.159}$$

La limite à droite donne :

$$\lim_{\epsilon\to 0^+}\frac{X(s,\epsilon,t)}{\epsilon}=\frac{\epsilon e^{s-t}}{\epsilon}=e^{s-t}. \tag{34.160}$$

La limite à gauche donne :

$$\lim_{\epsilon\to 0^-}\frac{X(s,\epsilon,t)}{\epsilon}=e^{t-s}. \tag{34.161}$$

Les deux limites n'étant pas égales, la limite (34.159) n'existe pas[5] et l'application $(s,x,t)\mapsto X(s,x,t)$ n'est pas dérivable par rapport à x. △

Lemme 34.27 ([387]).
Soit un application $A\colon \overline{B(t_0,\tau)}\times\overline{B(a,R)}\to\mathcal{L}(\mathbb{R}^n)$ continue par rapport à sa première variable ($t_0\in\mathbb{R}$ et $a\in\mathbb{R}^n$). Alors en posant l'équation

$$\begin{cases}\dfrac{\partial\psi}{\partial t}(t,b)=A(t,b)\psi(t,b) & \text{(34.162a)}\\ \psi(t_0,b)=\psi_0. & \text{(34.162b)}\end{cases}$$

Nous avons l'estimation

$$\|\psi(t,v)-\psi(t,w)\|\leqslant\|\psi_0\|\tau\max_{s\in\overline{B(t_0,\tau)}}\|A(s,v)-A(s,w)\|\times \\ \times\exp\left(\tau\max_{s\in\overline{B(t_0,\tau)}}\max\{\|A(s,v)\|,\|A(s,w)\|\}\right) \tag{34.163}$$

pour tout $t\in\overline{B(t_0,\tau)}$ et $v,w\in V$.

Théorème 34.28 (Régularité C^1 du flot [387]).
Soit un intervalle ouvert I de $\mathbb{R}$ et un ouvert connexe Ω de $\mathbb{R}^n$. Soit une fonction $f\in C^1\big(I\times\Omega,\mathbb{R}\big)$, $a\in\Omega$ et $t_0\in I$.

Il existe un voisinage $W\times V=\overline{B(t_0,\tau)}\times\overline{B(a,r)}$ de (t_0,a) dans $I\times\Omega$ et une unique application $\varphi\colon W\times V\to\Omega$ telle que

$$\begin{cases}\dfrac{\partial\varphi}{\partial t}(t,x)=f\big(t,\varphi(t,x)\big) & \text{(34.164a)}\\ \varphi(t_0,x)=x & \text{(34.164b)}\end{cases}$$

pour tout $x\in V$.

L'application $(t,x)\mapsto\varphi(t,x)$ est de classe C^1.

Problèmes et choses à faire

La preuve qui suit doit être lue avec beaucoup d'attention, en particulier sur les incohérences possibles de notations, et sur les oublis possibles de précautions oratoires type « quitte à encore réduire les voisinages V et W ».

Démonstration. En termes de notations, pour $x\in\Omega$ fixé nous écrivons $y_x(t)$ pour $\varphi(t,x)$ et pour $t\in I$ fixé nous notons $\varphi_t(x)$ pour $\varphi(t,x)$.

De plus lorsque nous écrirons des choses comme $g\colon\mathbb{R}\to\mathbb{R}$, nous n'entendrons pas que g est effectivement définie sur tout $\mathbb{R}$. La notation « $g\colon\mathbb{R}\to\mathbb{R}$ » indiquera seulement que la variable de g est réelle, et que nous comptons préciser le domaine plus tard. Cette remarque s'applique seulement à cette démonstration et non à l'ensemble du livre.

5. Si vous comptez donner ça à manger au jury d'un concours, soyez prudent et n'écrivez pas l'équation (34.159) au tableau. Réfléchissez comment rédiger cela correctement.

Nous considérons $R > 0$ tel que $\overline{B(a, 2R)} \subset \Omega$ et ensuite nous posons $V = \overline{B(a, R)}$. La fonction y_x, solution pour la condition initiale $y_x(t_0) = x$ est définie sur $W = [t_0 - \tau, t_0 + \tau]$ et prend ses valeurs dans $\overline{B(x, R)}$. Ceci est parce que y_x est continue, alors en prenant τ assez petit, la valeur de $y_x(t)$ ne va pas s'éloigner de x lorsque t ne s'éloigne pas de t_0.

Nous savons déjà de la proposition 34.25 que φ est C^1 en t et localement Lipschitz en sa seconde variable, avec une constante Lipschitz uniforme sur $W \times V$. Elle est donc continue en tant que fonction

$$\varphi\colon V \times W \to \mathbb{R}^d. \tag{34.165}$$

La différentielle partielle Df Pout t fixé nous notons $Df_{(t,x)}$ la différentielle de f par rapport à x. C'est à dire que

$$\begin{aligned} Df_{(t,x)}\colon \mathbb{R}^n &\to \mathbb{R}^n \\ u &\mapsto \frac{d}{ds}\Big[f(t, x + su)\Big]_{s=0}. \end{aligned} \tag{34.166}$$

C'est un élément de $\mathcal{L}(\mathbb{R}^n)$, l'ensemble des applications linéaires de $\mathbb{R}^n$ vers $\mathbb{R}^n$. Nous allons montrer que

$$(t, x) \mapsto Df_{(t,x)} \tag{34.167}$$

est continue en tant qu'application $\mathbb{R} \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^n)$. Pour cela nous introduisons l'application d'inclusion $i\colon \mathbb{R}^n \to \mathbb{R} \times \mathbb{R}^n$, $i(u) = (0, u)$. Elle donne

$$Df_{(t,x)}(u) = \frac{d}{ds}\Big[f\big((t, x) + s(0, u)\big)\Big]_{s=0} = df_{(t,x)} \circ i(u). \tag{34.168}$$

Autrement dit

$$Df_{(t,x)} = df_{(t,x)} \circ i. \tag{34.169}$$

Or l'application $(t, x) \mapsto df_{(t,x)}$ est continue par hypothèse (f est de classe C^1) et l'application

$$\begin{aligned} \mathcal{L}(\mathbb{R} \times \mathbb{R}^n, \mathbb{R}^n) &\to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \\ A &\mapsto A \circ i \end{aligned} \tag{34.170}$$

est également continue. Donc $(t, x) \mapsto Df_{(t,x)}$ est continue[6].

L'équation aux variations Soit $x \in \Omega$. Nous introduisons l'opérateur

$$\begin{aligned} S_x\colon \mathbb{R} \times \mathcal{L}(\mathbb{R}^n) &\to \mathcal{L}(\mathbb{R}^n) \\ S_x(t, \psi) &= Df_{(t,y_x(t))} \circ \psi. \end{aligned} \tag{34.171}$$

Par ce que nous avons raconté, cela est une fonction continue en sa première variable et Lipschitz en sa seconde variable. Nous identifions $\mathcal{L}(\mathbb{R}^n)$ à $\mathbb{R}^{2^n}$.

Toujours pour chaque x considéré nous posons l'équation différentielle ordinaire

$$\begin{cases} \dfrac{\partial \psi}{\partial t}(t, x) = S_x\big(t, \psi(t, x)\big) & \text{(34.172a)} \\ \psi(t_0, x) = \mathrm{Id}\,. & \text{(34.172b)} \end{cases}$$

qui est une équation différentielle ordinaire pour $\psi\colon \mathbb{R} \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^n)$ rentrant dans le cadre de Cauchy-Lipschitz.

Quel est le domaine de définition de ψ pour sa première variable ? C'est un ouvert autour de t_0. Nous réduisons W de telle sorte que la solution ψ soit définie sur W. Idem pour la variable x qui est dans un voisinage de a.

L'équation (34.172) s'appelle l'**équation aux variations**. Nous allons montrer dans la douleur que ψ est continue et est la différentielle de φ_t, c'est à dire que

$$(d\varphi_t)_b = \psi(t, b). \tag{34.173}$$

6. Si quelqu'un peut prouver ça de façon moins verbeuse, je suis preneur. Il me semble que quel que soit la façon dont on s'y prend, sous le capot, on passe par la continuité de l'application (34.170).

ψ est continue en (t,x) (début) Il s'agit de majorer les deux termes de

$$\|\psi(t_1,a_1)-\psi(t_2,a_2)\|\leqslant\|\psi(t_1,a_1)-\psi(t_2,a_1)\|+\|\psi(t_2,a_1)-\psi(t_2,a_2)\|. \tag{34.174}$$

Premier terme Nous avons

$$\begin{aligned}\|\psi(t_1,b)-\psi(t_2,b)\|&=\|\int_{[t_1,t_2]}\frac{\partial\psi}{\partial t}(s,b)ds\| &(34.175a)\\ &\leqslant\int_{[t_1,t_2]}\|Df_{(s,y_b(s))}\circ\psi(s,b)\|ds &(34.175b)\\ &\leqslant\int_{[t_1,t_2]}\|Df_{(s,t_b(s))}\|\|\psi(s,b)\|ds &(34.175c)\\ &\leqslant|t_1-t_2|\max_{s\in[t_1,t_2]}\|Df_{(s,y_b(s))}\|\max_{s\in[t_1,t_2]}\|\psi(s,b)\|. &(34.175d)\end{aligned}$$

Nous allons majorer le second maximum. Prenons $t\in[0,\tau]$; et posons $A(u,b)=Df_{(u,y_b(u))}$ pour alléger les notations. Par l'équation de définition de ψ nous avons

$$\psi(t,b)=\psi(0,b)+\int_{[0,t]}A(u,b)\psi(u,b)du, \tag{34.176}$$

et donc

$$\|\psi(t,b)\|\leqslant\|\psi_0\|+\int_{[0,t]}\|A(u,b)\|\|\psi(u,b)\|du. \tag{34.177}$$

En y appliquant le lemme de Grönwall dans sa version 34.4 nous trouvons

$$\begin{aligned}\|\psi(s,b)\|&\leqslant\|\psi_0\|\exp\left(\int_{[0,s]}\|A(u,b)\|du\right) &(34.178a)\\ &\leqslant\|\psi_0\|\exp\left(s\max_{u\in[0,s]}\|A(u,b)\|\right). &(34.178b)\end{aligned}$$

En retournant à (34.175d) nous avons $\psi_0=\mathrm{Id}$ et donc $\|\psi_0\|=1$ et

$$\max_{s\in[t_1,t_2]}\|\psi(s,b)\|\leqslant\max_{s\in[t_1,t_2]}\exp\left(s\max_{u\in[0,t]}\|Df_{(u,y_b(u))}\|\right) \tag{34.179}$$

Là dedans nous pouvons remplacer t par $\max\{|t_1|,|t_2|\}$. Posons enfin, pour alléger les expressions

$$a(t_1,t_2,b)=\max_{s\in[t_1,t_2]}\|Df_{(s,y_b(s))}\|. \tag{34.180}$$

La majoration que nous retenons est :

$$\|\psi(t_1,b)-\psi(t_2,b)\|\leqslant|t_1-t_2|a(t_1,t_2,b)\exp\big(\max\{|t_1|,|t_2|\}a(0,t,b)\big). \tag{34.181}$$

Cela tend vers zéro lorsque $t_1\to t_2$.

Deuxième terme En ce qui concerne le second terme,

$$\|\psi(t,b_1)-\psi(t,b_2)\| \tag{34.182}$$

nous utilisons le lemme 34.27 qui donne, pour $t\in[t_0-\tau,t_0+\tau]$,

$$\begin{aligned}\|\psi(t,b_1)-\psi(t,b_2)\|\leqslant\tau\max_{s\in\overline{B}(0,\tau)}\|Df_{s,y_{b_1}(s)}-Df_{s,y_{b_2}(s)}\|\times\\ \times\exp\left(\tau\max\{\|Df_{s,y_{b_1}(s)},\|Df_{s,y_{b_2}(s)}\|\|\}\right).\end{aligned} \tag{34.183}$$

Dans notre cas, $t_0=0$, donc $t\in[-\tau,\tau]$. Vu la continuité de Df, nous avons

$$\max_{s\in\overline{B}(0,\tau)}\|Df_{s,y_{b_1}(s)}-Df_{s,y_{b_2}(s)}\|\to0 \tag{34.184}$$

lorsque $b_1\to b_2$.

ψ est continue en (t,x) (fin) Les deux bons calculs faits, nous avons, en repartant de (34.174),

$$\lim_{(t_1,b_1)\to(t_2,b_2)} \|\psi(t_1,b_1)-\psi(t_2,b_2)\| = 0, \tag{34.185}$$

ce qui signifie que ψ est une fonction continue de ses deux variables en même temps.

Différentiabilité de φ (début) Nous montrons maintenant que $D\varphi(t,x)$ existe. Pour rappel, D est la différentielle par rapport à la seconde variable. Nous sommes à étudier l'existence de $D\varphi_{(t,b)} = d(\varphi_t)_b$. Nous posons

$$\theta(t,h) = \varphi(t,b+h) - \varphi(t,b) = y_{b+h}(t) - y_b(t) \tag{34.186}$$

où b est le point où nous étudions la différentiabilité. Il est dans un voisinage du point a fixé depuis le début et autour duquel il existe un voisinage qui donne un sens à tout ce que nous avons fait jusqu'à présent. La dépendance de θ en b est implicite. Vu que φ est Lipschitz en sa seconde variable, nous avons la majoration

$$\|\theta(t,h)\| \leqslant C\|h\| \tag{34.187}$$

dès que $t \in V$ et $b, b+h \in W$.

De plus, parce que t_0 est le temps de la condition initiale nous avons

$$\theta(t_0,h) = y_{b+h}(t_0) - y_b(t_0) = a + h - a = h. \tag{34.188}$$

Et aussi, par définition de ψ :

$$\psi(t,b) = \psi_0 + \int_{t_0}^{t} \frac{\partial\psi}{\partial t}(s,b)ds = \psi_0 + \int_{t_0}^{t} Df_{(s,y_b(s))} \circ \psi(s,b) \tag{34.189}$$

En appliquant à h et en se souvenant que $\psi_0 = \mathrm{Id}$,

$$\psi(t,b)h = h + \int_{t_0}^{t} \Big(Df_{s,y_b(s)} \circ \psi(s,b)\Big)h\, ds. \tag{34.190}$$

Puis on peut faire un calcul assez classique en se souvenant que $\theta(t_0,h) = h$:

$$\theta(t,h) = \theta(t_0,h) + \int_{t_0}^{t} \big[\frac{\partial\varphi}{\partial t}(s,b+h) - \frac{\partial\varphi}{\partial t}(s,b)\big]ds \tag{34.191a}$$

$$= h + \int_{t_0}^{t} \big[f\big(s,y_{b+h}(s)\big) - f\big(s,y_b(s)\big)\big]ds. \tag{34.191b}$$

On fait la différence entre les deux :

$$\theta(t,h) - \psi(t,b)h = -\int_{t_0}^{t} \big[Df_{s,y_b(s)} \circ \psi(s,b)h - f\big(s,y_{b+h}(s)\big) + f\big(s,y_b(s)\big)\big]ds. \tag{34.192}$$

Nous y ajoutons et soustrayons $Df_{s,y_b(s)}\theta(s,h)$ et nous retenons la majoration suivante :

$$\begin{aligned}\|\theta(t,h) - \psi(t,b)h\| \leqslant &\int_{t_0}^{t} \|Df_{(s,y_b(s))}\psi(s,b) - Df_{(s,y_b(s))}\theta(s,h)\|ds \\ &+ \int_{t_0}^{t} \|f(s,y_{b+h}(s)) - f(s,y_b(s)) + Df_{(s,y_b(s))}\theta(s,h)\|ds.\end{aligned} \tag{34.193}$$

Nous allons encore majorer ces deux termes séparément. Soit $\epsilon > 0$.

Premier terme Ce qui est dans la norme à majorer est

$$Df_{(s,y_b(s))}\big(\psi(s,b)h-\theta(s,h)\big). \tag{34.194}$$

Vu que Df est continue et que y_b est continue [7], l'application $s\mapsto Df_{s,y_b(s)}$ est continue et donc de norme majorée sur le compact $[t_0,t]$. Nous rapellons la notation

$$a(t_0,t,b)=\max_{s\in[t_0,t]}\|Df_{s,y_b(s)}\|, \tag{34.195}$$

et nous majorons encore et toujours. D'abord

$$\int_{t_0}^{t}\|Df_{(s,y_b(s))}\big(\psi(s,b)h-\theta(s,h)\big)\|ds\leqslant a(t_0,t,b)\int_{t_0}^{t}\|\psi(s,b)-\theta(s,h)\|ds. \tag{34.196}$$

Deuxième terme Pour traiter le deuxième terme, nous allons provisoirement noter $x=y_b(s)$ et $y=y_{b+h}(s)$; entre autres, $y-x=\theta(s,h)$. Ce qui est écrit dans le second terme de (34.193) est

$$f(s,y)-f(s,x)+Df_{(s,x)}\theta(s,h)=f(s,y)-f(s,x)+Df_{(s,x)}(y-x) \tag{34.197}$$

Comme D ne s'applique pas à la variable s, nous pouvons alléger la notation et déduire de la différentiabilité de f qu'il existe un $\eta>0$ tel que $x,y\in W$ avec $\|y-w\|\leqslant\eta$ implique

$$\|f(y)-f(x)-Df_x(y-x)\|\leqslant\epsilon\|y-x\|. \tag{34.198}$$

Prenons $\|h\|\leqslant\eta/C$ (le C de (34.187)) ; en déballant les notations,

$$\|f\big(s,y_{b+h}(s)\big)-f\big(s,y_b(s)\big)-Df_{(s,y_b(s))}\theta(s,h)\|\leqslant\epsilon\|\theta(s,h)\|\leqslant\epsilon C\|h\|. \tag{34.199}$$

Les deux termes ensemble En remettant les deux dans (34.193) nous trouvons la majoration

$$\|\theta(t,h)-\psi(t,b)h\|\leqslant|t-t_0|\epsilon C\|h\|+a(t_0,t,v)\int_{t_0}^{t}\|\psi(s,b)h-\theta(s,h)\|ds \tag{34.200}$$

qui est encore de la graine à Grönwall avec

$$\begin{cases} u(t)=\|\theta(t,h)-\psi(t,b)\| & \text{(34.201a)}\\ b(t)=|t-t_0|\epsilon C\|h\| & \text{(34.201b)}\\ a(s)=a(t_0,t,v), & \text{(34.201c)}\end{cases}$$

la troisième étant une fonction constante. Cela donne, pour $t\in[t_0-\tau,t_0+\tau]$,

$$\|\theta(t,h)-\psi(t,b)h\|\leqslant|t-t_0|\epsilon C\|h\|+\int_{t_0}^{t}(s-t_0)\epsilon C\|h\|a(t_0,t,b)\exp\left(\int_{s}^{t}a(t_0,t,b)du\right)ds. \tag{34.202}$$

En valeur absolue, la différence $s-t_0$ est majorée par τ, l'intégrale dans l'exponentielle vaut $(t-s)a(t_0,t,b)$, et restons avec

$$\|\theta(t,h)-\psi(t,b)h\|\leqslant\tau\epsilon C\|h\|+\tau\int_{t_0}^{t}\epsilon C\|h\|a(t_0,t,b)e^{a(t_0,t,b)(t-s)}ds. \tag{34.203}$$

En supposant $t>t_0$ nous pouvons calculer l'intégrale. Si vous m'avez suivi jusqu'ici, vous devriez avoir de tels maux de tête que je vous donne la réponse :

$$\int_{t_0}^{t}a(t_0,t,b)e^{(t-s)a(t_0,t,b)}ds=e^{(t_0-t)a(t_0,t,b)}-1. \tag{34.204}$$

7. Il faut encore réduire les voisinages V et W pour que ceci ait un sens.

En remettant dans l'expression,

$$\|\theta(t,h)-\psi(t,b)h\| \leqslant \tau\epsilon C\|h\| + \epsilon C\|h\|a(t_0,t,b)\tau\big(e^{(t-t_0)}-1\big) = \tau\epsilon C\|h\|e^{(t-t_0)a(t,t_0,b)}. \tag{34.205}$$

Nous pouvons majorer $t-t_0$ par τ et $a(t,t_0,b)$ par $a(t_0-\tau,t_0+\tau,b)$ pour avoir la majoration

$$\|\theta(t,h)-\psi(t,b)h\| \leqslant \tau\epsilon C\|h\|e^{\tau a(t_0-\tau,t_0+\tau,b)}. \tag{34.206}$$

Différentiabilité de $\varphi(t,b)$ (fin) Nous écrivons la définition 12.105 de la différentiabilité : nous voulons vérifier que

$$\lim_{h\to 0}\frac{\varphi(t,b+h)-\varphi(t,b)-\psi(t,b)h}{\|h\|}=0. \tag{34.207}$$

Nous remplaçons $\varphi(t,b+h)-\varphi(t,b)$ par $\theta(t,h)$ et prenons la norme avec les majorations données :

$$\lim_{h\to 0}\frac{\|\varphi(t,b+h)-\varphi(t,b)-\psi(t,b)h\|}{\|h\|} \leqslant \lim_{h\to 0}\tau\epsilon C e^{\tau e(t_0-\tau,t_0+\tau,b)}. \tag{34.208}$$

Cela étant valable pour tout ϵ, nous en déduisons la nullité de la limite.

Nous avons démontré que φ était différentiable par rapport à sa deuxième variable et que

$$D\varphi_{(t,b)} = \psi(t,b). \tag{34.209}$$

Conclusion : φ est de classe C^1 Nous avons déjà prouvé que $(t,b)\mapsto\psi(t,b)$ est continue. Donc de (34.209) nous déduisons que les dérivées partielles $(t,b)\mapsto\frac{\partial\varphi}{\partial x_i}(t,b)$ sont continues. Mais comme φ est Lipschitz en t, la dérivée partielle $(t,b)\mapsto\frac{\partial\varphi}{\partial t}(t,b)$ est également continue. La continuité de toutes les dérivées partielles de φ nous donne la classe C^1 pour φ par la proposition 14.245.

□

Proposition 34.29 (Régularité C^p du flot[388])**.**
Soit un intervalle ouvert I de $\mathbb{R}$ et un ouvert connexe Ω de $\mathbb{R}^n$. Soit une fonction $f\in C^p\big(I\times\Omega,\mathbb{R}\big)$, ainsi que $a\in\Omega$ et $t_0\in I$.

Il existe un voisinage $W\times V=\overline{B(t_0,\tau)}\times\overline{B(a,r)}$ de (t_0,a) dans $I\times\Omega$ et une unique application $\varphi\colon W\times V\to\Omega$ telle que

$$\begin{cases}\dfrac{\partial\varphi}{\partial t}(t,x)=f\big(t,\varphi(t,x)\big) & \text{(34.210a)}\\ \varphi(t_0,x)=x & \text{(34.210b)}\end{cases}$$

pour tout $x\in V$.

L'application $(t,x)\mapsto\varphi(t,x)$ est de classe C^p.

Démonstration. Nous savons déjà par le théorème 34.28 que $(t,x)\mapsto\varphi(t,x)$ est de classe C^1. Nous supposons que f est de classe C^p avec $p\geqslant 2$.

Vu que φ et f sont de classe C^1, nous avons aussi que l'application $(t,x)\mapsto f\big(t,\varphi(t,x)\big)$ est de classe C^1. L'équation donne alors immédiatement le fait que

$$(t,x)\mapsto\frac{\partial\varphi}{\partial t}(t,x) \tag{34.211}$$

est de classe C^1.

En ce qui concerne la régularité par rapport aux autres variables, il faudra travailler un peu plus.

Une équation différentielle pour le flot Nous allons commencer par un habile jeu d'écriture : la formule

$$\varphi(t,x)=x+\int_{t_0}^{t}f\big(s,\varphi(s,x)\big)ds \tag{34.212}$$

devient

$$\varphi_t(x)=x+\int_{t_0}^{t}f\big(s,\varphi_s(x)\big)ds. \tag{34.213}$$

Dans le même ordre d'idée nous notons $f_s(x)=f(s,x)$, et ce qui se trouve dans l'intégrale (34.213) n'est autre que la fonction

$$g_s(x)=(f_s\circ\varphi_s)(x). \tag{34.214}$$

Tout cela pour différentier l'égalité (34.213) par la proposition 19.35 :

$$(d\varphi_t)_x=\mathrm{Id}+\int_{t_0}^{t}(dg_s)_x ds \tag{34.215a}$$

$$=\mathrm{Id}+\int_{t_0}^{t}(df_s)_{\varphi_s(x)}\circ(d\varphi_s)_s ds. \tag{34.215b}$$

Nous dérivons ensuite cela par rapport à t :

$$\frac{\partial}{\partial t}\Big((d\varphi_t)_x\Big)=(df_t)_{\varphi_t(x)}\circ(d\varphi_t)_x. \tag{34.216}$$

Cela est une égalité dans $\mathcal{L}(\mathbb{R}^n)$.

Nous introduisons la fonction

$$\begin{aligned}F\colon I\times\mathcal{L}(\mathbb{R}^n)\times\Omega&\to\mathcal{L}(\mathbb{R}^n)\\(t,A,x)&\mapsto(df_t)_{\varphi_t(x)}\circ A.\end{aligned} \tag{34.217}$$

En fait, à la place de I et Ω il faut prendre des petits voisinages dans lesquels les choses ont un sens. Ce que dit l'équation 34.216 est que l'application

$$\begin{aligned}A\colon I\times\Omega&\to\mathcal{L}(\mathbb{R}^n)\\(t,x)&\mapsto(d\varphi_t)_x\end{aligned} \tag{34.218}$$

vérifie l'équation différentielle

$$\begin{cases}\dfrac{\partial A}{\partial t}(t,x)=F\big(t,A(t,x),x\big) & \text{(34.219a)}\\ A(t_0,x)=\mathrm{Id}\,. & \text{(34.219b)}\end{cases}$$

Une autre équation différentielle Nous n'oublions pas l'équation différentielle pour la dérivée par rapport à t :

$$\frac{\partial\varphi}{\partial t}(t,x)=f\big(t,\varphi(t,x)\big). \tag{34.220}$$

Réécriture pour la différentielle Nous allons récrire l'équation (34.219) de façon à ce que le paramètre x soit inclus dans la condition initiale. De cette manière, la solution pourra profiter de la régularité C^1 du flot déjà prouvée dans le théorème 34.28.

Soit

$$\begin{aligned}g\colon I\times\big(\mathcal{L}(\mathbb{R}^n)\times\Omega\big)&\to\mathcal{L}(\mathbb{R}^n)\times\Omega\\\big(t,(A,x)\big)&\mapsto\Big(F(t,a,x),0\Big).\end{aligned} \tag{34.221}$$

Nous posons $E=\mathcal{L}(\mathbb{R}^n)\times\Omega$; c'est cet espace qui va jouer le rôle de Ω. Nous considérons à présent l'équation différentielle suivante pour $z_x\colon I\to E$:

$$\begin{cases}\begin{pmatrix}z_1'(t)\\z_2'(t)\end{pmatrix}=z'(t)=g(t,z(t))=\begin{pmatrix}F\big(t,z_1(t),z_2(t)\big)\\0\end{pmatrix} & \text{(34.222a)}\\ z(t_0)=(\mathrm{Id},x). & \text{(34.222b)}\end{cases}$$

Il devrait y avoir un indice (Id, x) à z parce que c'est sa condition initiale. La fonction g est de classe C^p, donc cette équation admet une unique solution dont le flot est de classe C^1. Autrement dit, si S est dans un voisinage de Id, l'application

$$(t, x) \mapsto z(t) \tag{34.223}$$

est de classe C^1. Nous allons montrer qu'en posant $A(t, x) = z_1(t)$, nous avons une solution de (34.219) (l'unicité de la solution impose que cette solution est effectivement la différentielle de $d\varphi_t$). D'abord, la seconde ligne de l'équation différentielle est $z_2'(t) = 0$, c'est à dire $z_2(t) = x$ pour tout t.

Sachant cela, la première équation devient

$$\begin{cases} z_1'(t) = F\big(t, z_1(t), x\big) & \text{(34.224a)} \\ z_1(t_0) = \mathrm{Id}, & \text{(34.224b)} \end{cases}$$

qui est l'équation différentielle pour A. Rappel : il y a partout une dépendance de z en sa condition initiale x que nous n'avons pas écrite pour des raisons de légèreté notionnelle. Il n'en reste pas moins que le flot de l'équation différentielle pour z est C^1, c'est à dire que $(t, x) \mapsto z_1(t)$ est de classe C^1.

Par conséquent, $(t, x) \mapsto A(t, x)$ est également C^1.

Régularité C^2 du flot Le fait que A soit C^1 n'implique pas que le flot le soit parce que le flot suit la même équation différentielle que A ne signifie pas que il soit égal. Il y a un raisonnement à faire.

Le fait est que si A est une solution de (34.219), alors $z(t) = \big(A(t, x), x\big)$ est solution de (34.222). C'est l'unicité de cette dernière qui permet de déduire l'unicité de la solution pour A.

Nous avons donc que l'unique solution A du système (34.219) est égale à $A(t, x) = (d\varphi_t)_x$ et est de classe C^1 par rapport à (t, x).

Donc $(t, x) \mapsto \varphi(t, x)$ est de classe C^2.

Régularité C^p Nous avons vu que le flot de $y' = f(t, y)$ est de classe C^2 dès que f est de classe C^2. Supposons que f soit de classe C^p et montrons que si le flot est de classe C^k $(k < p)$ alors il est de classe C^{k+1}.

Vu que le flot d'une équation différentielle de classe C^p est de classe C^k, en particulier celui de (34.222) est de classe C^k. Donc aussi la solution pour $A(t, x) = (d\varphi_t)_x$ est de classe C^k. Et vu que $(t, x) \mapsto (d\varphi_t)_x$ est de classe C^k, l'application φ est de classe C^{k+1}.

□

34.30.

Le théorème d'inversion locale 19.57 nous permet de dire que, pour t fixé, le flot $x \mapsto \varphi_t(x)$ est un C^p-difféomorphisme local.

Proposition 34.31 (Cauchy-Lipschitz avec paramètre, régularité C^p[389, 390, 391])**.**
Soit un intervalle ouvert I de $\mathbb{R}$, un connexe ouvert Ω de $\mathbb{R}^n$ et un intervalle ouvert Λ de $\mathbb{R}^d$. Soit une fonction $f \in C^p(I \times \Omega \times \Lambda, \mathbb{R}^n)$ localement Lipschitz en Ω. Soient $t_0 \in I$, $y_0 \in \Omega$ et $\lambda_0 \in \Lambda$. Il existe un voisinage compact de (t_0, y_0, λ_0) sur lequel le problème

$$\begin{cases} y_\lambda'(t) = f\big(t, y_\lambda(t), \lambda\big) & \text{(34.225a)} \\ y_\lambda(t_0) = y_0 & \text{(34.225b)} \end{cases}$$

possède une unique solution. De plus $(t, \lambda) \mapsto y_\lambda(t)$ est de classe C^p par rapport à ses deux variables.

Démonstration. Nous récrivons immédiatement le problème pour la fonction $y\colon I \times \Lambda \to \mathbb{R}^n$ donné par $y(t, \lambda) = y_\lambda(t)$:

$$\begin{cases} \dfrac{\partial y}{\partial t}(t, \lambda) = f\big(t, y(t, \lambda), \lambda\big) & \text{(34.226a)} \\ y(t_0) = y_0. & \text{(34.226b)} \end{cases}$$

Nous allons montrer que ce problème est en réalité équivalent à un problème sans paramètre. Nous posons $E = \Omega \times \Lambda$ et

$$\begin{aligned} g\colon I \times E &\to E \\ (t,x) &\mapsto \big(f(t,x_1,x_2),0\big) \end{aligned} \tag{34.227}$$

où x_1 est la composante Ω de x et x_2 est la composante Λ de x. Pour une valeur $\mu \in \Lambda$ donnée nous considérons le problème au condition initiales

$$\begin{cases} x'(t) = g\big(t,x(t)\big) & \text{(34.228a)} \\ x(t_0) = (y_0,\mu). & \text{(34.228b)} \end{cases}$$

Le théorème de Cauchy-Lipschitz que nous prenons sous la forme 34.29 nous indique que ce problème admet une unique solution maximale et que le flot $\big(t,(y_0,\mu)\big) \mapsto x_{(y_0,\mu)}(t)$ est de classe C^p.

Nous passons maintenant à la résolution du problème (34.226)

Existence d'une solution C^1 Nous montrons à présent que la fonction y donnée par

$$y(t,\mu) = x_{(t_0,\mu)}(t)_1 \tag{34.229}$$

est solution de (34.226). Vu que x a deux composantes, nous pouvons un peu déballer l'équation. Afin d'éviter les notations laborieuses nous allons noter x pour $x_{(t_0,\mu)}$ et donc $x_1(t)$ pour $x_{(t_0,\mu)}(t)$. Nous avons l'équation différentielle

$$\begin{pmatrix} x_1'(t) \\ x_2'(t) \end{pmatrix} = \begin{pmatrix} f\big(t,x_1(t),x_2(t)\big) \\ 0 \end{pmatrix} \tag{34.230}$$

avec la condition initiale

$$\begin{pmatrix} x_1(t_0) \\ x_2(t_0) \end{pmatrix} = \begin{pmatrix} y_0 \\ \mu \end{pmatrix}. \tag{34.231}$$

La seconde ligne de l'équation donne immédiatement $x_2(t) = \mu$ pour tout t. En injectant dans la première ligne :

$$x_1'(t) = f\big(t,x_1(t),\mu\big). \tag{34.232}$$

Or vue la définition de y, le nombre $x_1'(t)$ n'est autre que $\frac{\partial y}{\partial t}(t,\mu)$. La fonction y que nous avons définie vérifie donc

$$\frac{\partial y}{\partial t}(t,\mu) = f\big(t,y(t,\mu),\mu\big) \tag{34.233}$$

et la condition initiale $y(t_0) = x_1(t_0) = y_0$. Elle est donc bien solution du problème initial. De plus l'application $(t,\mu) \mapsto y(t,\mu) = x_{(t_0,\mu)}(t)_1$ est de classe C^p.

Unicité Pour l'unicité, soit on invoque la proposition 19.52 qui donne l'unicité dans les fonctions continues et a fortiori dans les fonctions C^1. Soit on fait le jeu inverse : on prouve qu'à chaque solution de (34.226) correspond une solution de (34.228), et l'unicité de la solution x donne l'unicité du côté de y.

□

Lemme 34.32.
Soit le problème

$$\begin{cases} \dfrac{\partial y}{\partial s}(s) = f\big(y(s),s\big) & \text{(34.234a)} \\ y(t) = x & \text{(34.234b)} \end{cases}$$

avec t et x fixés. Nous supposons que f est de classe C^p.

Alors l'application $t \mapsto y_x(s)$ est de classe C^p.

Démonstration. Soit t fixé, et l'équation différentielle

$$\begin{cases} \frac{\partial z}{\partial s}(s)a- = f\big(z(s), t-s\big) & (34.235a) \\ z(0) = x. & (34.235b) \end{cases}$$

Par le théorème 34.29, La solution z est de classe C^p en (s,x). En posant $y(s) = z(t-s)$ il est vite vérifié que y est solution de (34.234). C'est alors bien de classe C^p en t. □

34.8.4 Stabilité de Lyapunov

Définition 34.33.
Dans le cas de l'équation différentielle $y'(t) = f\big(y(t),t\big)$ pour $y\colon \mathbb{R} \to \mathbb{R}^n$, un point $a \in \mathbb{R}^n$ est un ***point d'équilibre*** *lorsque la fonction constante $y(t) = a$ est une solution.*

Le point d'équilibre $a \in \mathbb{R}^n$ est ***stable*** *si pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que $\|y(0) - a\| < \delta$ implique $\|y(t) - a\| < \epsilon$ pour tout t.*

Théorème 34.34 (Théorème de stabilité de Lyapunov[1, 392, 144, 384])**.**
Soit l'équation différentielle

$$\begin{cases} y'(t) = f(y) & (34.236a) \\ y(0) = y_0 & (34.236b) \end{cases}$$

avec une fonction $f\colon \mathbb{R}^n \to \mathbb{R}^n$ de classe C^1 vérifiant $f(0) = 0$ et $y_0 \in \mathbb{R}^n$. Nous supposons que l'application linéaire df_0 n'a que des valeurs propres dont la partie réelle est strictement négative.

Alors

(1) Il existe $k > 0$ tel que si $\|y_0\| < k$ alors la solution maximale est définie sur $\mathbb{R}$,

(2) pour le même nombre $k > 0$, si $\|y_0\| < k$ alors $y(t) \xrightarrow{t\to\infty} 0$ exponentiellement vite,

(3) la solution $y = 0$ est un point d'équilibre attractif.

Démonstration. Placer ici une phrase intelligente[8].

Prolégomène Le théorème de Cauchy-Lipschitz 19.49 nous enseigne que l'équation différentielle considérée possède une unique solution maximale (entre autres parce qu'une fonction de classe C^1 est localement Lipschitz) et nous nommons J l'intervalle sur lequel elle est définie.

Système linéarisé Nous posons $A = df_0$. La fonction $y_L(t) = e^{tA}y_0$ est solution du système linéarisé

$$\begin{cases} y'(t) = Ay(t) & (34.237a) \\ y(0) = y_0. & (34.237b) \end{cases}$$

Pour évaluer la norme de y_L nous utilisons le lemme 17.122 : il existe un polynôme P tel que

$$\|y_L(t)\| \leqslant P\big(|t|\big) \sum_{i=1}^{r} e^{\operatorname{Re}\lambda_i t}\|y_0\|. \tag{34.238}$$

Mais par hypothèse, $\operatorname{Re}(\lambda_i) < 0$ et si nous posons $\lambda = \max\{\operatorname{Re}(\lambda_i)\}$ nous avons $\lambda < 0$ et

$$\|y_L(t)\| \leqslant P\big(|t|\big)e^{\lambda t}\|y_0\|. \tag{34.239}$$

Donc quel que soit y_0 nous avons $\lim_{t\to\infty} \|y_L(t)\| = 0$ c'est à dire $\lim_{t\to\infty} y_L(t) = 0$.

8. Parce que sinon l'environnement `description` qui suit donne un mauvais effet.

Une forme linéaire Nous définissons la forme bilinéaire suivante sur $\mathbb{R}^n$:

$$b(x,y)=\int_0^\infty \langle e^{tA}x,e^{tA}y\rangle dt. \tag{34.240}$$

D'abord cela est bien défini pour tout $x,y\in\mathbb{R}^n$ parce que

$$\big|\langle e^{tA}x,e^{tA}y\rangle\big|\leqslant \|e^{tA}x\|\|e^{tA}y\|\leqslant P_1\big(|t|\big)P_2\big(|t|\big)e^{2\lambda t}\|x\|\|y\|, \tag{34.241}$$

qui est intégrable entre 0 et ∞ à cause de la décroissance exponentielle [9]. Montrons que b est définie positive. Soit donc $x\neq 0$ et calculons

$$b(x,x)=\int_0^\infty \|e^{tA}x\|^2dt. \tag{34.242}$$

Ce qui est dans l'intégrale est forcément (pas strictement) positif pour tout t. Mais si $x\neq 0$ alors $\|x\|^2$ est strictement positif et sur un voisinage de $t=0$ nous avons aussi $\|e^{tA}x\|^2$ qui est strictement positif. Ergo $b(x,x)>0$ dès que $x\neq 0$, ce qui signifie que b est strictement définie positive (lemme 11.181).

Nous notons $q\colon V\to\mathbb{R}$ la forme quadratique associée à b et aussi la norme qui va avec : $\|x\|_q=\sqrt{q(x)}$. En ce qui concerne le gradient $\nabla q\colon V\to V$, nous avons le petit calcul suivant[392] qui se base sur une des nombreuses formules du lemme 14.203 [10] :

$$\begin{aligned}\nabla q(x)\cdot y&=\frac{d}{dt}\Big[q(x+ty)\Big]_{t=0} &(34.243a)\\ &=\frac{d}{dt}\Big[q(x)+t^2q(y)+2tb(x,y)\Big]_{t=0} &(34.243b)\\ &=2b(x,y). &(34.243c)\end{aligned}$$

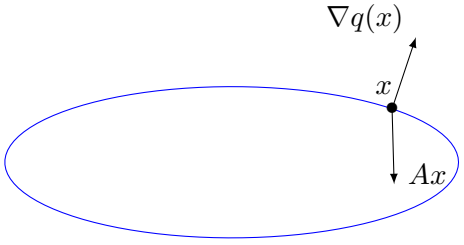
Nous avons aussi

$$\begin{aligned}\nabla q(x)\cdot Ax&=2b(x,Ax) &(34.244a)\\ &=2\int_0^\infty \langle e^{tA}x,e^{tA}Ax\rangle &(34.244b)\\ &=\int_0^\infty \frac{\partial}{\partial t}\Big(\langle e^{tA}x,e^{tA}x\rangle\Big)(t)dt &(34.244c)\\ &=\lim_{T\to\infty}\Big[\langle e^{tA}x,e^{tA}x\rangle\Big]_{t=0}^{t=T}. &(34.244d)\end{aligned}$$

Mais vu que $\|e^{tA}x\|\to 0$, pour $t\to\infty$ il ne reste que terme $t=0$ de la différence, c'est à dire

$$\nabla q(x)\cdot Ax=2b(x,Ax)=-\|x\|^2. \tag{34.245}$$

Étant donné que $\nabla q(x)$ est le vecteur dirigé vers l'extérieur de l'ellipsoïde de la courbe de niveau de q au point x, le vecteur Ax est dirigé vers l'intérieur.



9. Proposition 17.90.
10. Le fait que q soit différentiable est simplement le fait que b soit bilinéaire.

Majoration de $q\big(y(t)\big)'$ Nous posons

$$\begin{aligned} r\colon \mathbb{R}^n &\to \mathbb{R}^n \\ x &\mapsto f(x) - Ax. \end{aligned} \tag{34.246}$$

Soit y la solution maximale au problème (34.236) que nous pouvons aussi écrire sous la forme

$$y'(t) = r\big(y(t)\big) + Ay(t). \tag{34.247}$$

Calculons un peu ...

$$\begin{aligned} q\big(y(t)\big)' &= b\big(y(t), y(t)\big)' && \text{(34.248a)}\\ &= 2b(y, y') && \text{(34.248b)}\\ &= 2b\big(y, Ay\big) + 2b\big(y, r(y)\big) && \text{(34.248c)}\\ &= -\|y\|^2 + 2b\big(y, r(y)\big) && \text{(34.245) avec } x = y(t) \quad \text{(34.248d)}\\ &\leqslant -\|y\|^2 + 2\|y(t)\|_q \|r\big(y(t)\big)\|_q && \text{Cauchy-Schwarz : } |b(a,b)| \leqslant \|a\|_q\|b\|_q. \quad \text{(34.248e)} \end{aligned}$$

Chacun des deux termes peut encore être majoré. En ce qui concerne le premier, par équivalence des normes [11], il existe une constante C telle que $\|y\| \geqslant C\|y\|_q$. En renommant immédiatement C^2 en C, $\|y\|^2 \geqslant C\|y\|_q^2 = Cq(y)$.
Pour le second, nous allons utiliser la différentiabilité de r et le théorème des accroissements finis. Vu que $df_0 = A$ nous avons $dr_0 = df_0 - A = 0$ et de plus r est de classe C^1 parce que f l'est. Toutes les normes étant équivalentes [12] sur $\mathbb{R}^n$ nous pouvons exprimer la continuité de dr pour la norme $\|.\|_q$: si $\epsilon > 0$ est fixé alors il existe $\alpha > 0$ tel que $\|x\| < \alpha$ implique $\|dr_x\|_q < \epsilon$. Nous pouvons écrire les accroissements finis [13] pour la fonction r :

$$\|r(x) - r(0)\|_q \leqslant \sup_{a\in[0,x]} \|df_a\|\|x\|_q. \tag{34.249}$$

La chose facile à remarquer est que $r(0) = f(0) = 0$. En ce qui concerne les choses difficiles, vu que dr est continue (parce que r est C^1) il existe un $\delta > 0$ tel que $\|dr_a\|_q < \epsilon$ dès que $a \in B_q(0, \delta)$. Si nous prenons $\|x\|_q < \delta$ alors cette majoration est valable pour tous les éléments sur lequel est pris le supremum dans la formule (34.249). Donc

$$\|r(x)\|_q \leqslant \epsilon\|x\|_q \tag{34.250}$$

tant que $\|x\|_q \leqslant \delta$. Par conséquent, tant que $\|y(t)\|_q \leqslant \delta$ nous avons $\|r\big(y(t)\big)\| \leqslant \epsilon\|y(t)\|_q$. Nous continuons le calcul (34.248) :

$$\begin{aligned} q\big(y(t)\big)' &\leqslant Cq(y) + 2\epsilon\|y(t)\|_q^2 && \text{(34.251a)}\\ &= -(C - 2\epsilon)q(y). && \text{(34.251b)} \end{aligned}$$

Si ϵ est petit on a $C - 2\epsilon > 0$ et on pose $\beta = C - 2\epsilon$ pour écrire

$$q\big(y(t)\big)' \leqslant -\beta q\big(y(t)\big) \tag{34.252}$$

tant que $\|y(t)\|_q < \delta$.

Si $q(y_0) < \delta$ alors $q\big(y(t)\big) < \delta$ Nous posons [14]

$$\begin{aligned} t_1 &= \min\{t > 0 \text{ tel que } q\big(y(t)\big) = \delta\} && \text{(34.253a)}\\ t_2 &= \max\{t < 0 \text{ tel que } q\big(y(t)\big) = \delta\}. && \text{(34.253b)} \end{aligned}$$

11. Définition 12.3 et théorème 12.5.
12. Théorème 12.5.
13. Théorème 12.123.
14. t_1 est bien définit et est bien un minimum. J'en veux pour preuve [15] que si $q\big(y(t_s)\big) = \delta$, on peut prendre le minimum seulement sur les $t \in [0, t_s]$; or par continuité $q\big(y(t)\big) = \delta$ définit un fermé. Bref t_1 est un infimum sur un compact (fermé borné) et donc bien un minimum atteint.

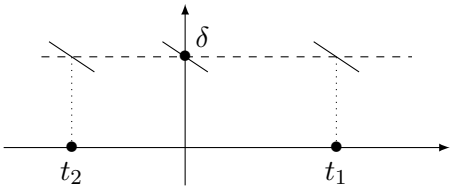
L'inégalité (34.252) est valable pour $t = 0$, $t = t_1$ et $t = t_2$; nous l'écrivons pour t_1 :

$$q\big(y(t)\big)'_{t=t_1} \leqslant -\beta q\big(y(t_1)\big) \leqslant -\beta\delta < 0 \tag{34.254}$$

Nous avons donc $q\big(y(t_1)\big) = \delta$ et $q\big(y(t)\big)'_{t=t_1} < 0$. Par conséquent pour tout t proche de t_1 avec $0 < t < t_1$ il y a $q\big(y(t)\big) > \delta$.

Pour la même raison, prise en $t = 0$ nous avons pour tout t proche de 0 avec $t > 0$ que $q\big(y(t)\big) < \delta$. Par continuité de $t \mapsto q\big(y(t)\big)$ cette fonction doit passer par la valeur δ dans $]t_2, 0[$ et $]0, t_1[$, ce qui contredit la maximalité de t_2 et la minimalité de t_1.

Ci-dessous, une partie de ce à quoi ressemble le graphe de $t \mapsto q\big(y(t)\big)$:



Deux conclusions :

— Vu que $q\big(y(t)\big)$ est borné pour tout $t \in \mathbb{R}$, nous sommes dans le cas (1) de l'alternative du théorème d'explosion en temps fini 34.18. Donc la solution $y(t)$ existe sur tout $\mathbb{R}$ pourvu que $\|y_0\|$ soit assez petit. Plus précisément par équivalence des normes, il existe un nombre $D > 0$ tel que $\|x\| \geqslant D\|x\|_q$ pour tout x. Si $\|y_0\| \leqslant D\delta$ alors

$$D\|y_0\|_q \leqslant \|y_0\| \leqslant D\delta, \tag{34.255}$$

qui donne immédiatement $\|y_0\|_q \leqslant \delta$, ce qui faut pour faire fonctionner l'existence de $y(t)$ pour tout t.

— Nous pouvons maintenant d'utiliser l'inégalité (34.252) pour tout $t \in \mathbb{R}$ sous la seule hypothèse que $q(y_0) < \delta$ au lieu de $q\big(y(t)\big) < \delta$.

La partie (1) de ce théorème est prouvée ; nous passons au reste à la partie (2). Pour cela nous supposons que $q(y_0) < \delta$.

À propos de $e^{\beta t}q(y)$ En sous-entendant la dépendance en t dans y nous avons

$$\Big(e^{\beta t}q(y)\Big)' = \beta e^{\beta t}q(y) + e^{\beta t}q(y)' = e^{\beta t}\big(\beta q(y) + q(y)'\big), \tag{34.256}$$

mais nous avons déjà prouvé que $q(y)' \leqslant -\beta q(y)$ (équation (34.252)), donc

$$\Big(e^{\beta t}q(y)\Big)' \leqslant 0 \tag{34.257}$$

Décroissance exponentielle Si $t \geqslant 0$, l'inégalité (34.257) donne

$$e^{\beta t}q\big(y(t)\big) \leqslant q(y_0), \tag{34.258}$$

c'est à dire

$$q\big(y(t)\big) \leqslant e^{-\beta t}q(y_0) \tag{34.259}$$

lorsque $t \geqslant 0$. Par équivalence des normes, nous avons des nombres D_1 et D_2 tels que

$$D_1\|x\|_q \leqslant \|x\| \leqslant D_2\|x\|_q \tag{34.260}$$

pour tout $x \in \mathbb{R}^n$. Nous avons donc pour tout $t \geqslant 0$ que

$$\|y(t)\| \leqslant D_2\|y(t)\|_q \leqslant D_2\|y_0\|_q e^{-\beta t}. \tag{34.261}$$

Pour rappel, $\beta > 0$, ce qui prouve la partie (2) du théorème.

Point d'équilibre Le point $y = 0$ est point d'équilibre (définition 34.33) parce que $f(0) = 0$, donc $y(t) = 0$ fonctionne. Dans ce cas, $y_0 = 0$.

Stabilité La stabilité est le fait que $\|y(t)\|_q \leqslant \delta$ dès que $\|y_0\|_q \leqslant \delta$.

□

34.8.5 Système proie-prédateur de Lotka-Volterra

Le système de **Lotka-Volterra** est l'équation différentielle suivante :

$$\begin{cases} x' = ax - bxy & (34.262a) \\ y' = -cy + dxy & (34.262b) \end{cases}$$

où a, b, c, d sont des constantes positives et avec la condition $x(t_0) > 0$, $y(t_0) > 0$.

En ce qui concerne l'interprétation des équations[393],

(1) $x(t)$ est le nombres proies

(2) $y(t)$ est le nombres prédateurs

(3) Les proies ont une reproduction rapide qui mène à une croissance exponentielle en absence de prédation (d'où le terme ax).

(4) Au contraire, les prédateurs meurent (ou migrent) rapidement lorsqu'ils n'ont pas de proies et nous supposons une décroissance exponentielle du nombre de prédateurs en l'absence de proies. D'où le terme $-cy$ avec le signe négatif.

(5) Les termes $-bxy$ et dxy sont les termes d'interaction entre le proies et les prédateurs. Ils sont proportionnels à la fréquence de leurs rencontres, lesquelles sont avantageuses pour les prédateurs et problématiques pour les proies.

Théorème 34.35 (Lotka-Volterra[144])**.**
Soient des constantes positives a, b, c, d et le système équations différentielles

$$\begin{cases} x' = ax - bxy & (34.263a) \\ y' = -cy + dxy & (34.263b) \\ x(t_0) > 0, y(t_0) > 0. & (34.263c) \end{cases}$$

Alors

(1) Les solutions sont positives sur leur domaines.

(2) Les solutions existent sur $\mathbb{R}$.

(3) Les solutions sont périodiques.

Démonstration. Nous divisions la preuve.

Comment théorème de Cauchy-Lipschitz s'applique Tel quel, le théorème de Cauchy-Lipschitz 19.49 ne s'applique pas parce qu'il demande une condition initiale pour avoir unicité. En ce qui concerne les notations, ce qui est noté « y » dans le théorème est ici le couple x, y et la fonction f est alors

$$f\left(t, \begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} ax - bxy \\ -cy + dxy \end{pmatrix}. \quad (34.264)$$

C'est une fonction continue localement Lipschitz partout par le lemme 14.257 et la proposition 14.256.

Nous savons cependant que les solutions sont de classe C^1 et que moyennant la donnée d'une condition initiale, la solution est unique.

Les solutions restent positives Supposons $x(s) = 0$ pour un certain $s > t_0$. Alors le solution

$$\begin{cases} x(t) = 0 & (34.265a) \\ y(t) = \exp(-ct) & (34.265b) \end{cases}$$

est une solution pour $[t_0, s + \epsilon]$. Par unicité de la solution avec condition initiale $s(s) = 0$, nous avons aussi $x(t_0) = 0$ pour toutes les solutions, ce qui contredit notre condition.

De la même façon, avoir $y(s) = 0$ donne une solution avec $y(t) = 0$ pour tout t et donc une contradiction.

Solutions sur $\mathbb{R}$ Nous montrons maintenant que les solutions sont définies sur $\mathbb{R}$.

Nous avons $x' < ax$, donc pour tout t où la solution est définie,

$$0 < x(t) < x(t_0)e^{a(t-t_0)}, \tag{34.266}$$

c'est à dire que la solution ne peut pas exploser en temps fini [16] : elle est bornée par le haut et le bas. Elle doit donc exister pour tout $t \in \mathbb{R}$. Par ailleurs, $y' < dxy$ donc

$$0 < y(t) < y(t_0)e^{d\int_{t_0}^{t} x(s)ds} \tag{34.267}$$

qui est également contraire à l'explosion en temps fini.

4 zones : monotonie Nous divisons $\mathbb{R}^2$ en quatre zones d'après les signes de $a-by$ et $c-dx$. Nous montrons que dans chacune de ces zones, les solutions sont monotones. Prenons par exemple la partie

$$\{(x,y) \in \mathbb{R}^2 \text{ tel que } a-by>0\} \times \{c-dx<0\}. \tag{34.268}$$

Vu l'équation $x' = x(a-by)$, tant que $\big(x(t),y(t)\big)$ est dans cette zone, la fonction x' a le signe de x et est donc positive. Donc x est croissante dans cette zone.

De la même façon, $y' = -y(c-dx)$ est y' a un signe constant dans la zone.

4 zones : on bouge Nous prouvons à présent qu'une solution ne reste pas dans une zone.

(1) Supposons que $\big(x(t_0),y(t_0)\big)$ soit dans la zone

$$\{a-by>0\} \quad \times \quad \{c-dx>0\} \tag{34.269a}$$
$$x'>0 \qquad\qquad y'<0 \tag{34.269b}$$

et que la solution reste dans cette zone (pour les $t > t_0$). Nous avons en particulier $x' > 0$, donc x est croissante tout en ayant la borne supérieure $x < c/d$. Par conséquent x a une limite que nous appelons $x_1 \in [0, \frac{c}{d}]$.

De la même façon, ; y est décroissante et bornée vers le bas par zéro. Donc y a une limite que nous notons $y_1 \in [0, y(t_0)]$.

Vu que x est bornée et de classe C^1 nous avons forcément $\lim_{t\to\infty} x'(t) = 0$. Mais vu que $x' = ax - bxy$ nous devons avoir

$$ax_1 - bx_1y_1 = 0. \tag{34.270}$$

Mais ni $x_1 > 0$ donc $a - by_1 = 0$, ce qui donne $y_1 = \frac{a}{b}$ et aussi $x_1 = \frac{c}{d}$. Bref, y est décroissante et tend vers a/b ; donc $y(t_0) > a/b$, ce qui contredit que $y(t_0)$ soit dans la zone considérée.

Étant donné que $x' > 0$ et $y' < 0$, la solution sort de la zone pour entrer dans la zone ...

(2) Supposons que $\big(x(t_0),y(t_0)\big)$ soit dans la zone

$$\{a-by>0\} \quad \times \quad \{c-dx<0\} \tag{34.271a}$$
$$x'<0 \qquad\qquad y'>0 \tag{34.271b}$$

et que la solution reste dans cette zone (pour les $t > t_0$). Les fonctions x et y sont convergentes. Par conséquent $\ln(y)$ converge aussi et vu que x est croissante,

$$\frac{y'}{y} = -c + dx \geqslant -x + dx(t_0) > 0 \tag{34.272}$$

Cela signifie que $\ln(y)'$ est toujours positive et bornée par le bas. Cela est impossible si y est borné.

Donc on sort de la zone pour entrer dans ...

16. Voir le corollaire 34.18.

(3) Supposons que $\big(x(t_0),y(t_0)\big)$ soit dans la zone

$$\{a-by<0\} \quad \times \quad \{c-dx<0\} \tag{34.273a}$$
$$x'<0 \qquad\qquad y'>0 \tag{34.273b}$$

et que la solution reste dans cette zone (pour les $t>t_0$).

Le même type de raisonnement fait passer à la zone...

(4) Supposons que $\big(x(t_0),y(t_0)\big)$ soit dans la zone

$$\{a-by<0\} \quad \times \quad \{c-dx>0\} \tag{34.274a}$$
$$x'<0 \qquad\qquad y'<0 \tag{34.274b}$$

et que la solution reste dans cette zone (pour les $t>t_0$). Encore une fois, cela nous fait sortir de la zone et retourne vers la première zone.

À ce moment nous voyons déjà que la relation entre proie et prédateurs, c'est un peu le mythe de Sisyphe

Une intégrale première Posons la fonction

$$H(x,y)=by+dx-a\ln(y)-c\ln(x). \tag{34.275}$$

Une simple dérivation montre que $x\mapsto H\big(x(t),y(t)\big)$ est constante. Nous considérons la fonction

$$\begin{aligned} f\colon\mathbb{R}&\to\mathbb{R}\\ s&\mapsto H\big(\frac{c}{d},s\big)\end{aligned} \tag{34.276}$$

dont la dérivée n'est autre que $f'(s)=b-\frac{a}{s}$. La fonction f est donc décroissante sur l'intervalle $[\frac{a}{b},\infty[$ et donc injective. Sur les changements de zones, il existe un t_0 tel que

$$x(t_0)=\frac{d}{c} \tag{34.277a}$$
$$y(t_0)>0. \tag{34.277b}$$

Pour cette valeur t_0 nous avons alors $H\big(x(t_0),y(t_0)\big)=f\big(y(t_0)\big)$. En posant $s_0=y(t_0)>0$ nous avons

$$H(x_0,y_0)=f(s_0) \tag{34.278}$$

et f étant injective, ce s_0 est la seule valeur de s à vérifier $H(x_0,y_0)=f(s)$.

Conclusion La fonction x passant d'une zone à l'autre, il existe un $t_1>t_0$ tel que $x(t_1)=a/b$. Nous avons évidemment

$$H\big(x(t_1),y(t_1)\big)=H(x_0,y_0) \tag{34.279}$$

parce que H est constante le long du mouvement. Cela se traduit par

$$H\big(\frac{a}{b},y(t_1)\big)=f(s_0), \tag{34.280}$$

et donc $y(t_1)=f(s_0)=y(t_0)$. Avec tout cela nous avons

$$\begin{cases} y(t_1)=y(t_0) & \text{(34.281a)}\\ x(t_1)=x(t_2)=\dfrac{a}{b}. & \text{(34.281b)}\end{cases}$$

Cela est donc un point par lequel la solution repasse. Par unicité de la solution, elle est donc périodique.

□

34.9 Équation du second ordre

34.9.1 Wronskien

Nous considérons ici une équation différentielle de la forme

$$y''(t) + q(t)y(t) = 0 \tag{34.282}$$

Dans ce point nous allons considérer la fonction q sans hypothèse de périodicité. L'équation de Hill (sous-section 34.9.4) sera la même équation, mais en supposant que q est périodique.

Nous commençons par argumenter que si q est continue, alors l'ensemble des solutions de l'équation (34.282) est un espace vectoriel de dimension deux. Pour cela il suffit d'appliquer la méthode de réduction de l'ordre (section 34.7) puis le théorème de dimension pour les systèmes linéaires (théorème 34.14). En effet si la fonction y_1 est solution de (34.282) si et seulement si le vecteur $Y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ est solution du système linéaire

$$Y'(t) = \begin{pmatrix} 0 & 1 \\ -q(t) & 0 \end{pmatrix} Y(t). \tag{34.283}$$

Soient deux solutions y_1 et y_2 de l'équation différentielle. Le **Wronskien** de ces deux solutions est le déterminant

$$W(t) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}. \tag{34.284}$$

Si nous considérons l'équation différentielle

$$y'' + py' + qy = 0, \tag{34.285}$$

le Wronskien peut être déterminé sans savoir explicitement y_1 et y_2 parce que $W = y_1y_2' - y_1'y_2$, et en dérivant,

$$\begin{aligned} W' &= y_1y_2'' + y_1'y_2' - y_1''y_2 - y_1'y_2' && \text{(34.286a)} \\ &= y_1(-py_2' - qy_2) - (-py_1' - qy_1)y_2 && \text{(34.286b)} \\ &= -p \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}, && \text{(34.286c)} \end{aligned}$$

c'est à dire

$$W' = -pW. \tag{34.287}$$

Il suffit donc de savoir une condition initiale pour obtenir une équation différentielle pour W.

34.9.2 Avec second membre

Une équation différentielle du second ordre avec un second membre se présente sous la forme

$$ay''(t) + by'(t) + cy(t) = v(t) \tag{34.288}$$

où $v(t)$ est une fonction donnée. Le truc est de commencer par résoudre l'équation différentielle sans second membre, c'est à dire trouver la fonction $y_H(t)$ telle que

$$ay_H''(t) + by_H'(t) + cy_H(t) = 0. \tag{34.289}$$

Cela se fait en utilisant la méthode du polynôme caractéristique.

Ensuite, il faut trouver une solution particulière $y_P(t)$ de l'équation avec le second membre. Une seule. Pour y parvenir, il faut du doigté et un peu de technique. Il faut faire des essais en fonction de ce à quoi ressemble le $v(t)$:

(1) Si $v(t)$ est un polynôme, alors il faut essayer un polynôme,
(2) Si $v(t) = \cos(\omega t)$ ou bien $v(t) = \sin(\omega t)$, alors essayer $y_P(t) = A\cos(t) + B\sin(\omega t)$,
(3) Si $v(t) = e^{\omega t}$, alors essayer $y_P(t) = Ae^{\omega t}$.

34.9.3 Équation $y'' + q(t)y = 0$

Nous allons donner quelques propriétés des solutions de l'équation

$$y'' + qy = 0 \tag{34.290}$$

en fonction de telle ou telle hypothèse sur q.

Proposition 34.36.
Si $q\colon \mathbb{R}^+ \to \mathbb{R}$ est continue et si

$$\int_0^\infty |q(t)|dt \tag{34.291}$$

converge, alors

(1) toute solution bornée de $y'' + qy = 0$ vérifie $\lim_{t\to\infty} y'(t) = 0$,

(2) l'équation $y'' + qy = 0$ admet des solutions non bornées.

Démonstration. Soit y une solution bornée, et intégrons l'équation différentielle entre 0 et ∞ :

$$\int_0^\infty y''(t)dt = -\int_0^\infty q(t)y(t)dt. \tag{34.292}$$

La fonction y étant bornée, l'hypothèse sur q permet de dire que l'intégrale de droite existe. Par ailleurs,

$$\int_0^\infty y'' = \lim_{a\to\infty}\int_0^a y'' = \lim_{a\to\infty} y'(a) - y'(0). \tag{34.293}$$

Cela justifie que la limite $\lim_{t\to\infty} y'(t)$ existe. Posons $\alpha = \lim_{t\to\infty} y'(t)$ et supposons par l'absurde que $\alpha \neq 0$. Soit $\epsilon > 0$ et λ assez grand pour que

$$\|y' - \alpha\|_{[\lambda,\infty[} < \epsilon. \tag{34.294}$$

Soit aussi $x > \lambda$. Nous avons

$$y(x) = y(\lambda) + \int_\lambda^x y'(t)dt \tag{34.295a}$$

$$\geqslant y(\lambda)\int_\lambda^x (\alpha - \epsilon) \tag{34.295b}$$

$$= y(\lambda) + \alpha x - \epsilon\lambda. \tag{34.295c}$$

En prenant la limite des deux côtés on voit que $y(x) \to \infty$ dès que $\alpha \neq 0$, ce qui est contraire aux hypothèses. Donc $\alpha = 0$.

Pour la seconde partie de la proposition, nous devons prouver que l'équation $y'' + qy = 0$ possède des solutions non bornées. Si l'équation a seulement des solutions bornées et si $\{u, v\}$ est une base de solutions, alors nous avons $u', v' \to 0$. Si nous reprenons l'équation (34.287) avec $p = 0$ nous savons que dans notre cas le Wronskien satisfait à $W' = 0$, c'est à dire qu'il est constant. Mais vu que u et v sont bornées et que les dérivées tendent vers zéro, nous avons $W(t) \to 0$ et donc $W(t) = 0$.

Or l'annulation identique du Wronskien contredit que $\{u, v\}$ serait une base de solutions. Donc il existe des solutions non bornées. □

Proposition 34.37.
Soit l'équation différentielle $y'' + qy = 0$. Si q est C^1, strictement positive et croissante, alors toutes les solutions sont bornées.

Démonstration. Soit y une solution et multiplions l'équation par $2y'$ (qui est non nulle par hypothèse) :

$$2y'y'' + 2qy'y = 0. \tag{34.296}$$

Nous allons intégrer cela en nous souvenant que $2y'y''$ est la dérivée de $(y')^2$. Pour tout $t>0$ nous avons

$$0=y'(t)^2-y'(0)^2+2\underbrace{\int_0^t q(t)y'(t)y(t)dt}_{\text{par partie}} \tag{34.297a}$$

$$=y'(t)^2-y'(0)^2+2\left([qy^2]_0^t-\int_0^t q'y^2\right) \tag{34.297b}$$

$$\tag{34.297c}$$

Le terme qui nous intéresse est celui qui contient $y(t)$:

$$2q(t)y(t)^2=-y'(t)^2+y'(0)^2+2q(0)y(0)^2+2\int_0^t q'y^2 \tag{34.298}$$

Nous pouvons majorer $-y'(t)^2$ par zéro et remplacer toutes les constantes par K :

$$q(t)y(t)^2\leqslant\int_0^t q'y^2+K=\int_0^t\frac{q'}{q}qy^2. \tag{34.299}$$

C'est le moment d'utiliser le lemme de Grönwall 34.4 avec $\phi=qy^2$ et $\psi=q'/q$. Les hypothèses de croissance et de positivité ont été posées exprès. Bref, on a

$$qy^2\leqslant K\exp\left(\int_0^t\frac{q'(s)}{q(s)}ds\right) \tag{34.300a}$$

$$=K\exp\left(\ln\frac{q(t)}{q(0)}\right) \tag{34.300b}$$

$$=K\frac{q(t)}{q(0)}. \tag{34.300c}$$

Notons que $q(0)$ est strictement positif. Nous déduisons que

$$y^2\leqslant\frac{K}{q(0)} \tag{34.301}$$

et donc y est bornée. □

34.9.4 Équation de Hill

L'équation **de Hill** est une équation différentielle de la forme

$$y''+qy=0 \tag{34.302}$$

où

(1) $q\in C^1(\mathbb{R},\mathbb{R})$,

(2) q est paire et π-périodique

Nous nous intéressons aux solutions complexes de cette équation différentielle.

Nous nommons $W\subset C^2(\mathbb{R},\mathbb{C})$ l'espace des solutions complexes de l'équation (34.302). Nous savons par ce qui a été dit en 34.9.3 que cet espace est de dimension deux. De plus avec le hypothèses faites ici sur q, nous savons que les solutions sont de classe C^3 parce que si y est une solution, alors l'équation $y''=qy$ nous indique que y est C^1 parce que y'' existe (y' est dérivable et donc continue). Mais si y est de classe C^1, alors le membre de droite qy est C^1 et donc y'' est C^1, ce qui prouve que y est de classe C^3. La récurrence ne va pas plus loin parce que q est seulement de classe C^1.

Nous considérons l'application de translation

$$\begin{aligned} T\colon C^2(\mathbb{R},\mathbb{C}) &\to C^2(\mathbb{R},\mathbb{C}) \\ (Ty)(x) &= y(x+\pi). \end{aligned} \tag{34.303}$$

En utilisant la règle de dérivation de fonctions composées, $(Ty)' = Ty'$ et $(Ty)'' = Ty''$, de telle sorte que si u est solution de l'équation (34.302), alors Tu est également solution. Donc W est un espace stable par T.

Le théorème 34.14 nous permet de choisir une base de W en imposant des conditions. Nous choisissons une base $\{y_1, y_2\}$ telles que

$$\begin{aligned} y_1(0) &= 1 \quad y_2(0) = 0 \\ y_1'(0) &= 0 \quad y_2'(0) = 1. \end{aligned} \tag{34.304}$$

Le théorème 34.14 nous assure que deux telles solutions existent et qu'elles forment une base de W parce que W est de dimension 2.

Lemme 34.38 ([383]).
Avec ce choix de base $\{y_1, y_2\}$ la matrice de T est donnée par

$$T = \begin{pmatrix} y_1(\pi) & y_2(\pi) \\ y_1'(\pi) & y_2'(\pi) \end{pmatrix}. \tag{34.305}$$

De plus la fonction y_1 est paire et la fonction y_2 est impaire.

Démonstration. Cherchons la matrice de T dans cette base en associant $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ à y_1 et $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ à y_2. Si $T = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, alors

$$Ty_1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix} = ay_1 + cy_2. \tag{34.306}$$

En évaluant cela en $t = 0$,

$$(Ty_1)(0) = ay_1(0) + cy_2(0) = a, \tag{34.307}$$

donc $a = (Ty_1)(0) = y_1(\pi)$. En dérivant (34.306), en tenant compte du fait que $(Ty_1)' = Ty_1'$ et en évaluant en $t = 0$, nous trouvons de même $c = y_1'(\pi)$. Puis le même cinéma avec y_2 donne

$$T = \begin{pmatrix} y_1(\pi) & y_2(\pi) \\ y_1'(\pi) & y_2'(\pi) \end{pmatrix}. \tag{34.308}$$

Passons maintenant à la parité de y_1 et y_2. Nous posons $\psi(t) = y_1(-t)$. Alors $\psi'(t) = -y_1'(-t)$ et $\psi''(t) = y_1''(t)$, tant et si bien que

$$\psi''(t) + q(t)\psi(t) = y_1''(-t) + q(t)y_1(-t) = 0. \tag{34.309}$$

donc ψ est une solution de l'équation. Mais

$$\begin{cases} \psi(0) = y_1(0) & \text{(34.310a)} \\ \psi'(0) = -y_1'(0) = 0, & \text{(34.310b)} \end{cases}$$

donc ψ a les mêmes conditions initiales que y_1. Par conséquent $\psi = y_1$ (par le l'unicité donnée dans le théorème de Cauchy-Lipschitz 19.49) et y_1 est paire. Nous procédons de même en partant de $\varphi(t) = -y_2(-t)$ pour trouver que $\varphi = y_2$ et que donc que y_2 est impaire.

□

Remémorons nous toutefois, pour calmer toute enthousiasme excessif, que T dépend de deux solutions et donc de la fonction q donnée dans l'équation.

Proposition 34.39 ([4]).
Nous considérons l'équation $y'' + qy = 0$ et sa base de solutions $\{y_1, y_2\}$ en suivant les notations données plus haut.

(1) Si $|\operatorname{Tr}(T)| < 2$, alors toutes les solutions de l'équation sont bornées.

(2) Si $|\operatorname{Tr}(T)| = 2$ alors nous avons une solution non bornée.

(3) Si $|\operatorname{Tr}(T)| > 2$ alors toutes les solutions de l'équation sont non bornées.

(4) Le cas $|\operatorname{Tr}(T)| = 2$ se présente si et seulement si $y_1'(\pi)y_2(\pi) = 0$.

Démonstration. Remarquons que le déterminant de la matrice T est égal au Wronskien des solutions y_1 et y_2 calculé en $t = \pi$. Calculons sa valeur :

$$W(y_1, y_2) = \det \begin{pmatrix} y_1 & y_2 \\ y_1' & y_2' \end{pmatrix} = y_1y_2' - y_1'y_2'. \tag{34.311}$$

En dérivant et en remplaçant y_i'' par $-qy_i$, nous trouvons tout de suite $W(y_1, y_2)' = 0$. Donc le Wronskien est constant et il est facile de le calculer en $t = 0$:

$$W(y_1, y_2)(0) = 1 - 0 = 1. \tag{34.312}$$

Donc pour tout t nous avons $W(y_1, y_2)(t) = 1$. En particulier

$$\det(T) = W(y_1, y_2)(\pi) = 1, \tag{34.313}$$

et notons au passage que T est inversible.

Nous écrivons le polynôme caractéristique de T sous la forme $\chi_T = X^2 - \operatorname{Tr}(T)X + \det(T)$, c'est à dire

$$\chi_T = X^2 - \operatorname{Tr}(T)X + 1, \tag{34.314}$$

dont le discriminant est $\Delta = \operatorname{Tr}(A)^2 - 4$.

Nous passons à présent aux différents points de la proposition.

(1) Si $|\operatorname{Tr}(T)| < 2$, alors $\Delta < 0$ et χ_T a deux racines complexes conjuguées que nous notons ρ et $\bar{\rho}$. De plus le produit des racines étant le terme indépendant, $\rho\bar{\rho} = 1$; en particulier $|\rho| = |\bar{\rho}| = 1$. Notons $\{u, v\}$ une base de vecteurs propres : $Tu = \rho u$ et $Tv = \bar{\rho}v$. Il est vite vu que la fonction $|u|$ est π-périodique :

$$|u|(t + \pi) = |u(t + \pi)| = |(Tu)(t)| = |(\rho u)(t)| = |\rho||u|(t) = |u|(t). \tag{34.315}$$

La fonction $|u|$ est continue[17] et périodique ergo bornée. La fonction $|v|$ est bornée pour la même raison et par linéarité, toutes les fonctions de W sont bornées.

(2) Si $\operatorname{Tr}(T) = \pm 2$, alors $\Delta = 0$ et χ_T a une racine réelle double[18] qui doit être ± 1. Soit u un vecteur propre de T pour la valeur propre ± 1. Nous avons

$$|u|(t + \pi) = |Tu(t)| = |\pm u(t)|, \tag{34.316}$$

ce qui prouve encore que $|u|$ est périodique et donc bornée.

Notons que nous n'avons pas d'informations sur le fait qu'une autre solution soit ou non bornée.

(3) Si $|\operatorname{Tr}(T)| > 2$, alors χ_T a deux racines réelles distinctes r et r' avec $rr' = 1$ (toujours les relations coefficients-racines). En raison de quoi $r' = r^{-1}$ et quitte à échanger r et r' nous supposons $|r| > 1$. L'opérateur est maintenant diagonalisable et nous considérons $\{u, v\}$ une base de vecteurs propres pour les valeurs propres r et r'. Une solution non nulle de l'équation s'écrit donc sous la forme

$$y = \alpha u + \beta v \tag{34.317}$$

avec $(\alpha, \beta) \neq (0, 0)$.

17. La fonction u elle-même n'est cependant pas garantie d'être périodique.

18. Ce qui n'implique pas le fait d'avoir deux vecteurs propres pour cette valeur propre, mais tout de même au moins un, voir l'exemple 11.136.

— Si $\alpha = 0$, alors $\beta \neq 0$ et nous choisissons une valeur t telle que $v(t) \neq 0$. Dans ce cas,

$$y(t + n\pi) = \beta v(t + n\pi) = \beta(T^n v)(t) = \beta(r')^n v(t), \tag{34.318}$$

et en faisant $n \to -\infty$ nous obtenons $\pm\infty$ suivant le signe de β.

— Si $\alpha \neq 0$, alors nous fixons [19] t tel que $u(t) \neq 0$. Alors

$$y(t + n\pi) = \alpha r^n u(t) + \beta(r')^n(t). \tag{34.319}$$

En faisant $n \to \infty$, nous avons $(r')^n \to 0$ tandis que le premier terme tend vers $\pm\infty$ suivant le signe de α.

(4) D'abord le théorème de Cayley-Hamilton 11.139 nous indique que $\chi_T(T) = 0$, c'est à dire que

$$T^2 - \mathrm{Tr}(T)T + 1 = 0. \tag{34.320}$$

Nous avons déjà mentionné le fait que T était inversible. Multiplions donc (34.320) par T^{-1} :

$$T + T^{-1} = \mathrm{Tr}(T)\mathbb{1}_2. \tag{34.321}$$

Vu que T^{-1} est l'endomorphisme $T^{-1}u(t) = u(t - \pi)$, sa matrice est donnée par

$$T^{-1} = \begin{pmatrix} y_1(-\pi) & y_2(-\pi) \\ y_1'(-\pi) & y_2'(-\pi) \end{pmatrix} = \begin{pmatrix} y_1(\pi) & -y_2(\pi) \\ -y_1'(\pi) & y_2'(\pi) \end{pmatrix} \tag{34.322}$$

où nous avons utilisé le fait que y_1 était paire et y_2 impaire (lemme 34.38). Si nous notons $T = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, alors $T^{-1} = \begin{pmatrix} a & -b \\ -c & d \end{pmatrix}$ et

$$T + T^{-1} = \begin{pmatrix} 2a & 0 \\ 0 & 2b \end{pmatrix}. \tag{34.323}$$

L'équation (34.321) donne alors, vu que $\mathrm{Tr}(T) = a + d$,

$$\begin{pmatrix} 2a & 0 \\ 0 & 2b \end{pmatrix} = \begin{pmatrix} a + d & 0 \\ 0 & a + d \end{pmatrix}, \tag{34.324}$$

ce qui donne immédiatement $a = d$. La matrice de T a donc comme forme $T = \begin{pmatrix} a & b \\ c & a \end{pmatrix}$ et $\mathrm{Tr}(T) = 2a$.

Donc $\mathrm{Tr}(T) = \pm 2$ si et seulement si $a = \pm 1$ et vu que $1 = \det(T) = a^2 - bc$, nous avons $a = \pm 1$ si et seulement si $bc = 0$, ce qui signifie exactement $y_1'(\pi)y_2(\pi) = 0$.

□

34.10 Différents types d'équations différentielles

34.10.1 Équation homogène

Une équation différentielle **homogène** est une équation de la forme

$$y' = f(t, y) \tag{34.325}$$

où $f(\lambda t, \lambda y) = f(t, y)$ pour tout $\lambda \neq 0$.

Elle se présente sous la forme

$$y' = \frac{\text{degré } n \text{ en } t, y}{\text{degré } n \text{ en } t, y}, \tag{34.326}$$

avec pas de y' à droite : juste du y et du t.

19. Mais pas trop hein ; nous aurons encore besoin d'assigner à t d'autres valeurs dans d'autres théorèmes.

Lemme 34.40.
L'équation $y' = f(t,y)$ est homogène si et seulement si $f(t,y)$ est une fonction de y/t seulement.

Pour résoudre l'équation homogène, on pose

$$z(t) = \frac{y(t)}{t}, \tag{34.327}$$

donc $tz = y$, et

$$y'(t) = tv'(t) + v(t), \tag{34.328}$$

à remettre dans l'équation de départ.

34.10.2 Équation de Bernoulli

C'est une équation du type

$$y' = a(t)y + b(t)y^\alpha \tag{34.329}$$

où $\alpha \neq 0$ ou 1. Pour la résoudre, on divise l'équation par y^α, et on pose $u = y^{1-\alpha}$, et on tombe sur une équation linéaire

$$u' = (1-\alpha)\big(a(t)u + b(t)\big). \tag{34.330}$$

34.10.3 Équation de Riccati

C'est une équation de la forme

$$y' = a(t)y^2 + b(t)y + c(t). \tag{34.331}$$

En général, on ne peut pas la résoudre, mais si on en connaît *a priori* des solutions particulières, alors on peut s'en sortir.

(1) Si on sait que $y_1(t)$ est une solution, alors on pose

$$y(t) = y_1(t) + \frac{1}{u(t)}, \tag{34.332}$$

et on obtient une équation linéaire

$$u' = -\big(2y_1(t)a(t) + b(t)\big)u - a(t). \tag{34.333}$$

(2) Si y_1 et y_2 sont solutions, alors nous avons y sous forme implicite

$$\frac{y - y_1}{y - y_2} = Ke^{\int a(t)\left(y_1(t) - y_2(t)\right)dt}. \tag{34.334}$$

Pour résoudre une équation de Ricatti, il faut donc d'abord deviner une ou deux solutions.

34.10.4 Équation différentielle exacte

34.10.4.1 Résolution lorsque tout va bien

Avant de vous lancer dans les équations différentielles exacte, vous devez lire la section sur les formes différentielles 14.18. Une équation différentielle exacte est de la forme $P(t,y) + Q(t,y)y' = 0$ que nous allons écrire sous la forme

$$P(t,y)dt + Q(t,y)dy = 0. \tag{34.335}$$

Nous savons que si $\partial_y P = \partial_t Q$, alors il existe une fonction $f(t,y)$ telle que $Pdt + Qdy = df$. Pour trouver une telle fonction, nous pouvons simplement intégrer la forme $Pdt + Qdy$. En effet, si $\gamma\colon [0,1] \to \mathbb{R}^2$ est un chemin tel que $\gamma(0) = (0,0)$ et $\gamma(1) = (t,y)$, alors en définissant

$$f(t,y) = \int_\gamma [Pdt + Qdt] = \int_0^1 \big[(P\circ\gamma)(u)dt + (Q\circ\gamma)(u)\big]\big(\gamma'(u)\big)du, \tag{34.336}$$

nous avons $df = Pdt + Qdy$. N'importe quel chemin fait l'affaire. Calculons avec $\gamma(u) = (tu, yu)$. La dérivée de ce chemin est donnée par

$$\gamma'(u) = t\begin{pmatrix}1\\0\end{pmatrix} + y\begin{pmatrix}0\\1\end{pmatrix}. \tag{34.337}$$

Étant donné que $dt\begin{pmatrix}a\\b\end{pmatrix} = a$ et $dy\begin{pmatrix}a\\b\end{pmatrix} = b$, nous avons

$$\begin{aligned} f(t,y) &= \int_0^1 [Pdt + Qdy]\big(\gamma(u)\big)\left(t\begin{pmatrix}1\\0\end{pmatrix} + y\begin{pmatrix}0\\1\end{pmatrix}\right)du \\ &= \int_0^1 P\big(\gamma(t)\big)tdu + \int_0^1 Q\big(\gamma(t)\big)ydu \\ &= \int_0^1 \big[tP(tu,uy) + yQ(tu,yu)\big]du. \end{aligned} \tag{34.338}$$

Nous retrouvons exactement la formule (22.302). Si ça t'étonne, c'est que tu n'as pas compris ;) Dans le cas où nous avons la fonction f qui vérifie $P = \partial_t f$ et $Q = \partial_y f$, l'équation (34.335) devient

$$\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}\frac{dy}{dt} = 0, \tag{34.339}$$

c'est à dire

$$\frac{d}{dt}\Big[f\big(t, y(t)\big)\Big] = 0, \tag{34.340}$$

dont la solution

$$f\big(t, y(t)\big) = C \tag{34.341}$$

donne la solution $y(t)$ sous forme implicite.

34.10.4.2 Facteur intégrant (quand tout ne va pas bien)

Si la forme $Pdt + Qdy$ n'est pas exacte, il n'existe pas de fonction f qui résolve l'affaire. Nous pouvons toutefois essayer de trouver un **facteur intégrant**. Nous cherchons une fonction M telle que

$$(MP)dt + (MQ)dy \tag{34.342}$$

soit exacte. Nous cherchons donc $M(t,y)$ telle que $\partial_y(MP) = \partial_t(MQ)$. En utilisant la règle de Leibnitz, nous trouvons l'équation suivante pour M :

$$M(\partial_y P - \partial_t Q) = Q(\partial_t M) - P(\partial_y M). \tag{34.343}$$

Cette équation est en générale extrêmement difficile à résoudre, mais dans certains cas particuliers, il est possible d'en trouver une solution à tâtons.

34.11 Distributions pour les équations différentielles

Nous commençons par définir l'espace $C^\infty\big(\mathbb{R}, \mathscr{S}'(\mathbb{R}^d)\big)$ en disant que $t \mapsto u_t$ est dans cet espace si

(1) pour tout $t \in \mathbb{R}$ nous avons $u_t \in \mathscr{S}'(\mathbb{R}^d)$,

(2) l'application $t \mapsto u_t$ est de classe C^∞.

Pour définir ce que nous entendons par une fonction de classe C^k à valeurs dans $\mathscr{S}'(\mathbb{R}^d)$ nous nous souvenons de la proposition 32.33.

34.11.1 Équation de Schrödinger

Théorème 34.41 (Équation de Schrödinger[4])**.**
Soit $g \in \mathscr{S}'(\mathbb{R}^d)$ et le problème

$$\begin{cases} \partial_t \tilde{u} - i\Delta\tilde{u} = 0 & (34.344a) \\ u_0 = g & (34.344b) \end{cases}$$

où $\tilde{u} \in C^\infty\big(\mathbb{R}, \mathscr{D}'(\mathbb{R}^d)\big)$ est lié à u par la remarque 32.55. Alors

(1) Il existe une unique solution dans $C^\infty\big(\mathbb{R}, \mathscr{S}'(\mathbb{R}^d)\big)$.

(2) Cette solution u vérifie de plus $\tilde{u} \in \mathscr{S}'(\mathbb{R} \times \mathbb{R}^d)$.

Démonstration. Nous allons donner explicitement une fonction $u \in C^\infty\big(\mathbb{R}, \mathscr{S}'(\mathbb{R}^d)\big)$ et nous allons vérifier l'équation (34.344a) en testant sur une fonction $\psi \in \mathscr{S}'(\mathbb{R}\times\mathbb{R}^d)$. Cela prouvera le point (2) ainsi que la partie existence de (1). Dans ce qui suit toutes les transformées de Fourier seront par rapport à la variable $x \in \mathbb{R}^d$ ou par rapport à ξ. Jamais par rapport à $t \in \mathbb{R}$.

Existence Pour $t \in \mathbb{R}$ nous posons [20]

$$u_t = \mathcal{F}^{-1}(f_t \hat{g}) \tag{34.345}$$

où $f_t \in \mathscr{S}(\mathbb{R}^d)$ est la fonction $f_t(x) = e^{-it\|x\|^2}$. Pour toute fonction $\varphi \in \mathscr{S}(\mathbb{R}^d)$ nous avons

$$u_t(\varphi) = (f\hat{g})\big(\mathcal{F}^{-1}(\varphi)\big) = \hat{g}\big(f\mathcal{F}^{-1}(\varphi)\big) = g\Big(\mathcal{F}\big(f\mathcal{F}^{-1}(\varphi)\big)\Big). \tag{34.346}$$

Le fait que $\mathcal{F}^{-1}(\varphi)$ soit une fonction Schwartz fait partie de la proposition 31.14. Pour chaque t nous avons bien $u_t \in \mathscr{S}'(\Omega)$.
De plus la fonction $h(t,x) = e^{-it\|x\|^2}(\mathcal{F}^{-1}\varphi)(x)$ est dans $C^\infty(\mathbb{R}\times\mathbb{R}^d)$, et par conséquent l'application

$$t \mapsto \hat{g}\big(h(t,.)\big) \tag{34.347}$$

est également C^∞ par la proposition 32.35. Ceci pour dire que $u \in C^\infty\big(\mathbb{R}, \mathscr{S}'(\mathbb{R}^d)\big)$. Il faut encore vérifier que cette fonction est bien une solution de notre problème. Nous testons cette équation sur $\psi \in \mathscr{S}(\mathbb{R}\times\mathbb{R}^d)$. Pour alléger les notations nous posons $\psi_t\colon x \mapsto \psi(t,x)$ et par conséquent aussi $(\partial_t\psi_t)(x) = (\partial_t\psi)(t,x)$. Nous avons :

$$\begin{aligned} \heartsuit &= (\partial_t\tilde{u} - i\Delta\tilde{u})(\psi) && (34.348a) \\ &= -\tilde{u}(\partial_t\psi) - i\tilde{u}(\Delta\psi) && (34.348b) \\ &= -\int_{\mathbb{R}} u_t\big((\partial_t\psi_t) + i(\Delta\psi_t)\big)dt && (34.348c) \end{aligned}$$

Ici nous nous souvenons du lemme 31.25 qui nous dit que nous pouvons permuter $\mathcal{F}^{-1}$ et ∂_t. Et pour l'autre terme il faut utiliser le lemme 31.13 avec $|\alpha| = 2$ et une somme pour obtenir que

$$\widehat{\Delta\varphi}(x) = -\|x\|^2\hat{\varphi}(x), \tag{34.349}$$

qui dans notre cas s'écrit sous la forme

$$\mathcal{F}^{-1}\Big((\Delta\psi_t)\Big)(x) = -\|x\|^2\mathcal{F}^{-1}\psi(t,x). \tag{34.350}$$

En remettant bout à bout,

$$\begin{aligned} \heartsuit &= -\int_{\mathbb{R}} (f_t\hat{g})\Big((\partial_t - i\|.\|^2)\mathcal{F}^{-1}\psi_t\Big)dt && (34.351a) \\ &= -\int_{\mathbb{R}} \hat{g}\Big(x \mapsto e^{-it\|x\|^2}(\partial_t - i\|x\|^2)(\mathcal{F}^{-1}\psi)(t,x)\Big)dt && (34.351b) \end{aligned}$$

20. En utilisant la définition (32.40) du produit d'une distribution par une fonction.

Pour alléger les notations nous notons $\check{\psi}_t(x) = (\mathcal{F}^{-1}\psi)(t,x)$. Nous avons

$$\partial_t\Big(e^{-it\|x\|^2}\check{\psi}_t(x)\Big) = -i\|x\|^2 e^{-it\|x\|^2}\check{\psi}_t(x) + e^{-it\|x\|^2}(\partial_t\check{\psi}_t),x) = e^{-it\|x\|^2}\big(\partial_t - i\|x\|^2\big)\check{\psi}_t(x); \tag{34.352}$$

cela nous permet d'un peu factoriser une dérivée dans $\heartsuit$:

$$\heartsuit = -\int_{\mathbb{R}} \hat{g}\Big(\partial_t\Big(e^{-it\|.\|^2}\check{\psi}_t(.)\Big)\Big)\,dt \tag{34.353a}$$

$$= -\int_{\mathbb{R}} \partial_t\hat{g}\Big(e^{-it\|.\|^2}\check{\psi}_t(.)\Big)\,dt \tag{34.353b}$$

$$= -\lim_{N\to\infty}\Big[\hat{g}\Big(e^{-i\|.\|^2}\check{\psi}_t(.)\Big)\Big]_{t=-N}^{t=N}. \tag{34.353c}$$

Histoire de bien comprendre les notations, il ne s'agit pas de calculer $\hat{g}\big(e^{-it\|.\|^2}\check{\psi}_t\big)$ pour un t général et de remplacer ensuite t par N et $-N$. En effet la valeur de $\hat{g}\big(e^{-it\|.\|^2}\check{\psi}_t\big)$ pour un t donné est celle qu'on obtient en calculant $\hat{g}(\dots)$ après avoir remplacé t par ce que l'on veut. Par conséquent, en posant $\varphi(t,\xi) = e^{-i\|\xi\|^2}\check{\psi}_t(\xi)$ nous avons :

$$\heartsuit = \lim_{N\to\infty}\Big[g\Big(x\mapsto\int_{\mathbb{R}} e^{-ix\xi}\varphi(t,\xi)d\xi\Big)\Big]_{t=-N}^{t=N} \tag{34.354a}$$

$$= \lim_{N\to\infty} g\Big(x\mapsto\int_{\mathbb{R}} e^{-ix\xi}\varphi(N,\xi)d\xi\Big) - \lim_{N\to\infty} g\Big(x\mapsto\int_{\mathbb{R}} e^{-ix\xi}\varphi(-N,\xi)d\xi\Big) \tag{34.354b}$$

La limite commute avec g parce que cette dernière est une distribution (continue). De plus la limite commute avec l'intégrale parce que ce qui est dedans est Schwartz. La fonction φ étant Schwartz, la limite est nulle. Donc

$$\heartsuit = 0. \tag{34.355}$$

Cela signifie que la fonction u proposée est bien une solution de l'équation de Schrödinger dans $C^\infty(\mathbb{R},\mathscr{S}'(\mathbb{R}^d))$.

Unicité Nous considérons deux solutions $u_1,u_2\in C^\infty\big(\mathbb{R},\mathscr{S}'(\mathbb{R}^d)\big)$ et la fonction $u=u_1-u_2$ doit satisfaire au problème

$$\begin{cases} (\partial_t\tilde{u} - i\Delta\tilde{u})(\psi) = 0 & \text{(34.356a)}\\ u_0 = 0. & \text{(34.356b)}\end{cases}$$

Nous allons montrer que seule la fonction $u_t = 0$ peut satisfaire à cela pour tout $\psi\in\mathscr{S}(\mathbb{R}\times\mathbb{R}^d)$. Nous allons même montrer qu'en imposant ces équations seulement sur la partie de $\mathscr{S}(\mathbb{R}\times\mathbb{R}^d)$ qui est à support compact par rapport à $\mathbb{R}$, la seule solution est $u_t=0$. Soit donc $\psi\in\mathscr{S}(\mathbb{R}\times\mathbb{R}^d)$ à support compact vis-à-vis de sa variable t. Alors

$$0 = -\tilde{u}(\partial_t\psi + i\Delta\psi) = -\int_{\mathbb{R}} u_t\Big((\partial_t\psi_t) + i(\Delta\psi_t)\Big)dt \tag{34.357}$$

où encore une fois $\partial_t\psi_t$ est la fonction $x\mapsto(\partial_t\psi)(t,x)$. Maintenant nous utilisons la proposition 32.58 pour dire que

$$\frac{d}{dt}\Big(u_t(\psi_t)\Big) = u_t^{(1)}(\psi_t) + u_t\left(\frac{\partial\psi}{\partial t}(t,.)\right) \tag{34.358}$$

pour écrire

$$0 = -\int_{\mathbb{R}}\frac{d}{dt}\big(u_t(\psi_t)\big) - u_t^{(1)}(\psi_t) + u_t\big(i(\Delta\psi)(t,.)\big)dt \tag{34.359}$$

Le premier terme est facile :

$$\int_{\mathbb{R}}\frac{d}{dt}\Big(u_t(\psi_t)\Big)dt = \lim_{N\to\infty}\Big[u_t(\psi_t)\Big]_{t=-N}^{t=N} = 0 \tag{34.360}$$

parce que ψ est à support compact par rapport à t. Nous restons donc avec

$$\int_{\mathbb{R}} u_t^{(1)}(\psi_t) - iu_t\big((\Delta\psi)(t,.)\big)dt = 0 \tag{34.361}$$

Nous traitons le terme en $u_t^{(1)}$ en utilisant le fait évident $T(\varphi) = (\mathcal{F}T)(\mathcal{F}^{-1}\varphi)$ et en remarquant le lemme 32.59 :

$$u_t^{(1)}(\psi_t) = (\mathcal{F}u_t^{(1)})(\mathcal{F}^{-1}\psi_t) = (\mathcal{F}u)_t^{(1)}(\mathcal{F}^{-1}\psi_t). \tag{34.362}$$

Pour l'autre terme on fait un peu la même chose en nous souvenant ce que fait la transformée de Fourier en traversant le laplacien :

$$u_t(\Delta\psi_t) = (\mathcal{F}u_t)(\mathcal{F}^{-1}\Delta\psi_t) = (\mathcal{F}u_t)\big(x \mapsto -\|x\|^2(\mathcal{F}^{-1}\psi_t)(x)\big). \tag{34.363}$$

En recollant encore :

$$\int_{\mathbb{R}} (\mathcal{F}u)_t^{(1)}(\mathcal{F}^{-1}\psi_t) + i(\mathcal{F}u_t)\big(\|.\|^2\mathcal{F}^{-1}\psi_t\big)dt = 0. \tag{34.364}$$

Cette équation est valable tant que $\psi \in \mathscr{S}(\mathbb{R}\times\mathbb{R}^d)$ avec support compact en t. Nous allons nous en créer une super cool. D'abord nous choisissons $\varphi \in \mathscr{S}(\mathbb{R}^d)$ et $\chi \in \mathscr{D}(\mathbb{R})$ et nous considérons [21]

$$\psi(t,x) = \mathcal{F}\Big(\xi \mapsto e^{it\|\xi\|^2}\varphi(\xi)\chi(t)\Big)(x). \tag{34.365}$$

Notons que la transformée de Fourier conserve le fait qu'une fonction soit Schwartz [22], mais pas le fait d'avoir support compact. Cependant nous ne prenons que la transformée de Fourier par rapport à x. Le résultat est donc une fonction ψ qui est Schwartz par rapport à x et support compact par rapport à t. Nous pouvons donc écrire (34.364) en utilisant la fonction (34.365) :

$$0 = \int_{\mathbb{R}} (\mathcal{F}u)_t^{(1)}\Big(x \mapsto e^{it\|x\|^2}\varphi(x)\chi(t)\Big) + i(\mathcal{F}u_t)\Big(x \mapsto \|x\|^2 e^{it\|x\|^2}\varphi(x)\chi(t)\Big)dt. \tag{34.366}$$

Là dedans, $\chi(t)$ peut sortir à la fois de la transformée de Fourier et de l'application des distributions ; il doit seulement rester dans l'intégrale. Dans le second terme nous allons utiliser l'égalité (due entre autre à la proposition 32.58) :

$$\begin{aligned}
\frac{d}{dt}\big(\hat{u}_t(e^{it\|.\|^2}\varphi)\big) &= \frac{d}{dt}\Big(u_t\big(\mathcal{F}e^{it\|.\|^2}\varphi\big)\Big) && (34.367a)\\
&= u_t^{(1)}\big(\mathcal{F}e^{it\|.\|^2}\varphi\big) + u_t\left(\frac{\partial}{\partial t}\mathcal{F}e^{it\|.\|^2}\varphi\right) && (34.367b)\\
&= (\mathcal{F}u_t^{(1)})\big(x \mapsto e^{it\|x\|^2}\varphi(x)\big) + (\mathcal{F}u_t)\big(x \mapsto i\|x\|^2 e^{it\|x\|^2}\varphi(x)\big) && (34.367c)\\
&= (\mathcal{F}u)_t^{(1)}\big(x \mapsto e^{it\|x\|^2}\varphi(x)\big) + (\mathcal{F}u_t)\big(x \mapsto i\|x\|^2 e^{it\|x\|^2}\varphi(x)\big). && (34.367d)
\end{aligned}$$

Et là, magie c'est exactement ce qui est dans (34.366). Donc

$$\int_{\mathbb{R}} \frac{d}{dt}\hat{u}_t\big(x \mapsto e^{it\|x\|^2}\varphi(x)\big)\chi(t)dt = 0 \tag{34.368}$$

pour toute fonctions à support compact χ. Donc la proposition 28.1 nous dit que

$$\partial_t\hat{u}_t\big(x \mapsto e^{it\|x\|^2}\varphi(x)\big) = 0. \tag{34.369}$$

C'est zéro partout et non seulement presque partout parce qu'en plus nous avons la continuité. Par conséquent pour tout $t \in \mathbb{R}$ nous avons

$$\hat{u}_t\big(x \mapsto e^{it\|x\|^2}\varphi(x)\big) = \hat{u}_0\big(x \mapsto \varphi(x)\big) = 0. \tag{34.370}$$

21. Le candidat qui parvient à effectivement présenter ça comme développement, il est fort.
22. Proposition 31.14.

Et cela est vrai pour toute fonction $\varphi \in \mathscr{S}(\mathbb{R}^d)$. Nous considérons donc $t_0 \in \mathbb{R}$ et une fonction $\theta \in \mathscr{S}(\mathbb{R}^d)$ pour construire

$$\varphi(x) = e^{-it_0\|x\|^2}\theta(x). \tag{34.371}$$

Nous avons alors $\hat{u}_{t_0}\big(x \mapsto \theta(x)\big) = 0$, ce qui signifie que $\hat{u}_{t_0} = 0$. Du coup pour tout $\theta \in \mathscr{S}(\mathbb{R}^d)$ nous avons $u_{t_0}(\mathcal{F}\theta) = 0$, mais comme la transformée de Fourier est une bijection de $\mathscr{S}(\mathbb{R}^d)$ (proposition 31.14) nous avons en fait $u_{t_0}(\theta) = 0$ pour tout $\theta \in \mathscr{S}(\mathbb{R}^d)$, c'est à dire $u_{t_0} = 0$ pour tout $t_0 \in \mathbb{R}$ et au final $u = 0$.

□

34.12 Équations différentielles du premier ordre

Définition 34.42 (Équation différentielle du premier ordre)**.**
Une ***équation différentielle du premier ordre*** *est une équation qui, sur un intervalle donné, I, décrit la relation entre une variable réelle, notée x ou t dans I, une fonction $y : I \to \mathbb{R}$, et la dérivée première de y qui on note y'.*

Souvent on écrit «$y'(x)$ = une formule contenante x et $y(x)$», c'est à dire

$$y'(x) = f(x, y(x)), \quad \text{pour } x \in I, \tag{34.372}$$

où f est une fonction de deux variables réelles.

Remarque 34.43.
La théorie des fonctions de deux variables ne sera pas abordée dans ce cours, nous allons nous contenter de prendre f dans (34.372) comme une simple notation.

On peut presque toujours omettre d'écrire la dépendance de y en x et écrire simplement (34.372) sous la forme $y' = f(x, y)$.

Définition 34.44 (Solution particulière d'une équation différentielle du premier ordre)**.**
Une ***solution particulière*** *de l'équation (34.372) sur l'intervalle I est une fonction $z : I \to \mathbb{R}$ telle que :*

(1) z est dérivable sur I ;

(2) $z'(x) = f(x, z(x))$, pour tout $x \in I$.

Définition 34.45 (Solution générale d'une équation différentielle du premier ordre)**.**
Résoudre une équation différentielle veut dire trouver l'ensemble qui contient toutes ses solutions particulières. Cet ensemble s'appelle ***solution générale*** *de l'équation.*

Exemple 34.46

(1) Résoudre une équation du type $y'(x) = f(x)$<++> revient à trouver l'ensemble des primitives de la fonction f, qui est donc la solution générale de cette équation. Il y a donc une infinité de solutions particulières, déterminées par une constante additive.
Si $f(x) = \sin(x)$ alors la solution générale sera $\mathcal{Y} = \{-\cos(x) + C \,:\, C \in \mathbb{R}\}$.

(2) L'équation

$$y' = y, \qquad x \in \mathbb{R}, \tag{34.373}$$

a peut-être été abordée dans votre cours de terminale lors de la définition de la fonction exponentielle. Sa solution générale est $\mathcal{Y} = \{Ce^x \,:\, C \in \mathbb{R}\}$. Ici aussi il y a une infinité de solutions particulières.

△

Remarque 34.47.
La solution générale d'une équation différentielle du premier ordre est une famille à un paramètre de fonctions.

Définition 34.48 (Équation differentielle du second ordre)**.**
Une ***équation différentielle du second ordre*** *est une équation qui, sur un intervalle donne, I, décrit la relation entre une variable réelle, notée x ou t dans I, une fonction $y : I \to \mathbb{R}$, et les dérivées première et seconde de y qui on note y' et y'' respectivement.*

On utilise la forme générale

$$y'' = f(x, y, y'), \quad \text{pour } x \in I. \tag{34.374}$$

où f est une fonction de trois variables réelles.

On peut définir de manière analogue les équations différentielles d'ordre supérieur. Les définitions de solution particulière et de solution générale se généralisent aux équations différentielles d'ordre supérieur à un.

Définition 34.49 (Trajectoire)**.**
La trajectoire tracée par une solution particulière y de l'équation (34.372) est le graphe de y en tant que fonction de x.

Exemple 34.50
Nous allons regarder de plus près l'équation (34.373), $y' = y$, pour tout $x \in \mathbb{R}$. Soient y_1 et y_2 deux solutions distinctes de cette équation. S'il existe un point $\bar{x}$ tel que $y_1(\bar{x}) = y_2(\bar{x})$ alors forcement $y_1(\bar{x})/y_2(\bar{x}) = 1$. Or, la solution générale de l'équation est $\mathcal{Y} = \{Ce^x : C \in \mathbb{R}\}$, donc $y_i(x) = C_i e^x$, $i = 1, 2$, où les C_i sont des constantes. Le rapport $y_1(\bar{x})/y_2(\bar{x})$ vaut C_1/C_2 et par conséquent $C_1 = C_2$. Ce résultat contredit l'hypothèse que les deux solutions soient distinctes. On a donc montré que *deux trajectoires distinctes de cette équations ne se croisent jamais.*

La figure 34.1 représente quelques trajectoires de l'équation. Si on les avait tracées toutes elles recouvriraient tout le plan x-y. Cela veut dire que *par tout point (x, y) passe une et une seule trajectoire de l'équation* (34.373).

△

Définition 34.51 (Condition initiale)**.**
Une ***condition initiale*** *pour l'équation (34.372) sur l'intervalle I est un point $(\bar{x}, \bar{y}) \in I \times \mathbb{R}$.*

On dit que la solution particulière z de (34.372) satisfait la condition initiale $(\bar{x}, \bar{y}) \in I \times \mathbb{R}$ si $z(\bar{x}) = \bar{y}$.

Définition 34.52 (Problème de Cauchy)**.**
L'association d'une équation différentielle et d'une condition initiale est appelée ***problème de Cauchy***

$$\begin{cases} y' = f(x, y), & x \in I, \\ y(\bar{x}) = \bar{y}. \end{cases} \tag{34.375}$$

Remarque 34.53.
Sous des conditions assez générales qui serons toujours vérifiées dans ce cours, tout problème de Cauchy admet une et une seule solution.

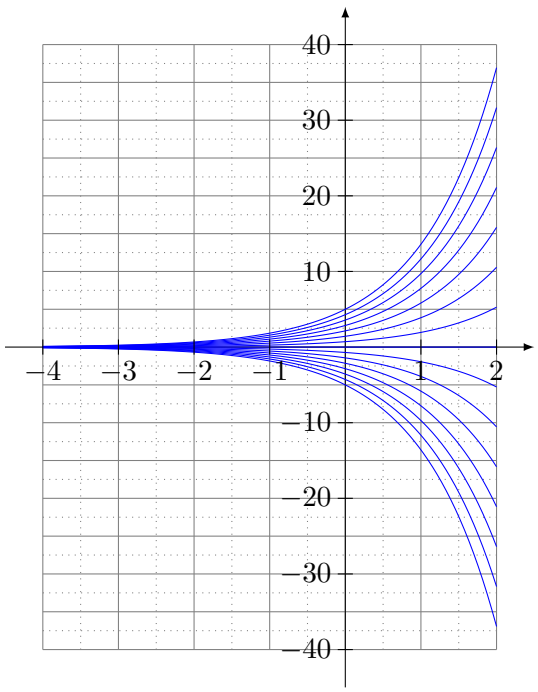


FIGURE 34.1 – Quelques trajectoires de l'équation $y' = y$.

Pour passer de la solution générale d'une équation différentielle de premier ordre à une solution particulière il faut choisir une valeur du paramètre. Comme il y a un seul paramètre une seule condition (la trajectoire de la solution doit passer par un point fixe du plan) peut suffire. Pour une équation différentielle de second ordre comme (34.374), nous aurons besoin de plus de conditions. Sans rentrer dans les détails, nous allons constater ce fait dans l'exemple suivant.

Exemple 34.54
La solution générale de l'équation

$$y'' = -y, \tag{34.376}$$

est $\mathcal{Y} = \{C_1 \cos(x) + C_2 \sin(x) \ : \ C_1, C_2 \in \mathbb{R}\}$. Remarquez que l'équation est du second ordre et que sa solution générale est une famille d'équations à deux paramètres réels. Ce sera toujours les cas pour les équations abordées dans la section 34.15. Pour déterminer une solution particulière de (34.376) il faut fixer les valeurs des deux paramètres et donc, en général, il sera nécessaire de donner deux conditions. △

Remarque 34.55.
Une condition comme $y(0) = 4$ nous dit que la constante $C_1 = 4$ mais elle ne nous permet pas de trouver C_2. Il y a donc une infinité de solutions de (34.376) qui satisfont à la condition $y(0) = 4$.

On peut fixer les deux conditions de deux manières différentes.

(1) Problème de Cauchy : on fixe une terne de valeurs réels $\bar{x}, \bar{y}, \bar{y}'$ et on cherche la solution telle que $y(\bar{x}) = \bar{y}$, $y'(\bar{x}) = \bar{y}'$.

Exemple 34.56
Les conditions $y(0) = 4$, $y'(0) = 15$ permettent de trouver la solution $z(x) = 4\cos(x) + 15\sin(x)$. △

(2) Problème aux bords : on fixe deux points dans le plan x-y, $A = (\bar{x}, \bar{y})$ et $B = (\tilde{x}, \tilde{y})$, et on cherche la solution dont la trajectoire passe par A et B, c'est à dire, on impose $y(\bar{x}) = \bar{y}$, $y(\tilde{x}) = \tilde{y}$.

Exemple 34.57

Les conditions $y(0) = 4$, $y(\pi/2) = 15$ permettent de trouver la solution $z(x) = 4\cos(x) + 15\sin(x)$. △

34.13 Premier ordre, variables séparables

Pour certaines équations différentielles la recherche d'une solution particulière se réduit à une recherche de primitive moyennant un changement de variables.

Définition 34.58 (Équation différentielle du premier ordre à variables separables)**.**
Une ***équation différentielle du premier ordre à variables séparables*** *est une équation qui, pour tout les x dans un intervalle donné, I, peut se mettre sous la forme*

$$f(y)y' = g(x), \tag{34.377}$$

où f et g sont deux fonctions de $\mathbb{R}$ dans $\mathbb{R}$.

Nous pouvons intégrer les deux côtes de l'égalité par rapport à x et obtenir

$$\int f(y(x))y'(x)\,dx = G(x) + C,$$

où G est une primitive de g et C une constante réelle. Il est facile à ce point d'effectuer une changement de variable dans le membre de gauche de l'équation en posant (sans surprise) $y = y(x)$ et donc $y'(x)\,dx = dy$.

$$\int f(y(x))y'(x)\,dx = \int f(y)\,dy = F(y(x)) + C,$$

où F est une primitive de f et C une constante réelle. En somme nous avons

$$F(y(x)) = G(x) + C,$$

et, si F admet une fonction réciproque, alors

$$y(x) = F^{-1}(G(x) + C). \tag{34.378}$$

Remarque 34.59.
L'expression de F^{-1} peut être difficile à calculer. Il sera alors préférable de garder y dans la forme implicite.

Exemple 34.60
L'équation

$$3y^2y' = x, \qquad \text{pour tout } x \in \mathbb{R}, \tag{34.379}$$

est une équation à variables séparables. Pour reprendre les notations du début du chapitre, ici $f(y) = 3y^2$ et $g(x) = x$. En intégrant de deux côtes on trouve

$$y^3 = \frac{x^2}{2} + C.$$

La fonction $F(y) = y^3$ est une bijection de $\mathbb{R}$ dans $\mathbb{R}$, donc nous pouvons écrire la solution générale de l'équation (34.379) dans la forme

$$\mathcal{Y} = \left\{ \left(\frac{x^2}{2} + C \right)^{1/3} \text{ tel que } C \in \mathbb{R} \right\}.$$

△

Exemple 34.61
En intégrant de deux côtes l'équation à variables séparables

$$2yy' = x, \text{pour tout } x \in \mathbb{R}, \tag{34.380}$$

on trouve

$$y^2 = \frac{x^2}{2} + C.$$

La fonction $F(y) = y^2$ est *n'est pas inversible* sur tout $\mathbb{R}$, et on sait que $\sqrt{y^2} = |y|$. Au moment de rendre y explicite on doit choisir entre

$$y = \left(\frac{x^2}{2} + C\right)^{1/2} \qquad \text{ou} \qquad y = -\left(\frac{x^2}{2} + C\right)^{1/2}.$$

Ce choix se fait suivant la condition initiale, si elle est donnée. S'il n'y a pas de condition initiale nous pouvons écrire que la solution générale est l'ensemble

$$\mathcal{Y} = \left\{ y : \mathbb{R} \to \mathbb{R} \text{ tels que } y^2 = \frac{x^2}{2} + C \text{ et } C \in \mathbb{R} \right\}.$$

△

Exemple 34.62
On considère le problème de Cauchy

$$\begin{cases} e^y y' = \frac{1}{x+3}, & x \in]-\infty, -3[, \\ y(-4) = 0. \end{cases} \tag{34.381}$$

En intégrant des deux côtes nous trouvons

$$e^y = \ln(|x+3|) + C.$$

Nous pouvons alors imposer la condition initiale et obtenir $e^0 = \ln(|-4+3|) + C$, c'est à dire $C = 1 - \ln(1) = 1$.

Remarque 34.63.
L'énoncé du problème de Cauchy dit que x peut varier dans $]-\infty, -3[$, mais nous voyons maintenant que la solution n'est pas définie sur toute la demi-droite, parce que e^y est toujours positif et $\ln(|x+3|) + 1$ est positif seulement pour $x < -(1/e + 3) \approx -3,3679$.

Donc la solution du problème de Cauchy est $y(x) = \ln(|x+3|)+1$ pour tout $x \in]-\infty, -(1/e+3)[$.
△

Exemple 34.64
Attention, cet exemple est le plus important de la section !
On considère l'équation à variables séparables

$$y' = \sin(x)y, \qquad x \in \mathbb{R}. \tag{34.382}$$

Dans ce cas, pour pouvoir écrire l'équation dans la forme (34.377) il faut pouvoir multiplier les deux côtés par $1/y$. Il faut donc éliminer tout de suite le cas où $y = 0$.

Si $y = 0$ alors $y' = 0$ et on a une solution constante (on dit souvent : une solution stationnaire) de l'équation. Par ailleurs les trajectoires des solutions ne peuvent pas se croiser ; donc si y_G est une solution non nulle de l'équation (34.382) alors $y_G(x) \neq 0$ pour tout x[23]. Il n'y a donc aucun danger à diviser par y dans la recherche d'une solution non identiquement nulle.

23. Ça vaut la peine de prendre un peu de temps pour bien comprendre cela.

Supposons maintenant que $y \neq 0$ et écrivons $y'/y = \sin(x)$. En intégrant des deux côtes on trouve

$$\ln(|y|) = -\cos(x) + C,$$

d'où

$$|y| = e^{-\cos(x)+C} = e^C e^{-\cos(x)}.$$

Si on avait impose une condition initiale alors on pourrait déterminer une solution particulière de l'équation en choisissant une valeur de la constante C. Nous pouvons observer cependant que la fonction exponentielle est bijective de $\mathbb{R}$ dans $\mathbb{R}^{+,\star}$ et par conséquent il n'y a pas de perte de généralité en disant que la solution générale de l'équation est

$$\mathcal{Y} = \left\{ y \,:\, |y| = Ke^{-\cos(x)}, \text{ pour } K \in \mathbb{R}^{+,\star} \right\} \cup \{y \equiv 0\}.$$

Il n'empêche qu'il serait plus élégant d'écrire la solution générale de l'équation sous une forme plus explicite, sans valeur absolue. Nous pouvons le faire en nous nous rappelant que

$$|x| = \begin{cases} x & \text{si } x \geqslant 0, \\ -x & \text{si } x < 0, \end{cases}$$

Il suffit alors d'autoriser K dans $\mathbb{R}^\star$ pour éliminer la valeur absolue.

Pour écrire la solution générale de façon encore plus compacte nous observons que si $K = 0$ alors $y \equiv 0$, c'est à dire, on retrouve la solution constante nulle.

Finalement, la solution générale de cette équation sera toujours écrite sous la forme suivante

$$\mathcal{Y} = \left\{ y = Ke^{-\cos(x)}, \text{ pour } K \in \mathbb{R} \right\}. \tag{34.383}$$

△

34.14 Équations différentielles linéaires du premier ordre

Définition 34.65 (Équation différentielle linéaire du premier ordre)**.**
Soit $I \subset \mathbb{R}$ un intervalle .

*Une **équation différentielle linéaire du premier ordre** est une équation différentielle de la forme*

$$a(x)y' + b(x)y = c(x), \quad \text{pour } x \in I, \tag{34.384}$$

où a, b, c sont des fonctions de $\mathbb{R}$ dans $\mathbb{R}$ et $a \neq 0$ pour tout $x \in I$.

On dit que a, b, c sont les coefficients de l'équation (34.384).

Remarque 34.66.
Une fonction $f : \mathbb{R} \to \mathbb{R}$ est dite *linéaire* si pour tout x_1, x_2 dans $\mathbb{R}$ et pour tout couple de constantes λ et μ on a

$$f(\lambda x_1 + \mu x_2) = \lambda f(x_1) + \mu f(x_2). \tag{34.385}$$

Ces équations différentielles sont dites linéaires parce que la partie de l'équation qui contient y (le membre de gauche) satisfait la propriété (34.385) par rapport à y. En effet par les propriétés de la dérivée nous avons que

$$a(x)(\lambda y_1 + \mu y_2)' + b(x)(\lambda y_1 + \mu y_2) = \lambda(a(x)y_1' + b(x)y_1) + \mu(a(x)y_2' + b(x)y_2).$$

Définition 34.67.
*L'équation (34.384) est dite **homogène** quand c est la fonction nulle. Si (34.384) n'est pas homogène on dit que l'équation*

$$a(x)y' + b(x)y = 0, \tag{34.386}$$

*est son **équation homogène associée**.*

Toute équation linéaire du premier ordre homogène est une équation du premier ordre à variables séparables, comme nous en avons vu l'exemple 34.64. Nous n'allons pas répéter les détails du procédé pour trouver sa solution générale, qui aura la forme suivante

À retenir 34.68

$$\mathcal{Y}_h = \left\{ Ke^{-\int \frac{b(x)}{a(x)}\,dx} \;:\; K \in \mathbb{R} \right\}. \tag{34.387}$$

Proposition 34.69. *(1) Soit y_p une solution particulière de l'équation (34.384) et y_h une solution particulière de l'équation homogène associé (34.386). Alors la fonction somme $z = y_p + y_h$ est encore une solution particulière de l'équation (34.384).*

(2) Soient y_1 et y_2 deux solutions particulières de (34.384). Alors la fonction différence $w = y_1 - y_2$ est un solution particulière de (34.386).

Démonstration. (1)

$$a(x)\left(y_p + y_h\right)' + b(x)\left(y_p + y_h\right) - c(x) = \big(a(x)y_p' + b(x)y_p - c(x)\big) + \big(a(x)y_h' + b(x)y_h\big) = 0. \tag{34.388}$$

(2)

$$a(x)\left(y_1 - y_2\right)' + b(x)\left(y_1 - y_2\right) = \big(a(x)y_1' + b(x)y_1 - c(x)\big) - \big(a(x)y_2' + b(x)y_2 - c(x)\big) = 0. \tag{34.389}$$

□

Cette proposition permet de démontrer le théorème suivant, qui est le plus important de cette section.

Théorème 34.70.
Soit y_p une solution particulière de l'équation (34.384) et $\mathcal{Y}_h$ la solution générale de l'équation (34.386), alors la solution générale de l'équation (34.384) est l'ensemble

$$\mathcal{Y} = \mathcal{Y}_h + y_p = \{z = y_h + y_p \;:\; y = h \in \mathcal{Y}_h\}. \tag{34.390}$$

À retenir 34.71
La résolution d'une équation différentielle linéaire du premier ordre comporte trois étapes :

(1) résolution de l'équation homogène associée ;
(2) recherche d'une solution particulière de l'équation non homogène ;
(3) somme de la solution générale de l'équation homogène et de la solution particulière trouvée au point précédent.

La partie qui nous manque encore est de savoir comment trouver une solution particulière de l'équation non homogène (34.384). Si la fonction c dans (34.384) est une constante ou un polynôme simple, ou une exponentielle alors on peut essayer de deviner. Cette méthode cependant n'est pas la plus sûre pour des débutants.

Exemple 34.72
On considère l'équation

$$y' - 5y = 10, \qquad x \in \mathbb{R}. \tag{34.391}$$

Comme tous les coefficients de l'équation sont constants on peut essayer de trouver une solution constante.

Toutes les fonctions constantes on dérivée nulle, par conséquent, si une solution constante existe elle doit satisfaire $-5y = 10$, ce qui veut dire que la solution constante est $y(x) \equiv -2$. △

Exemple 34.73
On considère l'équation

$$xy' + y = x + 1, \qquad x \in \mathbb{R}^{+,\star}. \tag{34.392}$$

Comme le membre de droite de l'équation est un polynôme de degré un on cherche une solution de la forme $y(x) = Ax + B$ avec A et B dans $\mathbb{R}$.

Par substitution on obtient $Ax + (Ax + B) = x + 1$, c'est à dire que une solution particulière de l'équation est $y(x) = x/2 + 1$. △

Exemple 34.74
L'équation

$$xy' - y = x + 1, \qquad x \in \mathbb{R}^{+,\star}. \tag{34.393}$$

ressemble beaucoup à celle de l'exemple précédent, cependant il n'existe pas un polynôme de degré un qui en soit solution.

Dans un cas comme celui-ci, il faut rapidement abandonner la divination et replier sur la méthode, plus technique mais plus sûre, dite *variation de la constante* △

34.14.1 Méthode de variation de la constante

— Soit $\mathcal{Y}_h$ la solution générale de l'équation homogène associé à (34.384). Il s'agit d'une famille à un paramètre de fonctions. La première étape de cette méthode consiste à construire un candidat solution particulière y_p en remplaçant le paramètre dans $\mathcal{Y}_h$ par une fonction $C : \mathbb{R} \to \mathbb{R}$ à déterminer.

Exemple 34.75
L'équation homogène associée à $y' - y = \cos(x)$ est $y' - y = 0$, dont la solution générale est $\mathcal{Y}_h = \{Ce^x \ : \ C \in \mathbb{R}\}$. Le candidat solution sera alors $y_p = C(x)e^x$, avec C fonction à déterminer. △

— La deuxième étape de cette méthode consiste à injecter y_p dans l'équation. Cela permet de trouver une équation différentielle à variables séparables pour C, en principe plus facile à résoudre que l'équation de départ.

Exemple 34.76
On continue avec l'exemple précédent. On a $y_p' = C'(x)e^x + C(x)e^x$, d'où

$$(C'(x)e^x + C(x)e^x) - C(x)e^x = \cos(x),$$

c'est à dire

$$C'(x) = \cos(x)e^{-x}.$$

△

— La troisième étape de la méthode consiste à trouver une solution particulière de l'équation différentielle pour C et, par conséquent déterminer une y_p.

Exemple 34.77
La solution générale de

$$C'(x) = \cos(x)e^{-x}.$$

est $\mathcal{C} = \left\{ e^{-1}\frac{(\sin(x)-\cos(x))}{2} + K \ : \ K \in \mathbb{R} \right\}$. Il nous suffit une solution particulière, nous pouvons donc choisir $K = 0$ et alors la solution particulière de (34.384) sera $y_p(x) = \frac{\sin(x)-\cos(x)}{2}$. △

Remarque 34.78.
Le plus souvent en intégrant l'équation pour C on en trouvera la solution générale. Dans ce cas on peut remplacer C par cette solution générale et obtenir d'un seul coup la solution générale de l'équation (34.384) , c'est à dire sans faire la somme entre la solution générale de l'homogène associée et la solution particulière.

Exemple 34.79
Dans l'exemple qu'on vient de voir la solution générale de (34.384) est

$$\mathcal{Y} = \mathcal{Y}_h + y_p = \left\{ Ce^x + \frac{(\sin(x) - \cos(x))}{2} \ : \ C \in \mathbb{R} \right\}. \tag{34.394}$$

On obtient le même résultat est écrivant $\mathcal{Y} = \left\{ e^{-x}\left(e^{-1}\frac{(\sin(x)-\cos(x))}{2} + K \right) \ : \ K \in \mathbb{R} \right\}$. Notez qu'on a changé le nom du paramètre de C à K seulement pour souligner qu'on obtient de même résultat par deux chemins différents, sinon les deux expressions sont équivalentes ! △

34.15 Équations différentielles linéaires du second ordre

Définition 34.80 (Équation différentielle linéaire du second ordre)**.**
Une ***équation différentielle linéaire du second ordre*** *est une équation différentielle de la forme*

$$a(x)y'' + b(x)y' + c(x)y = d(x), \quad \text{pour } x \in I, \tag{34.395}$$

où a, b, c et d sont des fonctions de $\mathbb{R}$ dans $\mathbb{R}$ et $a \neq 0$ pour tout $x \in I$.
On dit que a, b, c et d sont les coefficients de l'équation (34.395).

Dans ce cours nous allons étudier exclusivement le cas où a, b et c sont des fonctions constantes.

Définition 34.81 (Équation différentielle linéaire du second ordre homogène)**.**
Une ***équation différentielle linéaire du second ordre homogène*** *est une équation différentielle de la forme (34.395), telle que le coefficient d est nul.*

À toute équation de la forme (34.395) on peut associer une équation homogène exactement comme on a fait dans la section précédente pour les équations linéaires du premier ordre.

34.15.1 Équations différentielles linéaires du second ordre homogènes à coefficients constants

Remarque 34.82.
L'application qui à la fonction y fait correspondre $a(x)y'' + b(x)y' + c(x)y$ est linéaire, au sens de la remarque 34.66.

Cela nous dit en particulier, que si y_1 et y_2 sont deux solutions de l'équation homogène alors toute leur combinaison de la forme $z = \lambda y_1 + \mu y_2$, avec λ et μ dans $\mathbb{R}$, est encore une solution.

Jusqu'ici nous avons toujours travaillé avec des fonctions définies sur $\mathbb{R}$ et à valeurs dans $\mathbb{R}$. Dans cette section nous nous autorisons à passer par des fonctions définies sur $\mathbb{R}$ et à valeurs dans $\mathbb{C}$, mais cela sera uniquement une étape dans nos calculs. Au final toutes les solutions que nous allons considérer sont des fonctions à valeurs dans $\mathbb{R}$.

La solution générale **à valeurs dans les complexes** d'une équation de ce type a la forme

$$\mathcal{Y}_h^{\mathbb{C}} = \{C_1 e^{r_1 x} + C_2 e^{r_2 x} \,:\, C_1,\, C_2 \in \mathbb{C},\, x \in I\}\,, \tag{34.396}$$

où r_1 et r_2 sont aussi des nombres complexes. Remarquez que la solution générale est une famille à deux paramètres. Il faut aussi observer que en tout cas l'intervalle I dans lequel varie x est un intervalle dans $\mathbb{R}$, parce que I est une des données du problème.

À partir de cette information nous pouvons, pour toute équation donnée, chercher la solution générale **complexe** par substitution. Il suffit de remplacer y dans l'équation par e^{rx} et chercher les valeurs de r qui nous conviennent.

Si notre équation de départ est

$$ay'' + by' + cy = 0, \quad \text{pour } x \in I, \tag{34.397}$$

alors la substitution nous donne

$$e^{rx}\left(ar^2 + br + c\right) = 0.$$

Il est connu que la fonction exponentielle ne prend pas la valeur 0, par consequent ce qui s'annule est le polynôme de degré deux $ar^2 + br + c$. Il est donc très facile de trouver les valeurs de r qu'on pourra utiliser comme r_1 et r_2 dans la solution générale **complexe**.

Si $b^2 - 4ac > 0$ **:** le polynôme admet deux solutions réelles et distinctes, r_1 et r_2 ;

Si $b^2 - 4ac < 0$ **:** le polynôme admet deux solutions complexes conjuguées, $r_1 = \alpha + i\beta$ et $r_2 = \alpha - i\beta$;

Si $b^2 - 4ac = 0$ **:** le polynôme admet une solution réelle double $r = r_1 = r_2$.

Il faut maintenant écrire la solution générale **réelle** de l'équation, qui est celle que nous intéresse vraiment. La façon de l'obtenir est différente dans les trois cas.

Si $b^2 - 4ac > 0$ **:** la solution générale réelle a la même forme que la solution complexe, (34.396), il suffit de prendre les paramètres C_1 et C_2 dans $\mathbb{R}$ plutôt que dans $\mathbb{C}$.

$$\mathcal{Y}_h = \{C_1 e^{r_1 x} + C_2 e^{r_2 x} \,:\, C_1,\, C_2 \in \mathbb{R},\, x \in I\}\,, \tag{34.398}$$

Si $b^2 - 4ac < 0$ **:** le polynôme admet deux solutions complexes conjuguées, $r_1 = \alpha + i\beta$ et $r_2 = \alpha - i\beta$; Il faut alors utiliser les formules suivantes

$$\begin{aligned} e^{\alpha+i\beta} &= e^{\alpha}(\cos(\beta) + i\sin(\beta)) \\ e^{\alpha-i\beta} &= e^{\alpha}(\cos(\beta) - i\sin(\beta)). \end{aligned} \tag{34.399}$$

La somme $e^{r_1 x} + e^{r_2 x}$, où x est dans $I \in \mathbb{R}$, vaut

$$e^{(\alpha+i\beta)x} + e^{(\alpha-i\beta)x} = e^{\alpha x}(\cos(\beta x) + i\sin(\beta x)) + e^{\alpha x}(\cos(\beta x) - i\sin(\beta x)) = 2e^{\alpha x}\cos(\beta x)$$

et la différence $e^{r_1 x} - e^{r_2 x}$ vaut

$$e^{(\alpha+i\beta)x} - e^{(\alpha-i\beta)x} = e^{\alpha x}(\cos(\beta x) + i\sin(\beta x)) - e^{\alpha x}(\cos(\beta x) - i\sin(\beta x)) = 2e^{\alpha x}\sin(\beta x).$$

Par ces deux calculs élémentaires nous avons trouvé deux fonctions à valeurs dans $\mathbb{R}$ qui n'ont pas de zéros en commun. Elles sont les génératrices de la famille des solutions réelles de l'équation différentielle (la solution générale)

$$\mathcal{Y}_h = \{e^{\alpha x}\left(C_1\cos(\beta x) + C_2\sin(\beta x)\right) \,:\, C_1,\, C_2 \in \mathbb{R},\, x \in I\}\,, \tag{34.400}$$

Si $b^2 - 4ac = 0$ **:** le polynôme admet une solution réelle double $r = r_1 = r_2$. Dans ce cas la solution générale de l'équation est la famille

$$\mathcal{Y}_h = \{(C_1 + C_2 x)e^{rx} \,:\, C_1,\, C_2 \in \mathbb{R},\, x \in I\}\,. \tag{34.401}$$

Pour justifier cette formule nous observons d'abord que toute fonction $x \mapsto Ce^{rx}$, pour $C \in \mathbb{R}$ est une solution de l'équation différentielle (par construction). Ensuite nous utilisons la méthode de variation de la constante. On trouve rapidement que si une fonction de la forme $x \mapsto C(x)e^{rx}$ est une solution alors $C(x)$ est un polynôme de degré au plus 1, c'est à dire $C(x) = C_1 + C_2 x$ avec C_1 et C_2 dans $\mathbb{R}$.

34.15.2 Linéaires du second ordre à coefficients constants, non homogènes

Nous ne présentons pas une méthode générale pour la résolution de ces équations. Comme dans le cas des équations différentielles linéaires du premier ordre non homogènes, la solution générale de (34.395) est donnée par la somme d'une solution particulière et de la solution générale de l'équation homogène associée. La recherche d'une solution particulière est facilité par le fait que les coefficients de (34.395) sont supposés constants, c'est à dire que a, b et c sont des fonctions constantes. Il faut essayer de deviner la forme d'une solution particulière à partir de la forme du second membre de l'équation, la fonction d. Si d est un polynôme il faut essayer avec un polynôme du même degré, si d est une exponentielle, par exemple $d(x) = e^{5x}$, on pourra essayer avec un multiple de la même fonction exponentielle, dans l'exemple $f(x) = ke^{5x}$, avec k à determiner. Si d est une combinaison linéaire de sinus et cosinus, comme par exemple $12\cos(x) + 2\sin(x)$, on peut essayer avec $k_1\cos(x) + k_2\sin(x)$.

Exemple 34.83
On considère l'équation différentielle

$$y'' + 12y' + 36y = -192e^{2x}, \quad x \in \mathbb{R}. \tag{34.402}$$

Son équation homogène associée est

$$y'' + 12y' + 36y = 0, \tag{34.403}$$

dont le polynôme caractéristique est $r^2 + 12r + 36$. Ce polynôme admet une racine double, qui est -6, par conséquent la solution générale de (34.403) est

$$\mathcal{Y}_h = \left\{(C_1 + C_2x)e^{-6x} \,:\, C_1,\, C_2 \in \mathbb{R},\, x \in \mathbb{R}\right\}.$$

Le membre de droite de (34.402) est une fonction exponentielle, nous allons donc chercher une solution particulière de (34.402) de la forme $f(x) = ke^{2x}$. Par substitution nous trouvons

$$ke^{2x}(4 + 12 \times 2 + 36) = -192e^{2x},$$

ce qui veut dire que k doit être -3.

La solution générale de l'équation (34.402) est donc

$$\mathcal{Y} = \left\{(C_1 + C_2x)e^{-6x} - 3e^{2x} \,:\, C_1,\, C_2 \in \mathbb{R},\, x \in \mathbb{R}\right\}.$$

△

Exemple 34.84
Nous allons résoudre l'équation

$$y'' + 12y' + 36y = 12\cos(x) + 2\sin(x), \quad x \in \mathbb{R}. \tag{34.404}$$

Cette équation a comme homogène associée l'équation (34.403), comme dans l'exemple précédent. Il nous suffit donc de trouver une solution particulière de (34.402).

Nous pouvons essayer avec $f(x) = k_1\cos(x) + k_2\sin(x)$. Par substitution on trouve

$$\begin{aligned} -\left(k_1\cos(x) + k_2\sin(x)\right) + 12\left(-k_1\sin(x) + k_2\cos(x)\right) + 36\left(k_1\cos(x) + k_2\sin(x)\right) \\ = 12\cos(x) + 2\sin(x) \end{aligned}$$

Cette équation doit être satisfaite pour tout valeur de x, en particulier pour $x = 0$ et $x = \pi/2$. Cela revient à considère séparément les coefficients des fonctions sinus et cosinus. Il faut alors que k_1 et k_2 soient solutions du système

$$\begin{cases} -k_1 + 12k_2 + 36k_1 &= 12, \\ -k_2 - 12k_1 + 36k_2 &= 2. \end{cases}$$

On trouve $k_1 = 396/1369$ et $k_2 = 214/1369$, et la solution générale de notre équation est

$$\mathcal{Y} = \left\{ (C_1 + C_2 x)e^{-6x} + \frac{396}{1369}\cos(x) + \frac{214}{1369}\sin(x) \ : \ C_1, C_2 \in \mathbb{R}, \ x \in \mathbb{R} \right\}.$$

△

Exemple 34.85
Nous allons résoudre l'équation

$$y'' + 12y' + 36y = 10x^2 + 3, \quad x \in \mathbb{R}. \tag{34.405}$$

Cette équation a comme homogène associée l'équation (34.403), comme dans l'exemple précédent. Il nous suffit donc de trouver une solution particulière de (34.402).

Nous pouvons essayer avec $f(x) = k_1x^2 + k_2x + k_3$. Par substitution on trouve

$$(2k_1) + 12\left(2k_1x + k_2\right) + 36\left(k_1x^2 + k_2x + k_3\right) = 10x^2 + 3.$$

Pour trouver les bonnes valeurs des coefficients nous devons résoudre le système

$$\begin{cases} 36k_1 & = 10, \\ 24k_1 + 36k_2 & = 0, \\ 2k_1 + 12k_2 + 36k_3 & = 3, \end{cases}$$

ce qui donne $k_1 = 5/18$, $k_2 = -5/27$ et $k_3 = 7/54$. La solution générale de notre équation est

$$\mathcal{Y} = \left\{ (C_1 + C_2 x)e^{-6x} + \frac{5}{18}x^2 - \frac{5}{27}x + \frac{7}{54} \ : \ C_1, C_2 \in \mathbb{R}, \ x \in \mathbb{R} \right\}.$$

△

34.16 Fonction de Green

Soit l'équation différentielle

$$\begin{cases} y''(x) = g(x) & \text{(34.406a)} \\ y(0) = y(1) = 0 & \text{(34.406b)} \end{cases}$$

pour $x \in]0,1[$ et où g est continue sur $]0,1[$.

Nous définissons la fonction de Green

$$G(x,t) = \begin{cases} t(x-1) & \text{si } 0 \leqslant t \leqslant x \leqslant 1 \\ x(t-1) & \text{si } 0 \leqslant x \leqslant t \leqslant 1, \end{cases} \tag{34.407}$$

et nous allons montrer que

$$y(x) = \int_0^1 G(x,t)g(t)dt \tag{34.408}$$

est l'unique solution.

Unicité Si y_1 et y_2 sont des solutions, alors $y_1'' = y_2''$ et donc $y_1(x) = y_2(x) + ax + b$. Les conditions aux bords donnent alors $0 = y_1(0) = y_2(0) + b = b$. D'où $b = 0$. En imposant $y_1(1) = 0$ nous trouvons alors immédiatement $a = 0$, ce qui donne $y_1 = y_2$.

Existence Il est vite vérifié qu'avec (34.408) nous avons $y(0) = y(1) = 0$ parce que $G(0,t) = G(1,t) = 0$ pour tout t. Nous fixons une valeur pour $x \in]0,1[$ et nous découpons l'intégrale :

$$y(x) = \int_0^x G(x,t)g(t)dt + \int_x^1 G(x,t)g(t)dt. \tag{34.409}$$

Pour calculer $y'(x)$, il faut dériver à la fois à travers l'intégrale et dans la borne. Si vous connaissez une formule pour faire cela, c'est bien pour vois. Nous allons faire ça à la main et poser

$$I(x,y) = \int_0^y t(x-1)g(t)dt. \tag{34.410}$$

La dérivation de I par rapport à x se fait en utilisant le théorème 19.34 :

$$\frac{\partial I}{\partial x}(x,y) = \int_0^y tg(t)dt. \tag{34.411}$$

Pour la dérivation par rapport à y, il s'agit du théorème fondamental de l'analyse, plus précisément le lien primitive et intégrale de la proposition 16.211 :

$$\frac{\partial I}{\partial y}(x,y) = y(x-1)g(y). \tag{34.412}$$

Maintenant nous considérons la fonction $\varphi_I(x) = I(x,x)$. Elle satisfait à

$$\varphi_I'(x) = \frac{\partial I}{\partial x}(x,x) + \frac{\partial I}{\partial y}(x,x) = \int_0^x tg(t) + x(x-1)g(x). \tag{34.413}$$

Le même jeu avec $J(x,y) = \int_y^1 x(t-1)g(t)dt$ donne

$$\varphi_J'(x) = \int_0^x fg(t)dt + x(x-1)g(x). \tag{34.414}$$

En remettant les bouts ensemble,

$$y(x) = \int_0^x tg(t)dt + \int_1^x (1-t)g(t)dt. \tag{34.415}$$

Le calcul de la dérivée seconde donne alors

$$y''(x) = xg(x) + (1-x)g(x) = g(x). \tag{34.416}$$

Nous pouvons aussi, sur cette équation, estimer la variation de la solution en termes d'une variation de g. Soit donc une fonction continue δ_g sur $[0,1]$ et $\tilde{g} = g + \delta_g$. Nous considérons l'équation différentielle

$$\begin{cases} \tilde{y}''(x) = \tilde{g}(x) & \text{(34.417a)} \\ \tilde{y}(0) = \tilde{y}(1) = 0. & \text{(34.417b)} \end{cases}$$

Par ce que nous venons de faire, l'unique solution est

$$\tilde{y}(x) = \int_0^1 G(x,t)\tilde{g}(t)dt = \int_0^1 G(x,t)g(t)dt + \int_0^1 G(x,t)\delta_g(t)dt = y(x) + \delta_y(x) \tag{34.418}$$

où δ_y est une fonction continue ainsi définie :

$$\delta_y(x) = \int_0^1 G(x,t)\delta_g(t)dt. \tag{34.419}$$

Supposons que $\|\delta_g\|_\infty = \epsilon$. Alors des majorations donnent

$$|\delta_y(x)| \leqslant \epsilon \int_0^1 |G(x,t)|dt = \epsilon(1-x)\int_0^x tdt + \epsilon x \int_x^1 (1-t)dt = \frac{\epsilon}{2}x(1-x). \tag{34.420}$$

Mais la fonction $x \mapsto x(1-x)$ a son maximum en $x = \frac{1}{2}$, donc nous pouvons donner une majoration indépendante de x :

$$\|\delta_y\|_\infty \leqslant \frac{1}{8}\|\delta_g\|_\infty. \tag{34.421}$$

Notons que la majoration (34.421) en norme uniforme a l'air plus impressionnante, mais la majoration (34.420) donnant une majoration séparée pour chaque x est en réalité plus précise.

Chapitre 35

Équations aux dérivées partielles

35.1 Symbole principal, équation des caractéristiques

Soit l'équation différentielle semi-linéaire d'ordre k

$$\sum_{|\alpha|=k} a_\alpha(x)(\partial^\alpha u)(x) + F\Big(x, u(x), (Du)(x), \dots, (D^{k-1}u)(x)\Big) = 0 \tag{35.1}$$

pour la fonction $u\colon \mathbb{R}^d \to \mathbb{R}$.

Définition 35.1.
*Le **symbole principal** de l'équation (35.1) est l'application*

$$\begin{aligned} \sigma\colon \mathbb{R}^d \times \mathbb{R}^d &\to \mathbb{R} \\ (x,\xi) &\mapsto \sum_{|\alpha|=k} a_\alpha(x)\xi^\alpha \end{aligned} \tag{35.2}$$

où si $\alpha = (\alpha_1, \dots, \alpha_d)$ et $\xi = (\xi_1, \dots, \xi_d)$ alors $\xi^\alpha = \xi_1^{\alpha_1} \dots \xi_d^{\alpha_d}$.

Définition 35.2.
*Les **caractéristiques** de l'équation (35.1) est une surface S de $\mathbb{R}^d$ donné par une équation de la forme $\phi(x) = 0$ où ϕ satisfait à*

$$\sigma\big(x, \nabla\phi(x)\big) = 0 \tag{35.3}$$

et $\nabla\phi(x) \neq 0$ pour tout $x \in S$.

35.2 Méthode des caractéristiques pour l'ordre 1

Nous[394, 395] voulons étudier l'équation d'ordre 1

$$a(x,y)\frac{\partial u}{\partial x}(x,y) + b(x,y)\frac{\partial u}{\partial y}(x,y) + c(x,y)u(x,y) = f(x,y) \tag{35.4}$$

Le champ de vecteurs associé à cette équation est

$$v = \begin{pmatrix} a \\ b \end{pmatrix}, \tag{35.5}$$

et l'équation peut être écrite sous la forme

$$(v \cdot \nabla) + cu = f. \tag{35.6}$$

Définition 35.3.
*Le **flot** de ce champ de vecteurs sont les courbes paramétriques $\gamma(t) = \big(x(t), y(t)\big)$ vérifiant $\gamma' t(t) = v\big(\gamma(t)\big)$.*

Les équation du flot pour l'équation (35.4) sont

$$\begin{cases} x'(t) = a\big(x(y), y(t)\big) & \text{(35.7a)} \\ y'(t) = b\big(x(y), y(t)\big). & \text{(35.7b)} \end{cases}$$

Ce sont des équations différentielles ordinaires. Un système de deux équations couplées du premier ordre.

Quel est l'intérêt du flot ? Nous allons voir que sur la ligne $t \mapsto \gamma(t)$, la fonction u est constante. Or des solutions γ au système (35.7), il y en aura plusieurs : une pour chaque valeur des constantes d'intégration. Pour peu que ces lignes recouvrent tout le plan, nous pourrons résoudre l'équation de départ ligne par ligne.

Nous posons

$$\tilde{u}(t) = u\big(x(t), y(t)\big) \quad \text{(35.8a)}$$
$$\tilde{c}(t) = c\big(x(t), y(t)\big) \quad \text{(35.8b)}$$
$$\tilde{f}(t) = f\big(x(t), y(t)\big). \quad \text{(35.8c)}$$

La fonction $\tilde{u}$ est une fonction $\mathbb{R} \to \mathbb{R}$ normale qui se dérive normalement, en suivant la règle de dérivation des fonctions composées :

$$\begin{aligned} \tilde{u}'(t) &= \frac{\partial u}{\partial x}\big(x(t), y(t)\big)x'(t) + \frac{\partial u}{\partial y}\big(x(t), y(t)\big)y'(t) && \text{(35.9a)} \\ &= a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} && \text{(35.9b)} \\ &= f\big(x(t), y(t)\big) - c\big(x(t), y(t)\big)u\big(x(t), y(t)\big) && \text{(35.9c)} \\ &= \tilde{f}(t) - \tilde{c}(t)\tilde{u}(t). && \text{(35.9d)} \end{aligned}$$

Nous avons pour $\tilde{u}$ l'équation différentielle ordinaire

$$\tilde{u}' + \tilde{c}\tilde{u} = \tilde{f} \quad \text{(35.10)}$$

qui est résolue par la proposition 34.10.

35.2.1 Un exemple complet un peu minimal

Nous considérons l'équation différentielle[394]

$$\frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} - (x - y)u = 0. \quad \text{(35.11)}$$

Et nous allons la résoudre.

Les équations du flot, sont simples parce que les coefficients sont des constantes : $x'(t) = 1$, $y'(t) = -1$. Donc

$$x(t) = t + C_1 \quad \text{(35.12a)}$$
$$y(t) = -t + C_2. \quad \text{(35.12b)}$$

A priori nous avons une caractéristique pour chaque choix de (C_1, C_2) et nous espérons que le tout recouvre le plan $\mathbb{R}^2$. En fait seule une des deux constantes doit être laissée libre, l'autre consiste seulement en décaler le paramètre t. Nous posons donc $C_1 = 0$ et nous considérons les courbes caractéristiques

$$\gamma_C(t) = \begin{pmatrix} t \\ -t + C \end{pmatrix}. \quad \text{(35.13)}$$

Ces courbes recouvrent bien tout le plan. Pour savoir les valeurs de u sur la courbe γ_C, nous devons résoudre l'équation différentielle ordinaire

$$\tilde{u}_C' + \tilde{c}\tilde{u}_C = \tilde{f}, \quad \text{(35.14)}$$

en sachant que $\tilde{c}(t) = c\big(x(t), y(t)\big) = -\big(x(t) - y(t)\big) = 2t - C$. Cela se fait en suivant la méthode décrite dans l'exemple 34.8 et résumée dans la proposition 34.10.

En termes de notations, $\tilde{u}_C(t) = u\big(\gamma_C(t)\big)$. Récrivons l'équation :

$$\tilde{u}'(t) - (2t - C)\tilde{u}(t) = 0. \tag{35.15}$$

La méthode pour la résoudre est de mettre les $\tilde{u}$ d'un côté et les t de l'autre :

$$\frac{\tilde{u}'}{u} = 2t - C. \tag{35.16}$$

En intégrant par rapport à t des deux côtés,

$$\ln(\tilde{u}) = t^2 - Ct + K_C, \tag{35.17}$$

c'est à dire (avec redéfinition de K_C)

$$\tilde{u}(t) = K_C e^{t^2 - Ct} \tag{35.18}$$

ou encore

$$u\big(\gamma_C(t)\big) = K_C e^{t^2 - Ct} \tag{35.19}$$

où C est le paramètre que nous déterminons en sachant sur quelle caractéristique se trouve le point (x, y) où nous voulons calculer $u(x, y)$ et K est une constante (strictement positive parce que si vous avez suivi le mouvement, c'est une exponentielle) qui doit être déterminée par les conditions initiales. Dès que K est fixé pour un des points de la courbe γ_C, alors il est fixé pour tous les points.

Ce que nous avons obtenu est qu'il existe un K_C tel que pour tout t nous avons

$$u\big(\gamma_C(t)\big) = K_C e^{t^2 - Ct}. \tag{35.20}$$

Soit donc un point $(x_0, y_0) \in \mathbb{R}^2$. Nous devons d'abord déterminer où ce point se trouve par rapport aux caractéristiques, c'est à dire quelle est la valeur de C pour laquelle (x_0, y_0) est sur la courbe γ_C, et ensuite déterminer pour quelle valeur de t nous aurons $\gamma_C(t) = (x_0, y_0)$. À résoudre :

$$\gamma_C(t_0) = \begin{pmatrix} t_0 \\ -t_0 + C \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}. \tag{35.21}$$

Donc $t_0 = x_0$ et $C = x_0 + y_0$. En reprenant (35.19) nous avons

$$u\big(\gamma_C(t_0)\big) = K e^{x_0^2 - (x_0 + y_0)x_0} = K e^{-x_0 y_0}. \tag{35.22}$$

Pour peu que des conditions soient donnée sur chaque caractéristique, nous pouvons déterminer K. Attention : ce K est une constante d'intégration de l'équation différentielle ordinaire pour $\tilde{u}$. Donc elle n'est valable que sur chaque caractéristique séparément. Cela n'est donc pas du tout une constante sur $\mathbb{R}^2$.

Nous pouvons maintenant écrire la solution générale de l'équation de départ. L'équation cartésienne de la courbe γ_C est

$$x + y = C. \tag{35.23}$$

Donc K est une fonction de $x + y$, pas de x et y séparément. Cela est important à comprendre. A priori nous avons

$$u(x, y) = K(x, y)e^{-xy} \tag{35.24}$$

où $K(x, y)$ est constante sur la courbe γ_C contenant (x, y). Nous avons

— si $x_1 + y_1 = x_2 + y_2$,
— alors il existe C tel que (x_1, y_1) et (x_2, y_2) sont sur γ_C,
— alors $K(x_1, y_2) = K(x_2, y_2)$.

Donc il existe une fonction $\mathbb{R} \to \mathbb{R}$ telle que $K(x, y) = f(x + y)$.

Au final, la solution générale de l'équation est

$$u(x, y) = f(x + y)e^{-xy} \tag{35.25}$$

où f est une fonction à déterminer par les conditions initiales qui peuvent être données. Typiquement nous espérons que les conditions imposent une et une seule valeur de u sur chacune des courbes γ_C.

35.2.2 Un théorème d'existence et d'unicité

La méthode des caractéristiques donne essentiellement une preuve de l'unicité des solutions aux équations de transport, et une méthode pour construire cette solution. En effet, la procédure suivante permet de construire $u(x_0, y_0)$.

— Trouver la caractéristique passant par le point (x_0, y_0)
— Calculer en quel point elle passe par une condition initiale donnée.
— Attribuer à $u(x_0, y_0)$ la valeur trouvée sur la caractéristique là où elle passe par une condition initiale.

Rien ne permet a priori de savoir que cette procédure construit effectivement une solution. En particulier, comment calculer $\partial_x u$? Le quotient différentiel serait

$$\frac{\partial u}{\partial x}(x,y) = \lim_{\epsilon\to 0} \frac{u(x+\epsilon,y)-u(x,y)}{\epsilon}, \tag{35.26}$$

mais la caractéristique donnant la valeur de $u(x+\epsilon, y)$ est différente pour chaque ϵ. Rien a priori ne permet d'affirmer que le calcul soit simple, ni qu'il arrive à une solution du problème donné.

D'où la nécessité d'avoir un résultat un peu rigoureux donnant des conditions sous lesquelles les choses vont bien.

Proposition 35.4 (Équation de transport à coefficients variables[395, 386])**.**
Soit une fonction $c\colon \mathbb{R}^2 \to \mathbb{R}$ de classe C^2 en ses deux variables et uniformément Lipschitziennes en sa première variable[1] *et $g \in C^1(\mathbb{R})$. Alors l'équation aux dérivées partielles de premier ordre*

$$\begin{cases} \dfrac{\partial u}{\partial x}(x,t) + c(x,t)\dfrac{\partial u}{\partial t}(x,y) = 0 & \text{(35.27a)} \\ u(x,0) = h(y) & \text{(35.27b)} \end{cases}$$

admet une unique solution de classe C^1.

Cette solution est construite de la façon suivante[2]*. D'abord nous considérons la solution X au problème*

$$\begin{cases} \dfrac{\partial X}{\partial s}(s;x,t) = c\big(X(s;x,t),s\big) & \text{(35.28a)} \\ X(t;x,t) = x, & \text{(35.28b)} \end{cases}$$

et ensuite le problème (35.27) a pour unique solution

$$u(x,t) = h\big(X(0;x,t)\big). \tag{35.29}$$

Démonstration. Nous commençons par étudier l'existence et l'unicité de la fonction X définie par le problème 35.28. La fonction c ici est dans les hypothèses de la fonction f du théorème de Cauchy-Lipschitz global 19.50. D'où l'existence et l'unicité de la fonction $s \mapsto X(s;x,t)$ sur $\mathbb{R}$ pour chaque (x,y) donné[3].

Le lemme 34.32 nous dit que X est de classe C^2 en (s,x,t). Donc nous pourrons dériver et permuter les dérivées autant que nous voudrons (sans exagérer : ordre 2 au maximum).

Unicité Nous montrons que u doit être constante sur le chemin

$$\gamma_{(x,t)}(s) = \begin{pmatrix} X(s;x,t) \\ s \end{pmatrix}. \tag{35.30}$$

En effet, en posant

$$\varphi(s) = u\big(\gamma(s)\big) = u\big(X(s;x,t),s\big), \tag{35.31}$$

1. Dans [386], on ne demande que continue puis uniformément Lipschitz. Moi je crois que ce n'est pas assez pour assurer la dérivabilité de X par rapport à x, et encore moins pour permuter les dérivées dans $\partial^2_{tx}Y$.
2. Le fait que la construction ait un sens fait partie des choses à prouver
3. Avec l'énoncé tel que donné dans [395], il faut utiliser la technique de 34.20 pour l'existence globale, parce que la fonction b là-dedans n'est pas dans les mêmes hypothèses.

et en dérivant nous obtenons

$$\varphi'(s) = \frac{\partial u}{\partial t}\big(X(s;x,t),s\big)\frac{\partial X}{\partial s}(s;x,t) + \frac{\partial u}{\partial t}\big(X(s;x,t)\big) = 0. \tag{35.32}$$

Par conséquent la valeur commune de tous les $u\big(\gamma_{(x,t)}(s)\big)$ doit être celle en $\gamma_{(x,t)}(0) = h(X(0;x,t))$

Cela prouve l'unicité parce que la valeur de u est fixée en tout point. Nous devons encore vérifier que la fonction u ainsi construite est bien une solution du problème. C'est l'objet de la partie « existence » de la preuve.

Existence Même l'existence est divisée en plusieurs étapes.

Mise en place Nous prouvons que la fonction u donné par (35.29) est une solution du problème. Nous avons :

$$\frac{\partial u}{\partial t}(x,t) = h'\big(X(0;x,t)\big)\frac{\partial X}{\partial t}(0;x,t) \tag{35.33}$$

et

$$\frac{\partial u}{\partial x}(x,t) = h'\big(X(0;x,t)\big)\frac{\partial X}{\partial x}(0;x,t), \tag{35.34}$$

de sorte qu'en posant

$$g(s;x,t) = \frac{\partial X}{\partial t}(s;x,t) + c(x,t)\frac{\partial X}{\partial x}(s;x,t) \tag{35.35}$$

nous avons

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = h'\big(X(0;x,t)\big)g(0;x,t). \tag{35.36}$$

Une équation différentielle pour g Nous allons prouver que

$$\frac{\partial g}{\partial s}(s;x,t) = \alpha_{(x,t)}(s)g(s;x,t) \tag{35.37}$$

avec [4]

$$\alpha_{(x,t)}(s) = \frac{\partial c}{\partial x}\big(X(s;x,t),s\big). \tag{35.38}$$

D'abord nous avons

$$\frac{\partial g}{\partial s}(s;x,t) = \frac{\partial^2 X}{\partial s\partial t}(s;x,t) + c(x,t)\frac{\partial^2 X}{\partial s\partial x}(s;x,t). \tag{35.39}$$

Nous permutons les dérivées et nous tenons compte de (35.28) :

$$\frac{\partial g}{\partial s}(s;x,t) = \frac{\partial}{\partial t}\Big(c\big(X(s;x,t),s\big)\Big) + c(x,t)\frac{\partial}{\partial x}\Big(c\big(X(s;x,t),s\big)\Big). \tag{35.40}$$

Nous dérivons maintenant plus en profondeur. D'une part

$$\frac{\partial}{\partial t}\Big(c\big(X(s;x,t),s\big)\Big) = \frac{\partial c}{\partial x}\Big(X(s;x,t),s\Big)\frac{\partial X}{\partial t}(s;x,t) \tag{35.41}$$

et d'autre part,

$$\frac{\partial}{\partial x}\Big(c\big(X(s;x,t),s\big)\Big) = \frac{\partial c}{\partial x}\Big(X(s;x,t),s\Big)\frac{\partial X}{\partial x}(s;x,t), \tag{35.42}$$

de telle sorte que

$$\frac{\partial g}{\partial s}(s;x,t) = \frac{\partial c}{\partial x}\big(X(s;x,t),s\big)\Big[\frac{\partial X}{\partial t}(s;x,t) + c(x,t)\frac{\partial X}{\partial x}(s;x,t)\Big] \tag{35.43a}$$

$$= \alpha_{(x,t)}(s)g(s;x,t). \tag{35.43b}$$

4. La ligne suivante est une de celles qui me font penser qu'il manque des hypothèses dans [386]. Il faut bien pouvoir dériver c.

Une condition initiale pour g Nous montrons maintenant que $g(t;x,t)=0$. La condition initiale pour X est $X(t;x,t)=x$ pour tout $t,x\in\mathbb{R}$. Nous dérivons cette dernière par rapport à x et à t :

$$\frac{\partial X}{\partial x}(t;x,t)=1 \tag{35.44a}$$

$$\frac{\partial X}{\partial s}(t;x,t)+\frac{\partial X}{\partial t}(t;x,t)=0. \tag{35.44b}$$

Mais $\partial_s X(t;x,t)=c\big(X(t;x,t),t\big)$, donc la relation (35.44b) donne

$$\frac{\partial X}{\partial t}(t;x,t)=-c\big(X(t;x,t),t\big)=-c(x,t) \tag{35.45}$$

où nous avons tenu compte du fait que $X(t;x,t)=x$.
Voyons à présent ce que (35.44a) et (35.45) donnent pour $g(t;x,t)$:

$$g(t;x,t)=\frac{\partial X}{\partial t}(t;x,t)+c(x,t)\frac{\partial X}{\partial x}(t;x,t)=-c(x,t)+c(x,t)=0. \tag{35.46}$$

Conclusion pour g La fonction g vérifie l'équation différentielle

$$\begin{cases}\dfrac{\partial g}{\partial s}(s)=\alpha(s)g(s) & \text{(35.47a)}\\ g(t)=0. & \text{(35.47b)}\end{cases}$$

Bien entendu, $g(s)=0$ est une solution. Mais la solution à ce système est unique par Cauchy-Lipschitz 19.49. Ici nous utilisons le fait que

$$(s,y)\mapsto\alpha(s)y \tag{35.48}$$

est continue. C'est à dire entre autres que

$$s\mapsto\frac{\partial c}{\partial x}\big(X(s;x,t),s\big) \tag{35.49}$$

doit être continue. C'est le cas parce que c est de classe C^1 en sa première variable.

□

Les hypothèses de la proposition 35.4 sont loin d'être optimales. Voici un exemple dans lequel c n'est même pas dérivable par rapport à t et qui se passe très bien quand même.

Exemple 35.5
Soit l'équation différentielle

$$\begin{cases}\dfrac{\partial u}{\partial t}(x,t)+|t-1|\dfrac{\partial u}{\partial x}(x,t)=0 & \text{(35.50a)}\\ u(x,0)=h(x) & \text{(35.50b)}\end{cases}$$

où h est une fonction bien régulière ; mettons C^p. En suivant la méthode de la proposition nous devrions poser l'équation différentielle

$$\begin{cases}\dfrac{\partial X}{\partial s}(s;x,t)=|s-1| & \text{(35.51a)}\\ X(t;x,t)=x. & \text{(35.51b)}\end{cases}$$

Cela est la caractéristique passant par (x,t). Cependant il sera plus simple de chercher les caractéristiques en demandant qu'elles passent par $(x_0,0)$. Nous allons donc plutôt résoudre pour X_{x_0} l'équation différentielle

$$\begin{cases}\dfrac{\partial X}{\partial s}(s)=|s-1| & \text{(35.52a)}\\ X_{x_0}(0)=x_0 & \text{(35.52b)}\end{cases}$$

et les courbes caractéristiques seront les chemins

$$\gamma_{x_0}(s) = \begin{pmatrix} X_{x_0}(s) \\ s \end{pmatrix}. \tag{35.53}$$

La résolution donne d'abord

$$X_{x_0}(s) = \begin{cases} -\frac{s^2}{2} + s + K_1 & \text{si } s < 1 \\ \frac{s^2}{2} - s + K_2 & \text{si } s > 1. \end{cases} \tag{35.54}$$

Vu que la condition initiale est donnée pour $s = 0$, nous fixons K_1 pour la condition initiale et K_2 pour la continuité :

$$X_{x_0}(s) = \begin{cases} -\frac{s^2}{2} + s + x_1 & \text{si } s < 1 \\ \frac{1}{2} + x_0 & \text{si } s = 1 \\ \frac{s^2}{2} - s + 1 + x_0 & \text{si } s > 1. \end{cases} \tag{35.55}$$

Soit $(x, t) \in \mathbb{R}^2$. Quelle caractéristique passe par là ? Nous allons déterminer la fonction $x_0(x, t)$ qui donne le x_0 tel que la caractéristique γ_{x_0} passe par (x, t). Nous devons résoudre

$$\begin{pmatrix} X_{x_0}(s) \\ s \end{pmatrix} = \begin{pmatrix} x \\ t \end{pmatrix}. \tag{35.56}$$

Directement : $s = t$. Et ensuite $X_{x_0}(t) = x$. Nous avons

$$x_0(x, t) = \begin{cases} x + \frac{t^2}{2} - t & \text{si } t < 1 \\ x - \frac{1}{2} & \text{si } t = 1 \\ x - \frac{t^2}{2} + t - 1 & \text{si } t > 1. \end{cases} \tag{35.57}$$

Le truc presque étonnant est que x_0 est de classe C^1. En effet le calcul de

$$\frac{\partial x_0}{\partial t}(1) = \lim_{\epsilon \to 0} \frac{x_0(x, 1 + \epsilon) - x_0(x, 1)}{\epsilon} \tag{35.58}$$

se fait en séparant les limites $\epsilon \to 0^+$ et $\epsilon \to 0^-$. Le résultat est que $\partial_t x_0(1) = 0$. Nous avons donc

$$\frac{\partial x_0}{\partial t}(t) = \begin{cases} t - 1 & \text{si } t < 1 \\ 0 & \text{si } t = 1 \\ -t + 1 & \text{si } t > 1. \end{cases} \tag{35.59}$$

Cela étant continu, la fonction x_0 est de classe C^1 en t, et la dérivée en x étant toujours 1, elle est de classe C^1.

En ce qui concerne la solution de l'équation de départ,

$$u(x, t) = h\big(X_{x_0(x,t)}(0)\big) = h\big(x_0(x, t)\big). \tag{35.60}$$

Pourvu que h soit assez régulière, la fonction u est facilement de classe C^1. △

35.3 Méthode des caractéristique pour l'ordre 2

35.3.1 Principe général

Soit l'opérateur différentiel agissant sur $C^2(\mathbb{R}^2)$:

$$D = a(x, y)\frac{\partial^2}{\partial x^2} + b(x, y)\frac{\partial^2}{\partial x \partial y} + c(x, y)\frac{\partial^2}{\partial y^2}. \tag{35.61}$$

Nous voulons résoudre des équations du type $Du = 0$ pour $u\colon \mathbb{R}^2 \to \mathbb{R}$.

Pour commencer[394], et c'est le point crucial, nous voyons D comme un polynôme en ∂_x et ∂_y et nous le factorisons : si

$$aX^2 + bXY + cY = (\alpha X + \beta Y)(\gamma X + \delta Y) \tag{35.62}$$

alors nous avons

$$D = \left(\alpha\frac{\partial}{\partial x} + \beta\frac{\partial}{\partial y}\right)\left(\gamma\frac{\partial}{\partial x} + \delta\frac{\partial y}{\partial}\right) + \text{ termes d'ordre inférieurs.} \tag{35.63}$$

Les « termes d'ordre inférieurs » sont ceux de la forme $\alpha(x,y)\frac{\partial\delta}{\partial x}\frac{\partial}{\partial y}$.

L'astuce est de poser

$$v = (\gamma\partial_x + \delta\partial_y)u, \tag{35.64}$$

et de résoudre le système

$$\begin{cases} (\alpha\partial_x + \beta\partial_y)v = 0 & \text{(35.65a)} \\ (\gamma\partial_x + \delta\partial_y)u = v. & \text{(35.65b)} \end{cases}$$

Cela sont deux équations différentielles du premier ordre pour lesquelles nous avons déjà des techniques décrites en la section 35.2.

Afin que les fonctions α, β, γ et δ soient réelles, il faut que $b^2 - 4ac \geqslant 0$. Sachant que $a = \alpha\gamma$, $b = \alpha\delta + \beta\gamma$ et $c = \beta\delta$ cette condition sur a, b et c donne

$$(\alpha\delta + \beta\gamma)^2 - 4\alpha\gamma\beta\delta \geqslant 0. \tag{35.66}$$

Cela revient à

$$(\alpha\delta - \beta\gamma)^2 \geqslant 0. \tag{35.67}$$

Nous supposons à présent que l'inégalité soit stricte (cas hyperbolique). Nous avons en particulier que

$$\alpha\delta - \beta\gamma \neq 0. \tag{35.68}$$

Cette condition implique que les équations

$$\frac{dx}{dt} = \alpha(x,y) \quad \frac{dy}{dt} = \beta(x,y) \tag{35.69}$$

sont indépendantes des équations

$$\frac{dx}{dt} = \gamma(x,y) \quad \frac{dy}{dt} = \delta(x,y) \tag{35.70}$$

Ce sont les équations caractéristiques des équations (35.65).

35.3.2 Exemple : l'équation d'onde

Nous considérons l'équation aux dérivées partielles

$$\frac{\partial^2 u}{\partial t^2} - c^2\frac{\partial^2 u}{\partial x^2} = 0 \tag{35.71}$$

où c est une constante réelle. Nous en cherchons des solutions de classe C^2.

L'opérateur différentiel est donné par le polynôme $P(T,X) = T^2 - c^2X^2$ qui se factorise en

$$P = (T + cX)(T - cX), \tag{35.72}$$

c'est à dire que nous pouvons récrire l'équation des ondes sous la forme

$$(\partial_t + c\partial_x)(\partial_t - c\partial_x)u = 0. \tag{35.73}$$

Nous posons donc $v = (\partial_t - c\partial_x)u$ et nous avons le système[385]

$$\begin{cases} (\partial_t + c\partial_x)v = 0 & \text{(35.74a)} \\ (\partial_t - c\partial_x)u = v. & \text{(35.74b)} \end{cases}$$

La méthode des caractéristique est efficace pour résoudre la première, et pour trouver la solution générale de l'homogène associée à la seconde.

Nous nous lançons dans la résolution de (35.74a). Le flot est $v = \begin{pmatrix} 1 \\ c \end{pmatrix}$, et nous cherchons ses courbes intégrales sous la forme $\varphi(t) = \big(t, x(t)\big)$. Immédiatement, $x'(t) = c$, ce qui donne

$$\gamma_C(t) = \begin{pmatrix} t \\ ct + C \end{pmatrix}. \tag{35.75}$$

Cela donne une caractéristique pour chaque valeur de C. En posant $\tilde{v}_C(t) = v(t, ct + C)$ nous avons

$$\tilde{v}'_C(t) = \frac{\partial v}{\partial t}\big(\gamma_C(t)\big) + c\frac{\partial v}{\partial t}\big(\gamma_C(t)\big) = 0. \tag{35.76}$$

Donc $\tilde{v}_C$ est une fonction constante. Donc u est constant sur la courbe γ_C dont l'équation cartésienne est $x - ct = C$. Cela implique que

$$v(t, x) = f(x - ct) \tag{35.77}$$

où f est une fonction de classe C^1. En effet si (t_1, x_1) et (t_2, x_2) vérifient $x_1 - ct_1 = x_2 - ct_2$ alors $v(t_1, x_2) = v(t_2, x_2)$. Le fait que f soit C^1 est une demande que u soit au final dans C^2.

Nous devons maintenant résoudre l'équation (35.74b)

$$(\partial_t - c\partial_x)u = v. \tag{35.78}$$

Nous allons agir conformément à la stratégie expliquée par le lemme 34.6. Nous devons résoudre $Du = v$ avec

$$\begin{aligned} D\colon C^2(\mathbb{R}) &\to D(C^2(\mathbb{R})) \\ u &\mapsto (\partial_t - c\partial_x)u. \end{aligned} \tag{35.79}$$

Par la même méthode des caractéristiques que celle déjà menée plus haut nous trouvons $\ker(D)$ comme solution générale de $(\partial_t - c\partial_x)u_G = 0$. C'est à dire

$$u_G = g(x + ct) \tag{35.80}$$

où g est une fonction quelconque de classe C^2.

Il nous faut maintenant une solution particulière de

$$(\partial_t - c\partial_x)u_P(t, x) = f(x - ct). \tag{35.81}$$

Si F est une primitive de f alors

$$u_P(t, x) = -\frac{1}{2c}F(x - ct) \tag{35.82}$$

fonctionne. Vu que f est quelconque dans $C^1(\mathbb{R})$, la fonction F est un élément quelconque de $C^2(\mathbb{R})$. Au final, la solution générale de l'équation des ondes est

$$u(t, x) = g_1(x + ct) + g_2(x - ct) \tag{35.83}$$

où g_1 et g_2 sont des éléments de $C^2(\mathbb{R})$.

35.4 Classification des équations du second ordre

Soit une équation générale d'ordre 2 sur $\mathbb{R}^2$:

$$a\frac{\partial^2 u}{\partial x^2}+b\frac{\partial^2 u}{\partial x\partial y}+c\frac{\partial^2 u}{\partial y^2}+d\frac{\partial u}{\partial x}+e\frac{\partial u}{\partial y}+\beta u=f. \tag{35.84}$$

En ce qui concerne son symbole principal nous avons

$$\sigma(x,y,\xi_1,\xi_2)=a(x,y)\xi_1^2+b(x,y)\xi_1\xi_2+c(x,y)\xi_2^2, \tag{35.85}$$

ce qui donne l'équation des caractéristiques

$$a\left(\frac{\partial\phi}{\partial x}\right)^2+b\frac{\partial\phi}{\partial x}\frac{\partial\phi}{\partial y}+c\left(\frac{\partial\phi}{\partial y}\right)^2=0. \tag{35.86}$$

Si nous nous posons sur un point (x_0,y_0) tel que $\nabla\phi(x_0,y_0)=0$ et $\partial_x\phi(x_0,y_0)\neq 0$ alors, via le théorème de la fonction implicite [5], la condition $\phi(x,y)=0$ définit une fonction $y\mapsto x(y)$ vérifiant

$$\phi\big(x(y),y\big)=0 \tag{35.87}$$

pour tout y dans un voisinage de y_0.

Nous pouvons obtenir une équation différentielle ordinaire pour x de la façon suivante. D'abord nous posons $\varphi(y)=\phi\big(x(y),y\big)$ et ensuite nous calculons la dérivée de φ (qui est nulle par construction) :

$$0=\varphi'(y)=\frac{\partial\phi}{\partial x}x'+\frac{\partial\phi}{\partial y}. \tag{35.88}$$

Nous pouvons donc remplacer $\partial_y\phi$ par $x'\partial_x\phi$ dans l'équation des caractéristiques (35.86) :

$$a\left(\frac{\partial\phi}{\partial x}\right)^2+bx'\left(\frac{\partial\phi}{\partial x}\right)^2+c\left(\frac{\partial\phi}{\partial x}\right)^2(x')^2=0. \tag{35.89}$$

Vu que nous avons supposé $(\partial_x\phi)\neq 0$ sur un voisinage de (x_0,y_0) nous pouvons simplifier par $(\partial_x\phi)^2$ et avoir l'équation différentielle ordinaire

$$a\big(x(y),y\big)+b\big(x(y),y\big)x'(y)+c\big(x(y),y\big)x'(y)^2=0. \tag{35.90}$$

Notons que, conformément à ce que raconte le théorème des fonctions implicites, nous avons pris un voisinage de y_0 suffisamment petit pour que $x(y)$ reste dans un voisinage de x_0. Ce voisinage étant, nous pouvons le restreindre pour nous assurer du signe de a, b et c. Cela est évidemment très théorique parce que le théorème de la fonction implicite parle de l'existence de voisinages, mais pas de façon de les construire.

Nous donnons la classification suivante.

Définition 35.6.
*Si $b^2-4ac<0$ alors l'équation est **elliptique**.*
*Si $b^2-4ac>0$ alors l'équation est **hyperbolique**.*
*Si $b^2-4ac=0$ alors l'équation est **parabolique**.*

Exemple 35.7
Un exemple d'équation parabolique est l'équation de la chaleur

$$\frac{\partial u}{\partial t}-\alpha\frac{\partial^2 u}{\partial x^2}=0 \tag{35.91}$$

où $\alpha>0$ est une constante. Cette équation est avec $a=c=0$, donc elle est parabolique. △

5. Théorème 19.58.

Une équation aux dérivées partielles peut changer de nature selon le point.

Exemple 35.8
Soit l'équation

$$\frac{\partial^2 u}{\partial x^2} - (x^2 - y^2)\frac{\partial^2 u}{\partial y^2} = 0. \tag{35.92}$$

Son $b^2 - 4ac$ vaut $4(x^2 - y^2)$. Elle peut donc être hyperbolique, parabolique ou elliptique selon le point où l'on se trouve. △

35.4.1 Problème au limite

Dans les définitions qui suivent nous considérons un ouvert $\Omega \subset \mathbb{R}^d$ assez régulier et possédant en particulier un vecteur normal extérieur $n(x)$ pour tout point $x \in \partial\Omega$.

Définition 35.9.
Un problème aux limite ***de Dirichlet*** *est d'imposer la condition*

$$u(x) = g(x) \tag{35.93}$$

pour tout $x \in \Gamma \subset \partial\Omega$. C'est à dire imposer la valeur de u sur une partie du bord du domaine.

Définition 35.10.
Un problème aux limites de ***Von Neumann*** *est d'imposer*

$$\frac{\partial u}{\partial n} \cdot x = g(x) \tag{35.94}$$

pour tout $x \in \Gamma \subset \partial\Omega$. C'est à dire imposer les valeurs de la dérivée normale de u sur une partie du bord.

Exemple 35.11
Lorsqu'on veut imposer un flux de chaleur aux bords d'un domaine pour l'équation de la chaleur, il s'agit de poser des conditions de type Von Neumann. △

Définition 35.12.
Soit un domaine Ω de $\mathbb{R}^d$ et un opérateur différentiel L sur une partie de $\mathrm{Fun}(\Omega)$*. Soit une fonction g sur $\partial\Omega$. Un problème* ***aux limites stationnaires*** *est un problème du type : trouver u définie sur Ω telle que*

$$\begin{cases} L(u) = f & \text{(35.95a)} \\ u|_{\partial\Omega} = g. & \text{(35.95b)} \end{cases}$$

Définition 35.13.
Un problème ***aux limites d'évolution*** *est du type : trouver $u \in \mathrm{Fun}\left(]0,\infty[\times \Omega\right)$ tel que*

$$\begin{cases} \frac{\partial^m u}{\partial t^m} + L(u) = f & \text{sur }]0,\infty[\times \Omega \\ u(t,.) = g(t,.) & \text{sur }]0,\infty[\times \partial\Omega \\ u(0,.) = u_0 & \text{sur } \Omega \end{cases} \tag{35.96}$$

où u_0 est une fonction sur Ω.

L'opérateur L ne doit pas opérer sur la partie « t » de u.

Définition 35.14 (Problème bien posé au sens de Hadamard)**.**
Un problème aux limites est ***bien posé au sens de Hadamard*** *si*

(1) Il admet une unique solution.

(2) La solution dépend de façon continue en les données du problèmes.

La continuité est au sens des normes sur les fonctions sur $\partial\Omega$ et sur Ω en ce qui concerne les fonctions « données » du problème et des normes pour les fonctions sur Ω ou $\bar{\Omega}$ en ce qui concerne la solution.

Exemple 35.15(Un problème de Dirichlet bien posé)
Trouver la fonction u définie sur $\Omega = [0,1]^2$ telle que

$$\begin{cases} -\Delta u = 0 & \text{sur } \Omega \\ u = g & \text{sur } \partial\Omega. \end{cases} \tag{35.97}$$

△

35.5 Principe du maximum

Lemme 35.16 ([304]).
Soit un ouvert borné $\Omega \subset \mathbb{R}^n$. Soit une matrice symétrique strictement définie positive A telle que $A_{ij} \in C^O(\bar{\Omega})$ pour laquelle il existe $\lambda > 0$ minorant toutes les valeurs propres de toutes les matrices $A(x)$ pour $x \in \Omega$ [6].

Nous posons $L' = -\sum_{ij} A_{ij}\partial_{ij}$. Si $u \in C^2(\Omega)$ atteint un minimum local en $x_0 \in \Omega$, alors

$$(L'u)(x_0) \leqslant 0. \tag{35.98}$$

Démonstration. Nous allons bien entendu diagonaliser A. Si T est une matrice nous avons, en posant $u(x) = v(Tx)$:

$$\frac{\partial u}{\partial x_i}(x) = \sum_k \frac{\partial v}{\partial x_k}(Tx)T_{ki} \tag{35.99}$$

et

$$\frac{\partial^2 u}{\partial x_j \partial x_i}(x) = \sum_{kl} T_{ki}T_{lj}\frac{\partial^2 v}{\partial x_l \partial x_k}(Tx). \tag{35.100}$$

Si T est en particulier une matrice orthogonale diagonalisant A (théorème 11.174) nous avons $T^{-1} = T^t$ et

$$\sum_{ij} T_{ki}A_{ij}T_{jl}^{-1} = D_{kl} = \lambda_k\delta_{kl} \tag{35.101}$$

où les λ_k sont les valeurs propres de A. Notons que partout ici, tout est fonction de x sur Ω : tant A que T que les λ_k. Avec tous ces résultats nous calculons vite que

$$(Lu)(x_0) = -\sum_k \lambda_k(\partial_k^2 v)(Tx_0). \tag{35.102}$$

Si $x_0 \in \Omega$ est un minimum local de u, alors l'application v a un minimum local en $T^{-1}x_0$. Et donc

$$\frac{\partial^2 v}{\partial x_k^2}(Tx_0) \geqslant 0. \tag{35.103}$$

Du coup,

$$(Lu)(x_0) \leqslant 0. \tag{35.104}$$

□

6. Cela est plus que dire que toutes les $A(x)$ son symétriques strictement définie positive.

Lemme 35.17 ([304]).
Soit un ouvert borné $\Omega \subset \mathbb{R}^n$. Soit une matrice symétrique strictement définie positive A telle que $A_{ij} \in C^O(\bar{\Omega})$ pour laquelle il existe $\lambda > 0$ minorant toutes les valeurs propres de toutes les matrices $A(x)$ pour $x \in \Omega$.

Nous posons

$$L = -\sum_{ij} A_{ij}\partial^2_{ij} + \sum_i b_i\partial_i + c \tag{35.105}$$

où $b_i, c \in C^0(\bar{\Omega})$.

Soit $u \in C^0(\bar{\Omega}) \cap C^2(\Omega)$ telle que $Lu \geqslant 0$ sur Ω. Alors

(1) Si $c = 0$ alors

$$\min_{\bar{\Omega}}(u) = \min_{\partial\Omega}(u). \tag{35.106}$$

(2) Si $c \geqslant 0$ alors

$$\min_{\bar{\Omega}}(u) \geqslant \min_{\partial\Omega}(-u_-) \tag{35.107}$$

où u_- est défini par

$$u_-(x) = \begin{cases} 0 & si\ u(x) \geqslant 0 \\ -u(x) & si\ u(x) \leqslant 0. \end{cases} \tag{35.108}$$

Démonstration. D'abord, vu que Ω est borné, la fermeture $\bar{\Omega}$ est compacte et u y atteint son minimum. De plus $\partial\Omega$ est également compact (borné et le complémentaire est ouvert parce que Ω est ouvert). Donc u y atteint également son minimum. Cela pour dire que les minima écrits dans (35.106) et (35.107) ont un sens.

Pour (1).

$\underline{Lu \geqslant \eta > 0}$ Nous supposons qu'il existe $\eta > 0$ tel que $(Lu)(x) \geqslant \eta$ pour tout $x \in \Omega$. Soit x_0 le point minimum sur $\bar{\Omega}$. Si $x_0 \in \Omega$ alors il est intérieur et $\partial_i u(x_0) = 0$ par la proposition 19.78. Dans ce cas nous avons

$$(Lu)(x_0) = (L'u)(x_0) > 0, \tag{35.109}$$

ce qui contredit le lemme 35.16. Nous en déduisons que le minimum de u sur $\bar{\Omega}$ n'est pas atteint dans Ω, mais sur $\partial\Omega$. Pour la définition de la frontière, voir 9.43. Ici nous avons $\bar{\Omega}\backslash\Omega = \partial\Omega$.

Cela prouve (35.106) dans ce cas.

$\underline{Lu \geqslant 0 \textbf{ sur } \Omega}$ Nous prenons maintenant le cas général. Nous posons

$$u_{\gamma,\epsilon}(x) = u(x) - \epsilon e^{\gamma x_1}. \tag{35.110}$$

Nous avons[7]

$$L(e^{\gamma x_1}) = e^{\gamma x_1}\big(-A_{11}(x)\gamma^2 + b_1(x)\gamma\big). \tag{35.111}$$

Soit γ suffisamment grand pour que

$$\lambda\gamma^2 - \|b_1\|_{\bar{\Omega}}\gamma > 0. \tag{35.112}$$

Ici λ minore toutes les valeurs propres des $A(x)$ et nous notons que γ ne dépend pas de ϵ. En utilisant l'inégalité du lemme 11.18111.181, pour tout $\xi \in \mathbb{R}^n$ nous avons

$$\sum_{ij} A_{ij}\xi_i\xi_j \geqslant \lambda|\xi|^2, \tag{35.113}$$

ce qui donne avec $\xi = e_1$: $A_{11} \geqslant \lambda$, et même pour être plus précis : $A_{11}(x) \geqslant \lambda$ pour tout x. Ces inégalités donnent

$$-A_{11}(x)\gamma^2 + b_1(x)\gamma \leqslant -\lambda\gamma^2 + \|b_1\|_{\bar{\Omega}}\gamma < 0. \tag{35.114}$$

7. Nous abusons un peu de l'écriture parce que ce que nous calculons vraiment est $L\big(x \mapsto e^{\gamma x_1}\big)$.

Nous avons donc

$$L(e^{\gamma x_1}) = e^{\gamma x_1}\big(-A_{11}(x)\gamma^2 + b_1(x)\gamma\big) \leqslant -\lambda\gamma^2 + \|b_1\|_{\bar{\Omega}}\gamma < 0. \tag{35.115}$$

Nous posons

$$\eta = \epsilon\big(\lambda\gamma^2 - \|b_1\|_{\bar{\Omega}}\gamma\big)\min_{\bar{\Omega}}(e^{\gamma x_1}), \tag{35.116}$$

où le minimum a un sens parce que Ω est borné. Nous avons $\eta > 0$.
C'est le moment de calculer ce que u_ϵ peut pour nous :

$$\begin{aligned}
L(u_\epsilon) &= Lu - \epsilon L(e^{\gamma x_1}) && \text{(35.117a)}\\
&\geqslant -\epsilon L(e^{\gamma x_1}) && \text{(35.117b)}\\
&= -\epsilon e^{\gamma x_1}\big(-A_{11}(x)\gamma^2 + b_1(x)\gamma\big) && \text{(35.117c)}\\
&\geqslant -\epsilon e^{\gamma x_1}\big(\lambda\gamma^2 - \|b_1\|_{\bar{\Omega}}\gamma\big) && \text{(35.117d)}\\
&\geqslant \eta > 0. && \text{(35.117e)}
\end{aligned}$$

Justification :
— (35.117b) parce que $Lu \geqslant 0$.
— (35.117d) par (35.115).
La fonction u_ϵ est donc dans le cas précédent et nous avons

$$\min_{x\in\bar{\Omega}}\big(u(x) - \epsilon e^{\gamma x_1}\big) = \min_{x\in\partial\Omega}\big(u(x) - \epsilon e^{\gamma x_1}\big). \tag{35.118}$$

Mentionnons le fait que le choix fait de γ ne dépend pas de ϵ. Nous pouvons donc encore faire varier ϵ sans toucher à γ et en maintenant toutes les inégalités prouvées jusqu'ici.
Vu que l'expression $e^{\gamma x_1}$ est majorable sur $\bar{\Omega}$ nous avons convergence uniforme

$$u_\epsilon \xrightarrow{\|.\|_{\bar{\Omega}}} u \tag{35.119}$$

pour $\epsilon \to 0$.
Supposons qu'aucun point de $\partial\Omega$ ne réalise le minimum de u. Alors il existe $x_0 \in \Omega$ tel que $u(x_0) > u(x)$ pour tout $x \in \partial\Omega$. Mais comme $\partial\Omega$ est compact, il existe $\eta > 0$ tel que

$$u(x_0) < u(x) + \eta. \tag{35.120}$$

Soit ϵ tel que $\|u - u_\epsilon\|_{\bar{\Omega}} < \eta/2$. Nous avons

$$u(x_0) < u(x) - \eta \tag{35.121}$$

et donc aussi

$$u_\epsilon(x_0) < u(x_0) + \frac{\eta}{2} < u(x) - \frac{\eta}{2} < u_\epsilon(x), \tag{35.122}$$

ce qui signifierait que u_ϵ prend son minimum dans Ω. Or nous savons que ce n'est pas le cas.
Donc il existe un point de $\partial\Omega$ qui réalise le minimum de u.

Pour (2).
Supposons pour commencer que $u \geqslant 0$ sur Ω. Alors par continuité $u \geqslant 0$ sur $\bar{\Omega}$. Alors $u_- = 0$ et

$$\min_{\bar{\Omega}} u \geqslant 0, \tag{35.123}$$

ce qui fait que l'inégalité (35.107) est évidente.
Nous supposons donc que l'ensemble

$$\Omega_- = \{x \in \Omega \text{ tel que } u(x) < 0\} \tag{35.124}$$

est non vide. Notons que c'est également un ouvert.

Soit $\bar{L}u = Lu - cu$. Vu que $c \geqslant 0$ et que $Lu \geqslant 0$ nous avons $\bar{L}u \geqslant 0$ sur Ω_-. Par le point (1) nous avons

$$\min_{x\in\bar{\Omega}} u(x) = \min_{x\in\partial\Omega} u(x). \tag{35.125}$$

Mais le minimum de u est certainement atteint dans Ω_-, donc

$$\min_{x\in\bar{\Omega}_-} u(x) = \min_{x\in\bar{\Omega}} u(x). \tag{35.126}$$

Cela nous donne une première bonne égalité :

$$\min_{x\in\partial\Omega_-} u(x) = \min_{x\in\bar{\Omega}} u(x). \tag{35.127}$$

Nous pouvons la prolonger :

$$\begin{aligned}
\min_{x\in\bar{\Omega}} u(x) &= \min_{x\in\partial\Omega_-} u(x) && \text{(35.128a)}\\
&= \min_{x\in\partial\Omega_-} \big(-u_-(x)\big) && \text{(35.128b)}\\
&= \min_{x\in\partial\Omega_-\cap\partial\Omega} \big(-u_-(x)\big) && \text{(35.128c)}\\
&= \min_{x\in\partial\Omega} \big(-u_-(x)\big) && \text{(35.128d)}
\end{aligned}$$

Justifications :

— Pour (35.128c) nous avons la décomposition

$$\partial\Omega_- = \big(\partial\Omega_-\cap\Omega\big)\cup\big(\partial\Omega_-\cap\partial\Omega\big). \tag{35.129}$$

Or sur $\partial\Omega_-\cap\Omega$ nous avons $u(x) = 0$ et donc pas le minimum.

— Pour (35.128d). Sur $\partial\Omega$, le minimum est atteint dans la partie $\partial\Omega_-$ parce que le reste ne contient que des valeurs positives de u.

Cela prouve ce que nous voulions. □

Théorème 35.18.
(Principe du maximum fort) Soit un ouvert borné Ω de $\mathbb{R}^n$. Soit une matrice A dont

— *A_{ij} est dans $C^0(\bar{\Omega})$*

— *$A(x)$ est symétrique strictement définie positive pour tout x.*

— *Il existe $\lambda > 0$ minimisant toutes les valeurs propres des toutes les matrices $A(x)$ sur $\bar{\Omega}$.*

Soit $b_i, c \in C^0(\bar{\Omega})$ avec $c(x) \geqslant 0$ sur $\bar{\Omega}$.

Soit $u \in C^0(\bar{\Omega})\cap C^2(\Omega)$ telle que

$$\begin{cases} -\sum_{ij} A_{ij}\partial_{ij}u(x) + \sum_i b_i\partial_i u(x) \geqslant 0 & \text{(35.130a)}\\ u(x) \geqslant 0 \ \forall x \in \partial\Omega. & \text{(35.130b)}\end{cases}$$

Alors $u \geqslant 0$ sur $\bar{\Omega}$.

Démonstration. Nous appliquons le lemme 35.17(2) :

$$\min_{\bar{\Omega}} u \geqslant \min_{\partial\Omega}(-u_-). \tag{35.131}$$

Vu que $u(x) \geqslant 0$ sur $\partial\Omega$, nous avons $u_- = 0$ sur $\partial\Omega$ et donc $\min_{\bar{\Omega}} u \geqslant 0$. □

35.6 Quelques exemples

35.6.1 Un changement de variables

Soit l'équation différentielle

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \tag{35.132}$$

sur $\Omega =]0, \infty[\times \mathbb{R}$. Nous imposons la condition aux bords

$$u(0,x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x > 0. \end{cases} \tag{35.133}$$

Nous cherchons les solutions sous la forme

$$u(t,x) = f\left(\frac{x}{\sqrt{t}}\right). \tag{35.134}$$

Nous avons

$$\frac{\partial u}{\partial t} = -\frac{1}{2} x t^{-3/2} f'\left(\frac{x}{\sqrt{t}}\right)) \tag{35.135}$$

ainsi que

$$\frac{\partial u}{\partial x} = \frac{1}{\sqrt{t}} f'\left(\frac{x}{\sqrt{t}}\right) \tag{35.136}$$

et

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{t} f''\left(\frac{x}{\sqrt{t}}\right). \tag{35.137}$$

En remettant le tout dans l'équation de départ et en simplifiant par $1/t$ (qui est permis parce que $t > 0$) :

$$-\frac{1}{2} f'\left(\frac{x}{\sqrt{t}}\right) \frac{x}{\sqrt{t}} - f''\left(\frac{x}{\sqrt{t}}\right) = 0. \tag{35.138}$$

Nous résolvons l'équation différentielle

$$\frac{z}{2} g'(z) + g''(z) \tag{35.139}$$

pour la fonction g de la variable réelle z. Cela fait, la réponse sera $f = g \circ z$ où z serait la fonction

$$z(x,t) = x/\sqrt{t}, \tag{35.140}$$

Pour résoudre (35.139) nous commençons par résoudre pour la dérivée $h = g'$, c'est à dire l'équation différentielle

$$\frac{z}{2} h(z) + h'(z) = 0 \tag{35.141}$$

qui donne

$$\frac{g'(z)}{g'z} = -z/2. \tag{35.142}$$

Une intégration fournit $\ln\big(h(z)\big) = -z^2/4 + K$ et donc

$$h(z) = K e^{-z^2/4} \tag{35.143}$$

et

$$g(z) = C + K \int_0^z e^{-s^2/4} ds. \tag{35.144}$$

Nous ne pouvons pas aller plus loin parce que nous ne sommes pas capables de calculer la primitive demandée. Nous laissons donc g sous cette forme et nous posons

$$u(x,t) = g(x/\sqrt{t}), \tag{35.145}$$

en pleine confiance du fait que cela soit une solution de l'équation aux dérivées partielles (35.132).

Nous devons fixer K et C de telle façon à respecter les conditions aux bords (35.133). D'abord écrivons aussi explicitement que possible la fonction u :

$$u(t,x) = K + C\int_0^{x\sqrt{t}} e^{-s^2/4}ds. \tag{35.146}$$

Il faut calculer $u(0,x)$ en termes de C et K. Pour cela nous calculons, pour un x fixé :

$$\lim_{t\to 0^+} u(t,x) = K + C\lim_{t\to 0^+}\int_0^{x/\sqrt{t}} e^{-s^2/4}ds. \tag{35.147}$$

En utilisant un petit changement de variables sur l'intégrale gaussienne de l'exemple 16.242, et en remarquant que la fonction est symétrique,

$$\lim_{t\to 0^+} u(t,x) = \begin{cases} K + C\sqrt{\pi} & \text{si } x > 0 \\ K - C\sqrt{\pi} & \text{si } x < 0. \end{cases} \tag{35.148}$$

À résoudre :

$$\begin{cases} K + C\sqrt{\pi} = 1 & \text{(35.149a)} \\ K - C\sqrt{\pi} = 0. & \text{(35.149b)} \end{cases}$$

Solution : $C = 1/2\sqrt{\pi}$ et $K = 1/2$. Au final,

$$u(t,x) = \frac{1}{2} + \frac{1}{2\sqrt{\pi}}\int_0^{x/\sqrt{t}} e^{-s^2/4}. \tag{35.150}$$

Notons que cela donne une valeur pour $u(t,0)$:

$$u(t,0) = \frac{1}{2}. \tag{35.151}$$

Chapitre 36

Numérique

D'autres lectures agréables dans [396]

36.1 Introduction

À quels types de problèmes peut-on s'attendre lorsqu'on se lance dans du calcul numérique et en particulier dans la résolution numérique d'équations (algébrique, différentielles ou aux dérivées partielles) ?

Quelques réflexions en vrac sur ce sujet.

(1) Les erreurs de représentation de nombres : troncature et drift,

(2) Erreur de cancellation

(3) Conditionnement, stabilité : les réponses peuvent fortement dépendre des paramètres.

(4) Si on utilise une méthode itérative, comment savoir à quel moment on s'arrête ? Calculer la différence $|x_k - x_{k-1}|$ a-t-elle une erreur de cancellation ?

(5) Lors d'une implémentation, les matrices des systèmes à résoudre sont souvent très grandes et très creuses. Cela pose la question de la manière de les enregistrer.

(6) Pour la parallélisation, il faut faire attention au fait que parfois créer un nouveau processus demande plus de ressources que le mini-calcul qu'on voulait faire. Donc il ne faut pas toujours paralléliser tout ce qui est théoriquement parallélisable.

(7) Le fait que certaines méthodes son non-déterministes (Monté-Carlo) mène à des problèmes pour les tests unitaires des implémentations.

36.2 Représentations numériques

Dans cette section les séquences de chiffres écrites entre crochet sont à comprendre comme des séquences de chiffres qui représentent une quantité suivant un codage donné.

36.2.1 Entier relatif en complément à deux (binaire)

Si nous avons m bits pour coder un entier relatif, une idée serait de prendre le premier bit pour le signe (0 pour positif et 1 pour négatif) et les autres pour la valeur absolue. Deux inconvénients :

(1) Il y a deux codages pour le zéro, donc gaspillage.

(2) L'algorithme pour faire la somme passe mal. Par exemple pour faire $1 + (-1)$, le 1 est codé comme [001] et le -1 par [101] et la somme se ferait naïvement comme

$$\begin{array}{ccc} 0 & 0 & 1 \\ 1 & 0 & 1 \\ \hline 1 & 1 & 0 \end{array}$$

Donc le résultat est [110] qui s'interprète comme -2. Complètement faux.

Une solution est d'utiliser le **complément à deux**, qui est la façon usuelle de représenter des entiers signés.

Les entiers positifs se codent normalement, en laissant à zéro le premier bit (donc si nous disposons de m bits, nous codons sur $m-1$ bits).

Les entiers négatifs se codent en trois étapes.
- coder la valeur absolue
- inverser tous les bits (d'où le nom de « complément à deux »)
- soustraire 1.

Exemple 36.1
Pour coder -1 nous faisons
- Nous codons 1 : [001]
- Nous inversons tous les bits : [110]
- Nous faisons -1 : [101].

△

Avec ce système, la somme passe bien : calculer $1+(-1)$ donne

$$\begin{array}{ccc} 0 & 0 & 1 \\ 1 & 0 & 1 \\ \hline 1 & 1 & 0 \end{array}$$

La réponse est donc [110] qu'il faut interpréter via le complément à deux.

$$110 \xrightarrow{+1} 111 \xrightarrow{\text{complément}} 000. \tag{36.1}$$

Et ce dernier [000] s'interprète comme zéro.

Définition 36.2 (Entier signé en complément à deux[397])**.**
La suite de bits $[a_{m-1}\dots a_0]$ *s'interprète via la formule*

$$-a_{m-1}2^{m-1}+\sum_{i=0}^{m-2}a_i2^i. \tag{36.2}$$

Le premier bit donne effectivement le signe du nombre, mais l'interprétation d'un nombre n'est pas aussi simple que ce que l'on pourrait croire de prime abord.

Exemple 36.3(Entier signé en 8 bits)
Que pouvons nous faire avec 8 bits ? Le plus grand nombre est codé par [01111111] qui vaut $\sum_{k=0}^{6}2^k=2^7-1=127$. (avez-vous utilisé la somme (12.126) ?)

Le plus petit nombre codable en 8 bits n'est pas [11111111] mais bien [10000000] (cela est plus clair en regardant la formule (36.2) qu'en tentant de suivre la construction du complément à deux) qui signifie $-2^7=-128$.

Nous pouvons donc coder tous les nombres de -128 à 127. △

Plus généralement un système qui codes des entiers signés en N bits utilisant le complément à deux peut coder de $-(2^{N-1})$ à $2^{N-1}-1$.

36.4 (Le dépassement)**.**
Que se passe-t-il lorsque nous commettons un dépassement ? Calculons sur 3 bits la somme [011] + [001] qui revient à ajouter 1 au nombre le plus grand :

$$\begin{array}{ccc} 0 & 1 & 1 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \end{array}$$

qui signifie $-2^2 = -4$. Lors d'un dépassement, nous retombons automatiquement sur le plus petit.

Ce phénomène est bien connu des personnes qui programment sans faire attention dans certains languages de programmation qui ne font pas attention à votre place.

Définition 36.5 (Représentation en virgule fixe)**.**
Soit x un réel. On définit sa ***représentation en virgule fixe*** *par*

$$x = \{[x_n x_{n-1} ... x_0, x_{-1} ... x_{-m}], b, s\} \tag{36.3}$$

avec $b \in \mathbb{N}, b \geqslant 2$, $s \in \{0, 1\}$ et $x_j \in \mathbb{N}, x_j < b$ suivant la formule

$$x = (-1)^s \sum_{j=-m}^{n} x_j . b^j. \tag{36.4}$$

36.2.2 Représentation en virgule flottante

Définition 36.6 (Représentation en virgule flottante)**.**
La ***représentation en virgule flottante normalisée*** *en base b d'un nombre est la donnée de*

(1) Un bit s pour le signe

(2) Un entier non signé *q de e chiffres pour l'exposant*

(3) Une suite de chiffres $[a_1 \dots a_m]$ pour la mantisse.

Ces données s'interprètent via la formule

$$\mathrm{fl}(s, q, [a_1, \dots, a_m]) = (-1)^s \sum_{j=1}^{m} b^j a_j \times b^{q-d} \tag{36.5}$$

où $d = b^{e-1}$ est le ***décalage****.*

Une idée à retenir est que l'exposant est un entier non signé parce qu'il est plus simple d'introduire un décalage dans la formule (36.5) que de compliquer l'écriture de l'exposant.

36.2.3 Simple précision, IEEE-754

En écriture binaire, la représentation en virgule flottante est un peu différente parce qu'il y a une idée supplémentaire ; la simple précision que nous allons voir maintenant n'est donc pas un cas particulier de 36.6 avec $b = 2$.

Nous commençons par une description informelle de la précision simple avant de donner la définition. La représentation en **précision simple** d'un nombre se fait sur 32 bits répartis comme suit :

(1) 1 bit pour le signe,

(2) 8 bits pour l'exposant interprété comme nombre entier non signé

(3) 23 bits pour la mantisse

Soit le triple

$$\big(s, q, [a_1, \dots, a_{23}]\big) \tag{36.6}$$

Dans le cas générique, l'idée est de donner 24 bits pour la mantisse, mais en gardant en tête le fait que de toutes façons, le premier bit doit être 1, sinon il suffirait de décaler, c'est à dire changer l'exposant. Par conséquent la mantisse ne reçoit que 23 bits ; il y a un « 1 » sous-entendu en première position. Donc la mantisse $[a_1, \dots, a_{23}]$ est à lire comme le nombre

$$1, a_1 \dots a_{23} = 1 + \sum_{j=1}^{23} a_j 2^{-j}. \tag{36.7}$$

Exemple 36.7

La mantisse $[011100\dots0]$ signifie $1,0111 = 1 + 2^{-2} + 2^{-3} + 2^{-4} = 1 + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}$. △ Cela pour justifier la formule

$$\text{sp}\big(s, q, [a_1, \dots, a_{23}]\big) = (-1)^s\big(1 + \sum_{j=1}^{23} a_j 2^{-j}\big)2^{q-127}. \tag{36.8}$$

Notons :

(1) Le « 1+ » dans la parenthèse correspond au 1 implicite en première position de la mantisse.

(2) Il y a un décalage de 127 dans l'exposant, parce que q est un entier non signé.

Notons que cette règle du 1 implicite dans la mantisse empêche d'écrire le nombre 0, et ne permet pas d'écrire des nombres franchement petits parce que le 1 implicite est en *première* position dans la mantisse.

D'où l'idée de donner une règle particulière lorsque l'exposant vaut 0. Lorsque l'exposant est $q = 0$, alors nous ne considérons pas de 1 implicite dans la mantisse, et le décalage de l'exposant est -126 au lieu de -127. D'où la formule

$$\text{sp}(s, q = 0, [a_1 \dots a_{23}]) = (-1)^s 2^{-216} \sum_{j=1}^{23} a_j 2^{-j}. \tag{36.9}$$

En particulier, si $q = 0$ et $a = [0\dots0]$, nous avons le nombre zéro exact (il y a deux possibilités pour le code).

Enfin, nous avons des cas particuliers lorsque l'exposant est maximum, c'est à dire $q = [1111\,1111] = 2^8 - 1 = 255$. Dans ce cas, le nombre codé est soit $+\infty$ soit NaN. Nous posons $\text{sp}(s, q = 255, a = 0) = +\infty$ et $\text{sp}(s, q = 255, a \neq 0) = NaN$. Il y a en réalité plusieurs valeurs différentes de NaN, mais nous n'entrons pas dans ces détails[398].

Définition 36.8 (Représentation en simple précision (binaire))**.**
La représentation en ***précision simple*** *d'un nombre se fait sur* 32 *bits répartis comme suit :*

(1) 1 *bit pour le signe,*

(2) 8 *bits pour l'exposant interprété comme nombre entier non signé*

(3) 23 *bits pour la mantisse*

Un nombre est représenté par un triple

$$\big(s, q, [a_1, \dots, a_{23}]\big) \tag{36.10}$$

Selon que l'exposant $q - d$ soit égal à 0, $2^8 - 1 = 255$ ou autre chose, les règles d'interprétation sont différentes. Il y a donc trois cas.

Exposant q générique[399] *Si $q \neq 0$ et $q \neq 255$ alors le nombre est* ***normalisée****. La règle de lecture est alors*

$$\text{sp}\big(s, q, [a_1, \dots, a_{23}]\big) = (-1)^s\big(1 + \sum_{j=1}^{23} a_j 2^{-j}\big)2^{q-127}. \tag{36.11}$$

Exposant q égal à 0 *Le nombre est dit* ***dénormalisé*** *et la règle de lecture est*

$$\text{sp}\big(s, q, [a_1, \dots, a_{23}]\big) = (-1)^s 2^{-126} \sum_{j=1}^{23} a_j 2^{-j}. \tag{36.12}$$

Exposant q égal à 255 *La règle de lecture est alors au cas pas cas ou à peu près.*

(1) $\text{sp}(s, q = 255, a = 0) = +\infty$.

(2) $\text{sp}(s, q = 255, a \neq 0) = \text{NaN}$.

Vous pouvez jouer avec la simple précision dans [400].

Exemple 36.9(Plus petit normalisé)
Pour faire un nombre normalisé, il faut au minimum $q = 1$. En prenant $a_j = 0$ nous obtenons le plus petit nombre normalisé possible en simple précision. La formule (36.11) donne

$$\text{sp}(1, q = 1, a = 0) = 2^{1-127} = 2^{-126} \simeq 1.17549435082229 \times 10^{-38}. \tag{36.13}$$

△

Exemple 36.10(Plus grand normalisé)
L'exposant q ne peut pas être maximum, sous peine de tomber dans les règles spéciales de $+\infty$ ou NaN. Donc $q = [1111\,1110] = 2^8 - 2 = 254$. En ce qui concerne la mantisse, il faut la prendre maximale, c'est à dire $a_j = 1$ pour tout j. Nous avons alors le nombre

$$\text{sp}(1, q = 254, a = [1 \dots 1]) = \big(1 + \sum_{j=1}^{23} 2^{-j}\big)2^{254-127} = (1 - \frac{1}{2^{24}})2^{128} \tag{36.14a}$$

$$= 3.40282346638528859811704183484516925440 \times 10^{38} \tag{36.14b}$$

où nous avons utilisé la somme (12.126) (et Sage pour le dernier calcul). △

Notons ceci avec Sage :

```
1
2  SageMath Version 7.0, Release Date: 2016-01-19
3  Type "notebook()" for the browser-based notebook interface.
4  Type "help()" for help.
5
6 sage: A=(1- (1/2**24)  )*2**(128)
7 sage: type(A)
8 <type 'sage.rings.rational.Rational'>
```

tex/sage/sageSnip003.sage

La précisions du nombre donné en (36.14b) aurait été embarrassante si le type avait été un nombre en simple précision. Précision technique : en Python, le type `int` n'a pas de limite supérieure à part la mémoire.

Exemple 36.11(Plus petit non nul dénormalisé)
Pour être dénormalisé il faut $q = 0$ (ce qui est toutefois assez logique si nous voulons un petit nombre), et pour ne pas être nul, il faut une mantisse non nulle. Donc $a = [0 \dots 01]$. La formule (36.12) donne alors

$$\text{sp}(s = 0, q = 0, a = [0 \dots 01]) = 2^{-126}2^{-23} = 2^{-149} \simeq 1.40129846432482 \times 10^{-45}. \tag{36.15}$$

△

Exemple 36.12(Plus grand dénormalisé)
Pour être dénormalisé il faut toujours $q = 0$, mais cette fois nous prenons la plus grande mantisse possible :

$$\text{sp}(s = 0, q = 0, a = [1 \dots 1]) = 2^{-126} \sum_{j=1}^{23} 2^{-j} = 2^{-216}(1 - 2^{-23}) = 1.17549421069244 \times 10^{-38} \tag{36.16}$$

△

Notons ceci avec Sage :

```
sage: B=2**(-126)*(1-2**(-23))
sage: A=2**(-126)
sage: n(A-B)
1.40129846432482e-45
```

tex/sage/sageSnip004.sage

Vu que $2^{-23} \simeq 1.2 \times 10^{-7}$, approximer la parenthèse par 1 donne une faute sur la septième décimale, ce qui est visible en simple précision.

36.3 Problèmes pour écrire des nombres

Définition 36.13.
*L'**erreur relative** commise en remplaçant un nombre réel x par une valeur approchée $\hat{x}$ est définie par*

$$\epsilon_x := \left| \frac{x - \hat{x}}{x} \right|. \tag{36.17}$$

L'erreur relative n'est pas influencée par l'ordre de grandeur de x. En effet, l'ordre de grandeur de $\hat{x}$ est certainement la même que celle de x, dans la majorité des cas sans problèmes. Du coup si $x' = 200x$ alors $\hat{x'} \simeq 200\hat{x}$ et le 200 se simplifie.

Le nombre de chiffres significatifs correct dans l'approximation est donné par $-\log_{10}(\epsilon_x)$. La partie entière de ce nombre est le nombre de chiffres tout à fait exacts et la partie décimale donne une idée sur le fait que le chiffre suivant est plus ou moins bien.

Remarque 36.14.
Si nous voulons donner $x \in \mathbb{R}$ à un ordinateur, nous sommes soumis à deux erreurs :

(1) D'abord, vu que nous ne pouvons pas taper sur le clavier toutes les décimales de x, nous faisons une **erreur de troncature**.

(2) L'ordinateur devant convertir cela en base deux, il commet une seconde erreur, dite **erreur d'assignation**.

36.3.1 Troncature : la base

Supposons que nous voulions écrire le nombre (écrit ici en base 10)

$$0.4567894251 \tag{36.18}$$

de façon plus facile à lire, on peut demander de ne laisser que t chiffres significatifs. Disons $t = 3$.

Technique de troncature On garde 3 chiffres significatifs : 0.456. Facile.

Technique d'arrondi Vu que le premier qu'on supprime est un 7, le dernier qu'on garde est majoré de 1 : on écrit 0.457.

Que faire si le premier chiffre rejeté est un 5 ? En première approximation, nous pouvons prendre la règle suivante : si le premier chiffre rejeté est un 5, il faut augmenter de 1 de dernier chiffre gardé parce qu'il y a presque certainement encore un chiffre non nul derrière.

Remarque 36.15.
Les ordinateurs travaillent tous en mode d'arrondi.

Exemple 36.16
Si on doit entrer le nombre 0.38358546 dans un ordinateur qui ne garde que 3 chiffres significatifs, il faut taper 0.384 au clavier (erreur classique dans les exercices). △

36.3.2 Troncature : le drift

Soit une machine ne pouvant retenir que 3 chiffres significatifs et effectuant les arrondis vers le haut lorsque le chiffre à éliminer est un 5. Nous notons $\oplus$ et $\ominus$ les opérations d'addition et soustraction avec arrondis[401]. Les égalités comprenant plus de trois chiffres significatifs sont des égalités au sens de la machine. Nous écrirons donc sans états d'âme :

$$1 \oplus 0.555 = 1.555 = 1.56. \tag{36.19}$$

Considérons la suite numérique

$$\begin{cases} x_0 = 1.00 & \text{(36.20a)} \\ x_n = (x_{n-1} \ominus y) \oplus y & \text{(36.20b)} \end{cases}$$

avec $y = -0.555$.

Nous avons

$$x_1 = (1 \oplus 0.555) \ominus 0.555 = 1.56 \ominus 0.555 = 1.005 = 1.01 \tag{36.21}$$

et ensuite

$$x_2 = (1.01 \oplus 0.555) \ominus 0.555 = 1.565 \ominus 0.555 = 1.57 \ominus 0.555 = 1.015 = 1.02. \tag{36.22}$$

Et ainsi de suite. La suite est donc croissante alors que la définition nous donnerait envie d'avoir $x_n = x_0$ pour tout n.

Remarque 36.17.
En réalité, cette suite se stabilise à $x_n = 10$ pour tout n à partir de $n = 845$. En effet,

$$(10 \oplus 0.555) \ominus 0.555 = 10.555 \ominus 0.555 = 10.6 \ominus 0.555 = 10.045 = 10. \tag{36.23}$$

Le fait est qu'à ce moment, l'erreur de troncature est assez loin dans les décimales pour que le premier chiffre négligé soit un “0” au lieu d'un “5”.

Notons toutefois que cette stabilité n'est pas là pour nous rassurer parce qu'elle n'en est pas moins complètement fausse.

La règle de troncature adoptée dans Sage est d'arrondir au nombre pair le plus proche lorsque le premier nombre à négliger est un 5. Donc 12.5 s'arrondit à 12 plutôt que 13.

Exemple 36.18
Soient les expressions (algébriquement égales) :

(1) $A = x(x + 1)$

(2) $B = x^2 + x$

Nous savons que

$$x = \text{fl}(x) = 10^{-30} \tag{36.24}$$

et

$$1 = \text{fl}(1) \tag{36.25}$$

parce que pour 1 et 10^{-30}, il n'y a pas d'erreurs d'assignation.

En précision simple, $10^{-30} + 1 = 1$ parce qu'en précision simple, il n'y a que 7 ou 8 chiffres significatifs [1].

Nous avons $A = 10^{-30}$, mais x^2 donne un `underflow` parce que 10^{-60} ne peut pas être représenté en précision simple. En pratique, beaucoup de logiciels en font 0. Dans ce cas, en réalité B donne effectivement 10^{-30} après avoir fait $x^2 + x = 0 + x = 10^{-30}$. △

1. Erreur de « relation normale ».

36.3.3 Quelques bonnes règles

(1) Si on a plusieurs nombres à additionner ou soustraire, il vaut mieux commencer par sommer ou soustraire ceux dont on sait qu'ils ont le même ordre de grandeur. Il n'y a donc pas tout à fait « associativité » des erreurs.

(2) Les opérations délicates sont l'addition et la soustraction. La multiplication et la division sont sans dangers, à part l'erreur de dépassement du maximum. Dans une multiplication, on perd au pire quelques chiffres significatifs, mais certainement les derniers, pas les premiers.

36.3.4 Erreur de "cancellation"

Lorsque deux nombres sont de même ordre de grandeur, avec plusieurs nombres significatifs identiques. La cancellation est le fait que, suite à la soustraction, tous les chiffres significatifs ou presque se sont simplifiés et qu'il ne reste plus que des chiffres non significatifs.

Exemple 36.19([402])
Sur une machine ne gardant que 4 chiffres significatifs, faire

$$0.5678 \times 10^6 - 0.5677 \times 10^6 = 0.0001 \times 10^6 = 0.1000 \times 10^3. \tag{36.26}$$

Le fait est que les trois derniers zéros ne sont pas significatifs, mais maintenant la machine nous fait croire qu'ils le sont.

Une autre façon de voir ce problème est d'imaginer qu'il faille calculer la différence

$$0.5678\,289798 \times 10^6 - 0.5677\,3136907 \tag{36.27}$$

sur cette machine. Certes la machine nous autorise à avoir 4 chiffres significatifs, donc au moment d'entrer les nombres nous perdons un beau paquet de chiffres. Mais au moment de faire la différence, nous perdons (presque) tout le reste. Donc là où nous pouvions espérer avoir 4 chiffres significatifs de la différence, nous n'en avons que 1. Les trois derniers zéros de la réponse (0.1000×10^3) sont faux. △

Remarque 36.20.
L'erreur de cancellation provoque des chiffres significatifs faux, mais ne provoque pas de faute dans l'*ordre de grandeur* des réponses[2]. Donc si nous voulons nous assurer que a et b sont égaux « à erreur numérique près », le test

$$|a - b| < \epsilon \tag{36.28}$$

est valide, malgré l'erreur de cancellation qui ne manquera pas de se produire dans le calcul de la différence.

Exemple 36.21
Soit à résoudre l'équation $ax^2 + bx + c = 0$ avec $a, b, c \neq 0$ et $b^2 - 4ac > 0$. Solution :

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{36.29}$$

Supposons que $|4ac| \ll b^2$ avec tout de même pas tellement petit qu'on se perd dans la précision. Bref, on suppose que seules quelques dernières décimales de $b^2 - 4ac$ sont différentes de zéro.

On a :

$$\sqrt{b^2 - 4ac} = \sqrt{\tilde{b}} = |\tilde{b}| \tag{36.30a}$$

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \tag{36.30b}$$

$$x_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{36.30c}$$

2. Est-ce bien vrai, cela ?

Si $b > 0$, nous avons une erreur de cancellation dans x_2 parce qu'on fait la différence entre deux nombres presque égaux. Donc x_2 mal calculé. Par contre x_1 est bien calculé.

Si par contre $b < 0$, c'est le contraire.

Avec $a = 10^{-3}$, $b = 0.8$, $c = -1.2 \times 10^{-5}$. À la main nous obtenons : $x_1 = -800$, $x_2 = 1.5 \times 10^{-5}$, et un ordinateur se tromperait ...

```
1
2  SageMath Version 7.0, Release Date: 2016-01-19
3  Type "notebook()" for the browser-based notebook interface.
4  Type "help()" for help.
5
6 sage: f(x)=10**(-3)*x**2+0.8*x-1.2*10**(-5)
7 sage: solve(f(x)==0,x)
8 [x == -1/50*sqrt(400000030) - 400, x == 1/50*sqrt(400000030) - ↩
     400]
9 sage: numerical_approx(-1/50*sqrt(400000030))
10 -400.000015000000
11 sage: numerical_approx( 1/50*sqrt(400000030) - 400  )
12 0.0000149999996779115
```

tex/sage/sageSnip001.sage

Donc Sage ne tombe pas dans le piège. △

Comment résoudre ce problème ? Ou, autre façon de poser la question : comment Sage a fait pour résoudre le problème ?

Utilisons les relations coefficients-racines :

$$x_1 + x_2 = -b/a \tag{36.31a}$$
$$x_1 x_2 = c/a \tag{36.31b}$$

La première lie les deux racines par des opérations de addition et soustractions, et donc n'est pas intéressantes. La seconde est bien. Si nous connaissons x_1, nous calculons

$$x_2 = \frac{c}{ax_1}. \tag{36.32}$$

Quitte à redéfinir x_1 et x_2, la solution bien calculée est :

$$x_1 = \frac{-b - \operatorname{sgn}(b)\sqrt{b^2 - 4ac}}{2a}. \tag{36.33}$$

Exemple 36.22
Nous considérons :

$$f(x) = cos(x + \delta) - \cos(x). \tag{36.34}$$

Cela a une erreur de cancellation lorsque $|\delta| \ll |x|$. On élimine l'erreur de cancellation par

$$f(x) = -2\sin(\delta/2)\sin\left(x + \frac{\delta}{2}\right). \tag{36.35}$$

Problèmes et choses à faire

Pourquoi la condition pour avoir l'erreur est $\delta \ll x$ et non simplement $\delta \ll 1$?

△

Exemple 36.23
Pour

$$f(x) = \sqrt{x+\delta} - \sqrt{x}. \tag{36.36}$$

On fait la coup du binôme conjugué :

$$f(x) = \frac{\delta}{\sqrt{x+\delta} + \sqrt{x}}. \tag{36.37}$$

Plus d'erreur de cancellation, vu qu'au dénominateur nous avons une somme de deux positifs. △

Les erreurs de cancellation ne se résolvent pas en augmentant la précision des nombres donnés.

Exemple 36.24(Dans la vie réelle)
La préparation de l'exemple 19.54 nous a porté à calculer la différence entre $\exp(x)$ et $f_{30}(x)$ où f_{30} est censée être une bonne approximation de l'exponentielle. Des erreurs de cancellation sont donc à craindre.

Et en effet, le code suivant produit un résultat non déterministe :

```
1 f=1/15244417230585693025075200000*x^28 + ↩
    1/10888869450418352160768000000*x^27 + ↩
    1/155112100433309859840000000*x^25 + ↩
    1/3102242008666197196800000*x^24 + 1/2585201673888497664000000*x↩
    ^23 + 1/510909421717094400000*x^21 + 1/1216451004088320000*x^20↩
     + 1/121645100408832000*x^19 + 1/3556874280960000*x^17 + ↩
    1/104613949440000*x^16 + 1/13076743680000*x^15 + 1/622702080*x↩
    ^13 + 1/239500800*x^12 + 1/39916800*x^11 + 1/3628800*x^9 + ↩
    1/201600*x^8 + 1/50400*x^7 + 1/1200*x^5 + 1/12*x^4 + 1/6*x^3 + x ↩
    - cos(x) + 2 -exp(x)
2 a=numerical_approx(10)
3 print(f(a))
```

tex/sage/sageSnip016.sage

Voir la question ici :
https://ask.sagemath.org/question/37946/undeterministic-numerical-approximation/

△

36.3.5 Calcul d'une dérivée

Pour calculer la dérivée de f en a, il est loisible d'utiliser la formule

$$f'(a) = \lim_{h\to 0} \frac{f(a+h) - f(a)}{h}. \tag{36.38}$$

Le numérateur est alors sujet à une erreur d'absorption dans le calcul de $a+h$ et ensuite une erreur de cancellation dans le calcul de la différence.

En utilisant la formule

$$f'(a) = \lim_{h\to 0} \frac{f(a+h) - f(a-h)}{2h} \tag{36.39}$$

nous pouvons espérer avoir une erreur de cancellation plus petite.

36.3.6 Erreur d'absorption

L'addition d'un nombre avec un nombre très différent peut faire perdre de l'information sur le plus petit. Par exemple avec 4 chiffres significatifs,

$$0.5678 \oplus 0.0001237 = 0.5679 \tag{36.40}$$

où nous avons perdu presque toute l'information du petit nombre.

Une situation particulièrement ennuyeuse est celle où justement c'est le petit nombre qui nous intéresse parce que le grand est censé se simplifier :

$$(0.0001327 \oplus 0.5678) \ominus 0.5678 = 0.5679 \ominus 0.5678 = 0.0001 \tag{36.41}$$

qui ne possède qu'un seul chiffre significatif correct alors que voyant le calcul, la réponse aurait pu être trouvée.

Moralité : si certains manipulations algébrique peuvent faire apparaître des simplifications avant de passer le calcul à la machine, il est bon de les effectuer.

36.4 Conditionnement et stabilité

Définition 36.25.
Soit F une fonction à valeurs réelles définie sur $X \times D$ où X et D sont des espaces vectoriels réels normés. Le problème de la recherche des solutions de

$$F(x, d) = 0 \tag{36.42}$$

*est dit **stable** autour de $d_0 \in D$ si*

(1) la solution $x = x(d)$ existe et est unique pour tout d ;

(2) Pour tout $\eta > 0$, et pour tout d_0, il existe un nombre $K > 0$ tel que $\|d - d_0\| < \eta$ entraine $\|x(d) - x(d_0)\| \leqslant K \|d - d_0\|$.

La seconde condition est le fait que x soit Lipschitz[3] sur un voisinage de d_0.

Exemple 36.26(Stabilité de la différence)
Prenons le problème qui consiste à calculer la différence entre deux nombres : $x = a - b$. Cela se traduit par

$$\begin{aligned} F\colon \mathbb{R} \times \mathbb{R}^2 &\to \mathbb{R} \\ x &\mapsto x - a + b. \end{aligned} \tag{36.43}$$

Nous avons :

$$\begin{aligned} \big|x(a,b) - x(a',b')\big| &= |a - b - a' + b'| && \text{(36.44a)} \\ &\leqslant |a - a'| + |b - b'| && \text{(36.44b)} \\ &= \|(a,b) - (a',b')\|_1 && \text{(36.44c)} \end{aligned}$$

où nous avons utilisé la norme $\|.\|_1$ sur $\mathbb{R}^2$. Par la proposition 12.4 sur les équivalences de normes, le nombre $K = \sqrt{2}$ fonctionne pour toute valeurs de η.

La problème de la différence est donc un problème stable. △

Exemple 36.27(Stabilité de la multiplication)
Si a est fixé, le problème de calculer ab (b est la donnée) est stable. En effet ce problème est donné par la fonction $F(x, b) = x - ab$, dont la solution est $x(b) = ab$. Nous avons donc

$$\big|x(b) - x(b')\big| = |ab - ab'| = |a||b - b'|. \tag{36.45}$$

La constante de Lipschitz de ce problème est donc $|a|$. △

3. Définition 14.250.

Définition 36.28.
Le nombre

$$K_{abs}(d_0,\eta) := \sup_{d \text{ tel que } |d_0-d|<\eta} \frac{\|x(d)-x(d_0)\|_X}{\|d-d_0\|_D} \tag{36.46}$$

est appelé le ***conditionnement absolu*** *du problème autour de d_0.*

Soit $F(x,d)=0$ un problème stable de conditionnement absolu $K_{abs}(d,\eta)$. Le conditionnement relatif est défini par

$$K_{rel}(d,\eta) := K_{abs}(d,\eta)\frac{\|d\|_D}{\|x(d)\|_X}. \tag{36.47}$$

Le problème est dit ***bien conditionné*** *près de d si $K_{rel}(d,\eta)$ est petit.*

Exemple 36.29(Mauvais conditionnement de la différence)
Reprenons le problème de la différence, mais en fixant a. Nous avons donc $x(b)=a-b$ et le conditionnement absolu est

$$\sup \frac{|x(b)-x(b_0)|}{|b-b_0|} = 1 \tag{36.48}$$

Le conditionnement relatif est :

$$K_{rel}(b_0,\eta) = \frac{|b|}{|a-b|}. \tag{36.49}$$

Et donc le problème est mal conditionné autour de a.

Autrement dit, si a' est un nombre proche de a, calculer la différence $a-a'$ est un problème mal conditionné. △

Exemple 36.30(Bon conditionnement de la multiplication)
Pour le problème $F(x,b)=x-ab$ nous avons

$$K_{abs} = \sup_{b'} \frac{|ab-ab'|}{|b-b'|} = |a|. \tag{36.50}$$

Et aussi

$$K_{rel} = a\frac{|b|}{|ab|} = 1. \tag{36.51}$$

Le conditionnement relatif du problème de la multiplication est donc toujours 1. Il est donc un toujours un problème bien conditionné. △

Ne pas confondre :

Le conditionnement provient du problème lui-même.

La stabilité provient de l'algorithme de résolution.

Exemple 36.31(Un problème mal conditionné)
Le système

$$\begin{cases} 2.1x + 3.5y = 8 & \text{(36.52a)} \\ 4.19x + 7.0y = 15 & \text{(36.52b)} \end{cases}$$

Solution : $x=100$, $y=-57.714285\ldots$ (périodique)

Perturbons : nous remplaçons 4.19 par 4.192. L'erreur relative est : 4.77×10^{-4}.

Solution : $\bar{x}=125$, $\bar{y}=-72.714285\ldots$, avec donc erreur relative de 0.26. Autrement dit : l'erreur relative sur la solution est grande même avec une petite erreur relative sur la donnée.

C'est un problème mal conditionné.

Le fait est que c'est une intersection de deux droites presque parallèles. Donc effectivement une petite perturbation d'une des deux droites donne une grande perturbation du point d'intersection.

Le fait est qu'un ordinateur effectue *toujours* une perturbation, au moins de l'ordre 10^{-16} pour ne fut-ce que représenter les nombres. C'est à dire une perturbation sur les six nombres définissant le système. Il n'y a donc pas d'espoir d'obtenir un algorithme donnant une bonne réponse. △

Un résultat pratique pour étudier le conditionnement d'un problème est le suivant.

Corollaire 36.32.
Soit $x = x(d)$ un problème stable. Supposons $\mathbb{D}$ de dimension finie, supposons que U est ouvert dans $\mathbb{D}$. Supposons encore $x : U \to \mathbb{R}$ différentiable en d_0. Alors quand η est petit, on a

$$K_{abs}^{\eta}(d_0) \sim \|\nabla x(d_0)\|. \tag{36.53}$$

Lemme 36.33.
Tout problème de la forme $x = x(d)$ avec $d \in \mathbb{R}$ et $x \in C^1(\mathbb{R})$ est stable.

Démonstration. Il faut démontrer qu'une fonction C^1 sur $\mathbb{R}$ vérifie automatiquement la condition (2) de la définition de la stabilité. Pour cela, remarquons qu'une fonction C^1 possède une dérivée continue, et donc bornée sur tout compact [4]

Prenons $\eta > 0$ et $d_0 \in \mathbb{R}$ et puis un d tel que $|d - d_0| < \eta$. Par le théorème des bornes atteintes, la fonction x' est bornée sur l'intervalle $[d_0 - \eta, d_0 + \eta]$. Appelons K un majorant de x' sur cet intervalle. La fonction

$$f(d) = x(d_0) + K|d - d_0| \tag{36.54}$$

majore $x(d)$, et donc on a

$$\big|x(d) - x(d_0)\big| \leqslant K|d - d_0|. \tag{36.55}$$

Attention : vérifier si ce raisonnement est correct avec $d_0 > d$, et adapter au besoin. □

Exemple 36.34
Un exemple de problème stable de la forme $x = x(d)$ avec $d \in \mathbb{R}$ et $x \in C^0(\mathbb{R})\backslash C^1(\mathbb{R})$.

La fonction

$$x(d) = \begin{cases} 0 & \text{si } x \geqslant 0 \\ x & \text{si } x > 0 \end{cases} \tag{36.56}$$

est continue, mais pas C^1 (non dérivable en $x = 0$). La dérivée est partout bornée par 1, et donc le problème est stable.

Un autre exemple très classique serait de prendre $x(d) = |d|$. Dans ce cas, on peut prendre n'importe que η et $K = 1$. Le calcul est que

$$|x(d) - x(d_0)| < K|d - d_0| \tag{36.57a}$$
$$\big||d| - |d_0|\big| < |d - d_0|. \tag{36.57b}$$

Cette dernière inéquation est correcte, comme on peut le voir en mettant au carré les deux membres.

△

Exemple 36.35
Un exemple de problème instable de la forme $x = x(d)$ avec $d \in \mathbb{R}$ et $x \in C^0(\mathbb{R})$.

Un exemple assez classique de fonction dont la dérivée n'est pas bornée sans pour autant que la fonction aie un comportement immoral [5] est $x \mapsto \sqrt{x}$. Afin d'avoir une fonction définie sur $\mathbb{R}$ tout entier, nous regardons la fonction

$$x(d) = \sqrt{|d|}. \tag{36.58}$$

4. Un compact est un ensemble fermé et borné, typiquement un intervalle du type $[a, b]$.
5. Penser à $x \mapsto x \sin(1/x)$.

Si nous considérons maintenant $d_0 = 0$ et n'importe quel η, nous avons

$$\frac{|x(d) - x(d_0)|}{|d - d_0|} = \frac{\sqrt{d}}{d} = \frac{1}{\sqrt{d}}. \tag{36.59}$$

Il n'est pas possible de trouver un K qui majore ce rapport. Le problème est donc mal conditionné.

Attention : dans ce calcul nous avons supposé $d > 0$. Pensez à adapter au cas $d < 0$. △

Exemple 36.36(Problème bien conditionné avec algorithme instable)
Soit à calculer

$$I_n = \frac{1}{e}\int_0^1 x^n e^x dx \tag{36.60}$$

avec $n \geqslant 0$. Par partie, nous obtenons :

$$I_n = 1 - nI_{n-1}. \tag{36.61}$$

D'autre part, $I_0 = \frac{e-1}{e}$, $I_1 = \frac{1}{e}$. Puis par récurrence, c'est tout en main.

Du côté de l'ordinateur, nous lui donnons forcément une approximation de I_1, parce que nous lui donnons une approximation de e. Soit l'erreur ϵ_1 sur I_1.

Sans démonstration :

Lemme 36.37.
Nous avons $\lim_{n\to\infty} I_n = 0$.

Mais numériquement, il n'est pas possible de rester longtemps sous ϵ_1 parce que nous n'espérons pas avoir une erreur plus petite que ça. Donc à partir du moment où $I_n < \epsilon_1$, les valeurs sont toutes complètement fausses. Cela est le mieux que l'on puisse espérer. Mais la réalité est pire.

En réalité, en lançant le calcul sur un ordinateur, les valeurs sont même croissantes avec n à partir d'un certain moment.

On peut étudier l'erreur et montrer que l'erreur est donnée par :

$$\epsilon_n = (-1)^{n-1} n! \epsilon_1. \tag{36.62}$$

Mais comme la factorielle est tellement forte que c'est sans espoir d'aller loin en essayant très fort de donner une petite erreur sur ϵ_1.

△

Il existe heureusement un algorithme stable pour cette intégrale. La formule est :

$$I_{n-1} = \frac{1}{n}(1 - I_n). \tag{36.63}$$

Si nous savons un I_N avec N grand, cette formule donne les I_i avec $i = N, N-1, \ldots, 2$. Posons donc $I_N = a \in \mathbb{R}$ n'importe comment. Donc ϵ_N est grand. Mais il se trouve que l'erreur sur ϵ_1 est donnée par

$$\epsilon_1 = \frac{(-1)^{N-1}}{N!}\epsilon_N. \tag{36.64}$$

Donc même en prenant vraiment n'importe quoi pour I_N, nous obtenons de bonnes approximations pour I_i avec les petits i. Même avec $I_{20} = 1000$ (qui est complètement faux), nous trouvons énormément de chiffres significatifs corrects pour I_1.

36.4.1 Comment choisir et penser le K ?

La formule (36.46) contient une formule qui ressemble étrangement à la dérivée. La stabilité d'un problème est très liée à la dérivée de F. La stabilité et la dérivée ne sont pas les mêmes choses,

mais il n'est pas mauvais de penser au K de la stabilité comme la dérivée. Ou plus précisément : le supremum de la dérivée.

Un fil conducteur du lemme 36.33 et des exemples 36.34, 36.35 est que l'on a un K qui fonctionne lorsque la dérivée est bornée sur l'intervalle $]d_0 - \eta, d_0 + \eta[$. Dans le cas où ce supremum existe, le prendre en guise de K fonctionne souvent.

Il faut cependant parfois faire acte d'imagination. La fonction $x \mapsto |x|$ n'est pas dérivable en 0. Il n'empêche que $K = 1$ fait fonctionner la définition de la stabilité. Remarquez que $K = 1$ est le supremum de la dérivée là où elle existe.

À partir du moment où c'est clair que le K est le supremum de la dérivée, on comprend pourquoi c'est le gradient qui arrive dans le corollaire 36.32. En effet, le gradient indique la direction de plus grande pente. C'est donc bien dans cette direction qu'il faut chercher la « plus grande dérivée ».

Proposition 36.38.
Pour le problème stable $x = x(d)$ avec $x \in C^1(\mathbb{R}^n, \mathbb{R})$, on a

$$K_{abs}(d) \sim \|dx_d\| \tag{36.65}$$

où dx_d désigne la différentielle de x en d et la norme est la norme opérateur.

36.5 Un peu de points fixes

36.5.1 Choix de la fonction à point fixe

Pour l'équation $f(x) = 0$, il existe une infinité de fonctions g pour lesquelles l'équation est équivalente à $x = g(x)$.

Exemple : $f(x) = x^2 - 2 - \ln(x)$, nous pouvons faire

(1) $x = x^2 - 2 - \ln(x) + x$

(2) Poser $x^2 = 2 + \ln(x)$ et donc

$$x = -\sqrt{2 + \ln(x)} \tag{36.66a}$$
$$x = \sqrt{2 + \ln(x)}. \tag{36.66b}$$

(3) Ou encore

$$x = \frac{2 + \ln(x)}{x} \tag{36.67}$$

où nous savons déjà que $x \neq 0$ parce que $x = 0$ n'est pas dans le domaine de f.

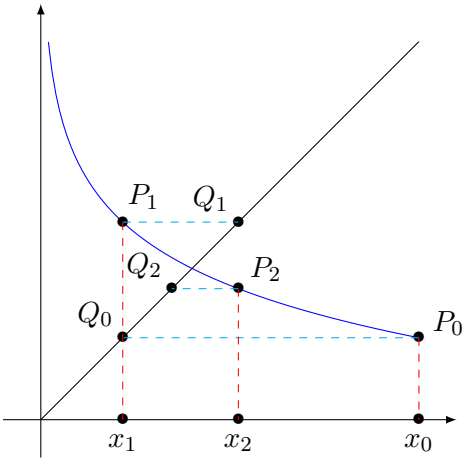
(4) Ou par l'exponentielle :

$$x = e^{x^2 - 2}. \tag{36.68}$$

Dans tous ces cas nous pouvons construire une suite (x_n) en posant un nombre arbitraire pour x_0 et ensuite la récurrence

$$x_{n+1} = g(x_n). \tag{36.69}$$

Graphiquement, la solution de l'équation est l'intersection entre les courbes $y = x$ et $y = g(x)$. Un petit dessin pour montrer la convergence :



Attention : cette méthode ne converge pas toujours. Parfois elle converge de façon monotone, et parfois pas. Le choix de la fonction g qui fait $x = g(x)$ peut énormément changer la vitesse de convergence.

Théorème 36.39 (Condition suffisante pour existence d'un point fixe)**.**
Une fonction continue $f : [a, b] \to [a, b]$ admet au moins un point fixe dans $[a, b]$.

Théorème 36.40 (Condition suffisante pour l'unicité)**.**
Soit f continue sur $[a, b]$ avec $g(x) \in [a, b]$ pour tout $x \in [a, b]$. Supposons qu'il existe $0 < k < 1$ tel que pour tout $x \in [a, b]$ nous ayons $|g'(x)| \leqslant k$ alors

(1) La fonction g possède un unique point fixe dans $[a, b]$.

(2) Pour tout $x_0 \in [a, b]$, tous les termes de la suite $x_{n+1} = g(x_n)$ sont dans $[a, b]$.

(3) Ladite suite (x_n) converge vers le point fixe.

Théorème 36.41.
Soit f continue sur $[a, b]$ avec $g(x) \in [a, b]$ pour tout $x \in [a, b]$. Supposons

(1) qu'il existe $0 < k < 1$ tel que pour tout $x \in [a, b]$ nous ayons $|g'(x)| \leqslant k$ et

(2) g est p fois dérivable sur $[a, b]$.

(3) $g'(\alpha) = g''(\alpha) = \ldots = g^{(p-1)}(\alpha)$ et $g^{(p)}(\alpha) \neq 0$ où α est l'unique point fixe.

Alors la suite (x_n) converge avec un ordre p.

Exemple 36.42
Nous reprenons

$$f(x) = x^2 - 2 - \ln(x). \tag{36.70}$$

Et nous voulons résoudre $f(x) = 0$. Graphiquement c'est l'intersection entre $y = x^2 - 2$ et $y = \ln(x)$. Il est vite tracé de savoir qu'il y a deux solutions : $\alpha_1 \in [0, 1]$ et $\alpha_2 \in [\sqrt{2}, 2]$.

Déjà un petit problème : l'intervalle $[0, 1]$ ne va pas parce que f n'y est pas continue. Un petit raffinement d'analyse nous fournit $\alpha_1 \in [e^{-2}, 1]$.

Nous avons au moins les fonctions de points fixes suivantes :

$$g_1(x) = \sqrt{2 + \ln(x)} \tag{36.71a}$$

$$g_2(x) = e^{x^2-2}. \tag{36.71b}$$

Pour la première, il y avait un $\pm$ qui a été négligé parce que nous savons que les deux solutions cherchées sont positives. Travaillons avec la première. D'abord

$$g_1'(x) = \frac{1}{2x\sqrt{2 - \ln(x)}}. \tag{36.72}$$

Nous avons $\lim_{x\to e^{-2}} g_2'(x) = +\infty$. Il ne sera donc pas possible de trouver $0 < k < 1$ tel que $|g'(x)| \leqslant k$. Tentons quand même la méthode :

$$x_0 = 0.5 \tag{36.73}$$

Il se fait que cela est plus proche de α_1 que de α_2. Mais en réalité la suite converge vers α_2.

Passons à la seconde méthode.

$$g_2'(x) = 2xe^{x^2-2}. \tag{36.74}$$

Sur l'intervalle $[e^{-2}, 1]$, g_2' est croissante et prend toutes ses valeurs dans $[e^{-2}, 1]$. Nous pouvons prouver que

$$|g_2'(x)| \leqslant 2e^{-1} < 1. \tag{36.75}$$

Donc poser $k = 2e^{-1}$ fait fonctionner la proposition. Donc quel que soit le x_0 pris dans cet intervalle, nous aurons une suite convergente vers un point fixe à l'intérieur de l'intervalle. C'est à dire convergente vers α_1.

Cela est un exemple de problème pour lequel changer de fonction g change réellement la vie. △

36.5.2 Convergence quadratique

Définition 36.43.
*Une suite (x_n) a une convergence **quadratique** vers α si elle converge vers α et s'il existe un C tel que pour tout n nous ayons*

$$\|x_{n+1} - \alpha\| \leqslant C\|x_n - \alpha\|^2. \tag{36.76}$$

Il est bien entendu possible de parler de convergence quadratique si la relation (36.76) a lieu seulement à partir d'un certain indice.

Le lemme suivant donne l'importance du choix de point de départ lorsqu'on utilise une méthode itérative dont la convergence est quadratique.

Lemme 36.44.
Soit une suite $x_n \to \alpha$ de convergence quadratique. Si $\|x_0 - \alpha\| \leqslant r$ alors

$$\|x_n - \alpha\| \leqslant \frac{1}{C}(Cr)^{2^n} \tag{36.77}$$

Démonstration. Nous pourrions directement prouver la formule (36.77) par récurrence, mais nous allons la reconstruire un peu. Nous cherchons

$$\|x_n - \alpha\| \leqslant C^{k(n)}r^{2^n}. \tag{36.78}$$

Nous avons les inégalités

$$\|x_{n+1} - \alpha\| \leqslant C\|x_n - \alpha\|^2 \tag{36.79a}$$
$$\leqslant CC^{2k(n)}r^{2^{n+1}} \tag{36.79b}$$
$$= C^{2k(n)+1}r^{2^{n+1}} \tag{36.79c}$$

d'où nous voyons que la fonction k doit vérifier

$$\begin{cases} k(0) = 0 & \text{(36.80a)} \\ k(n+1) = 2k(n) + 1 & \text{(36.80b)} \end{cases}$$

La première équation est l'hypothèse $\|x_0 - \alpha\| \leqslant r$ comparée à la formule (36.78). Il est vite vérifié que $k(n) = 2^n - 1$. D'où le résultat. □

Si le point de départ est choisi de façon à avoir $Cr < 1$ alors nous avons là un très bon majorant parce qu'il s'agit d'un majorant convergeant très rapidement vers zéro. Si au contraire $Cr > 1$ alors ce majorant ne sert à rien.

36.45.
Le fait d'avoir une convergence quadratique signifie que le nombre décimales correctes double (environ) à chaque itération, dans n'importe quelle base. En effet supposons que x_n ait k décimales correctes ; cela signifie que $|x_n-\alpha|\sim 10^{-k}$. Donc

$$|x_{n+1}-\alpha|\lessapprox M10^{-2k}. \tag{36.81}$$

Cela est le double de décimales correctes de $|x_n-\alpha|$, moins l'ordre de grandeur de M.

Pour la méthode de bisection, le nombre de décimales augmente de 1 à chaque itération, mais seulement en base 2. En base 10, de façon générique[6] il faut entre 3 et 4 itérations pour avoir une décimale de plus.

36.46 (Condition d'arrêt[403])**.**
D'autre part, lorsqu'une méthode a une convergence quadratique, nous avons un test d'arrêt. Pour ce voir, nous avons la limite

$$\lim_{n\to\infty}\frac{|x_{n+1}-\alpha|}{|x_n-\alpha|}\leqslant\lim_{n\to\infty}\frac{C|x_n-\alpha|^2}{|x_n-\alpha|}=0. \tag{36.82}$$

Cette limite est alors également valable sans les valeurs absolues et si nous soustrayons $x_n-\alpha$ au numérateur, la limite devient -1 :

$$-1=\lim_{n\to\infty}\frac{x_{n+1}-x_n}{x_n-\alpha}. \tag{36.83}$$

Ou encore

$$\lim_{n\to\infty}\frac{x_n-x_{n+1}}{x_n-\alpha}=1. \tag{36.84}$$

Cela a pour conséquence que si n est grand,

(1) x_{n+1} a le même ordre de grandeur que $x_n-\alpha$.

(2) x_n-x_{n+1} et $x_n-\alpha$ ont le même signe.

Donc si nous voulons une approximation de α avec une erreur ϵ, il suffit d'arrêter le calcul lorsque $|x_{n+1}-x_n|\leqslant\epsilon$. Et ce faisant nous savons de plus si l'approximation est par excès ou par défaut.

36.5.3 Convergence

Proposition 36.47 (Convergence d'une méthode de point fixe[403])**.**
Soit $g:\mathbb{R}\to\mathbb{R}$ de classe C^1 et α un point fixe attractif[7] *de g. Soit k tel que $|g'(\alpha)|<k<1$ et δ tel que $\|g'\|_{B(\alpha,\delta)}<k$.*

Alors

(1) La fonction g est k-contractante[8] *sur $B(\alpha,\delta)$.*

(2) Nous avons $g\big(B(\alpha,\delta)\big)\subset B(\alpha,\delta)$.

(3) Pour tout $x_0\in B(\alpha,\delta)$ la suite $x_{n+1}=g(x_n)$ converge vers α et

$$|x_n-\alpha|\leqslant|x_0-\alpha|k^n. \tag{36.85}$$

Si de plus $g'(\alpha)=0$ et g est de classe C^2 alors nous avons convergence quadratique (définition 36.43).

Démonstration. Vu que α est un point fixe attractif de g nous pouvons considérer un k tel que $|g'(\alpha)|<k<1$. Et comme g est de classe C^1, la fonction g' est continue et donc bornée sur toute boule du type $\overline{B(\alpha,\delta)}$. Soit δ le plus grand nombre tel que $\|g'\|_{\overline{B(\alpha,\delta)}}\leqslant k$. Nous notons $I=\overline{B(\alpha,\delta)}$ pour cette valeur de δ.

6. C'est à dire sauf coup de malchance ou coup de chance.
7. Définition 19.39.
8. Définition 19.42

Pour tout $x \in I$ nous avons, en utilisant le théorème des accroissements finis 14.136(2) :

$$\begin{aligned} |g(x) - \alpha| &= |g(x) - g(\alpha)| && \text{(36.86a)}\\ &\leqslant \sup_{t\in I} |g'(t)||x - \alpha| && \text{(36.86b)}\\ &\leqslant k|x - \alpha| && \text{(36.86c)}\\ &< \delta && \text{(36.86d)} \end{aligned}$$

parce que $k < 1$ et $|x - \alpha| \leqslant \delta$. Par conséquent $g(x) \in B(\alpha, \delta)$. Cela prouve le point (2). Pour le point (1), soient $x, y \in B(\alpha, \delta)$ et

$$|g(x) - g(y)| \leqslant \sup_{a\in I} |g'(a)||x - y| \leqslant k|x - y|. \tag{36.87}$$

Pout le point (3) nous avons $|g(x_n) - \alpha| \leqslant k|x_n - \alpha|$, c'est à dire

$$|x_{n+1} - \alpha| \leqslant k|x_n - \alpha|. \tag{36.88}$$

Le résultat annoncé s'obtient par récurrence sur n.

En ce qui concerne la convergence quadratique, c'est du Taylor (proposition 14.287). Développons $g(x_n)$ autour de $g(\alpha)$:

$$g(x_n) = g(\alpha) + g'(\alpha)(x_n - \alpha) + \frac{1}{2}(x_n - \alpha)^2\epsilon(x_n - \alpha) \tag{36.89}$$

avec $\lim_{t\to 0} \epsilon(t) = 0$. En posant $C = \frac{1}{2}\sup_{t<\delta} |\epsilon(t)|$ nous avons $|g(x_n) - g(\alpha)| \leqslant C|x_n - \alpha|^2$, c'est à dire

$$|x_{n+1} - \alpha| \leqslant C|x_n - \alpha|^2. \tag{36.90}$$

□

Ce corollaire est une paraphrase de la proposition 36.47. Il en retient seulement les points intéressants en pratique.

Corollaire 36.48.
Soit α une solution de l'équation $x = g(x)$, avec g continue sur un voisinage de α et dérivable dans l'intérieur. Nous supposons que

$$|g'(\alpha)| < 1. \tag{36.91}$$

Alors il existe un rayon δ tel que si $x_0 \in B(\alpha, \delta)$, la suite (x_n) converge vers α.

Certes cette proposition demande moins d'hypothèses, mais en réalité, il ne donne pas de vrais moyens de choisir un point de départ x_0. Avec les deux théorèmes précédents, nous pouvions prendre x_0 n'importe où dans $[a, b]$. Le fait est que pour choisir x_0 nous pouvons tracer et donner à la main un x_0 proche de ce qui semble être α. Si ça ne converge pas, il faut donner un x_0 plus proche. La proposition nous assure que si nous jouons bien à choisir x_0 très proche, la suite finira par converger.

Notons que le corollaire 36.48 a encore l'inconvénient de demander de calculer $g'(\alpha)$ alors que α est inconnu. La résolution de l'inéquation $|g'(x)| < 1$ nous donne un certain nombre d'intervalles dans $\mathbb{R}$.

Soient I_n les intervalles solutions de l'inéquation. Si $\alpha \in I_n$ alors la méthode converge. Sinon, c'est pas garantit. En tout cas nous ne devons pas savoir réellement α pour appliquer le théorème. Il suffit de savoir que α est dans un des I_n.

36.6 Méthode de Newton

L'objectif de la méthode de Newton est d'évaluer une racine α de l'équation $f(x) = 0$ lorsque nous avons déjà une approximation x_0 de la racine α.

Définition 36.49.
Le nombre α est une ***racine simple*** *de l'équation $f(x) = 0$ si $f(\alpha) = 0$ et $f'(\alpha) \neq 0$. Le nombre α est une* ***racine multiple*** *d'ordre r de $f(x) = 0$ si*

$$f(\alpha) = f'(\alpha) = \ldots = f^{(r-1)(\alpha)} = 0 \tag{36.92}$$

et $f^{(r)}(\alpha) \neq 0$.

Exemple 36.50
La fonction $x \mapsto x^3$ en $x = 0$ est un racine d'ordre 3. △

36.6.1 « Justification » par la formule par Taylor

Soit une fonction f continue et dérivable sur $[a, b]$. Soit α une racine de f et x_n une de ses approximations. Nous notons l'erreur θ et nous avons $\alpha = x_n + \theta$. Du coup nous avons $f(x_n + \theta) = f(\alpha) = 0$.

Écrivons la série de Taylor du théorème 14.277 autour de x_n : il existe une fonction $\epsilon\colon \mathbb{R} \to \mathbb{R}$ telle que $\lim_{t\to 0} \epsilon(t) = 0$ telle que

$$f(\alpha) = f(x_n + \theta) = f(x_n) + \theta f'(x_n) + \frac{\theta^2}{2}\epsilon(\theta). \tag{36.93}$$

Nous isolons le θ du terme d'ordre 1 en nous souvenant que le membre de gauche est nul :

$$\theta = -\frac{f(x_n) - \theta^2\epsilon(\theta)}{f'(x_n)} \tag{36.94}$$

Vu que $\alpha = x_n + \theta$, nous pouvons écrire

$$\alpha = x_n - \frac{f(x_n) + \theta^2\epsilon(\theta)}{f'(x_n)}. \tag{36.95}$$

Il est donc raisonnable de poser

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{36.96}$$

en espérant que cela soit une meilleur approximation de α que x_n.

En tout cas l'erreur sur x_{n+1} est

$$\alpha - x_{n+1} = x_n + \theta - x_n + \frac{f(x_n) + \theta^2\epsilon(\theta)}{f'(x_n)} = \theta + \frac{f(x_n) + \theta^2\epsilon(\theta)}{f'(n_n)}, \tag{36.97}$$

qui ne doit pas être fondamentalement plus grand que θ dès que θ est petit, surtout que si x_n est une approximation de α, nous pouvons espérer que $f(x_n)$ soit également petit. Là où les choses peuvent déraper en grand, c'est si $f'(x_n)$ est petit.

Cette méthode de Newton ne converge pas toujours. Le pire est lorsque par malheur il y a une bosse pas loin de la racine. Alors il y a un risque de tomber sur $f'(x_{n+1}) = 0$ ou en tout cas très proche de zéro. Dans ce cas le point x_{n+2} est envoyé très loin.

36.6.2 « Justification » par points fixes

Nous savons que pour résoudre $f(x) = 0$ par une méthode de point fixe, il y a de nombreux choix possibles de fonctions g telles que $g(x) = x$ donne la même solution que $f(x) = 0$. Soit α une solution de $f(x) = 0$ et cherchons une fonction g de la forme

$$g(x) = x - kf(x). \tag{36.98}$$

Nous savons par la proposition 36.47 que la fonction g donne une convergence quadratique lorsque $g'(\alpha)=0$. Pour la forme (36.98) nous avons $g'(\alpha)=1-kf'(\alpha)$, ce qui nous donne l'idée de poser $k=\frac{1}{f'(\alpha)}$.

Le fait est que $f'(\alpha)$ n'est pas connu, mais nous pouvons l'approximer par $f'(x)$ lorsque x est proche de α. D'où l'idée de considérer la fonction

$$g(x)=x-\frac{f(x)}{f'(x)}, \tag{36.99}$$

et donc la suite $x_{n+1}=g(x_n)$ c'est à dire

$$x_{n+1}=x-\frac{f(x_n)}{f'(x_n)}. \tag{36.100}$$

Dès que x_n est proche de α, sous l'hypothèse (raisonnable par continuité) que $f'(x_n)$ soit proche de $f'(\alpha)$, la méthode devrait donner une convergence quadratique.

Remarque 36.51.
Cette justification par points fixes n'est pas vraiment différente de celle par Taylor parce que Taylor est utilisé dans la preuve de la proposition 36.47.

Définition 36.52 (Méthode de Newton)**.**
*La **méthode de Newton** pour la fonction f est la suite définie par récurrence*

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}. \tag{36.101}$$

Cette définition ne précise pas la valeur de x_0, ni de condition d'arrêt.

36.6.3 Convergence de la méthode de Newton

Théorème 36.53 (Convergence quadratique de la méthode de Newton[403])**.**
Soit f une fonction continue vérifiant $f(\alpha)=0$ et $f'(\alpha)\neq 0$. Nous considérons la fonction

$$g(x)=x-\frac{f(x)}{g'(x)} \tag{36.102}$$

que nous supposons être de classe C^2.

Si C est une majoration de $\|g''\|$ sur un intervalle contenant α, alors en posant $\delta=1/C$ nous avons

(1) La boule $B(\alpha,\delta)$ est préservée par g : $g\big(B(\alpha,\delta)\big)\subset B(\alpha,\delta)$.

(2) Pour tout $x_0\in B(\alpha,\delta)$ nous avons convergence quadratique vers α de la suite définie par $x_{n+1}=g(x_n)$.

(3) Nous avons l'estimation

$$|x_n-\alpha|\leqslant\frac{1}{C}\big(C|x_0-\alpha|\big)^{2^n} \tag{36.103}$$

où C est la constante de la définition de convergence quadratique.

Démonstration. Nous commençons par calculer la dérivée de g :

$$g'(x)=-\frac{f(x)f''(x)}{f'(x)^2}, \tag{36.104}$$

d'où nous déduisons que $g'(\alpha)=0$. Ensuite nous utilisons abondamment la formule des accroissements finis (théorème 14.248) en commençant par

$$|g(t)-g(\alpha)|\leqslant\|g'\|_{[t,\alpha]}|t-\alpha| \tag{36.105}$$

où par $\|f\|_A$ nous entendons la norme uniforme de f sur A, c'est à dire $\|f\|_A = \sup_{x\in A}\|f(x)\|$. Note : nous écrivons $[t,\alpha]$, mais ça pourrait être $[\alpha,t]$.

Si $x \in [t,\alpha]$ alors

$$\begin{aligned}|g'(x)| &= |g'(x) - g'(\alpha)| && (36.106a)\\ &\leqslant \|g''\|_{[x,\alpha]}|x-\alpha| && (36.106b)\\ &\leqslant \|g''\|_{[x,\alpha]}|t-\alpha| && (36.106c)\\ &\leqslant \|g''\|_{[t,\alpha]}|t-\alpha|. && (36.106d)\end{aligned}$$

En particulier, $\|g'\|_{[t,\alpha]} \leqslant \|g''\|_{[t,\alpha]}|t-\alpha|$, et nous pouvons continuer les majorations (36.105) :

$$|g(t) - g(\alpha)| \leqslant \|g''\|_{[t,\alpha]}|t-\alpha|^2. \tag{36.107}$$

La fonction g étant de classe C^2, la dérivée seconde g'' est bornée (nous supposons déjà travailler sur un compact contenant α). Soit C une borne. Nous sommes en mesure de prouver le point (1) avec $\delta = 1/C$. En effet si $t \in B(\alpha, 1/C)$ alors

$$|g(t) - \alpha| = |g(t) - g(\alpha)| \leqslant C|t-\alpha|^2 \leqslant C\frac{1}{C^2} = \frac{1}{C}, \tag{36.108}$$

ce qui prouve que $g(t) \in B(\alpha, 1/C)$.

Le point (2) se prouve de la même manière : si $x_n \in B(\alpha, 1/C)$ alors

$$|x_{n+1} - \alpha| = |g(x_n) - g(\alpha)| \leqslant C|x_n - \alpha|^2, \tag{36.109}$$

ce qui est bien la convergence quadratique.

La majoration du point (3) s'obtient par récurrence sur n. Pour $n = 0$, la relation (36.103) devient $|x_0 - \alpha| \leqslant |x_0 - \alpha|$ qui est vraie. Ensuite par la convergence quadratique et la récurrence,

$$|x_{n+1} - \alpha| \leqslant C|x_n - \alpha|^2 \leqslant C\Big[\frac{1}{C}(C|x_0-\alpha|)^{2^n}\Big]^2 = \frac{1}{C}\big[M|x_0-\alpha|\big]^{2^{n+1}}. \tag{36.110}$$

□

36.54.
Dans le cas pratiques, nous commençons souvent par résoudre l'équation $f(x) = 0$ par dichotomie. Au moment où nous sommes assez proche de la solution nous commençons Newton.

La raison est que la dichotomie fonctionne toujours : nous allons toujours nous approcher de la solution. Si par contre le point de départ est mal choisit, la méthode de Newton peut envoyer n'importe où, y compris très loin de la solution.

La proposition suivante nous indique que dans le cas d'une fonction convexe, le choix de point de départ de la méthode de Newton n'est pas tellement crucial parce que il sont tous bons. De plus la convergence se faisant de façon décroissante (si on part de la droite), nous savons que le résultat sera une approximation par excès de α.

Proposition 36.55 (Newton dans le cas convexe)**.**
Soit f de classe C^2 et une racine α telle que $f'(\alpha) > 0$. Soit $b > \alpha$ tel que f soit convexe sur $[\alpha, b]$.

Alors pour tout $x_0 \in [\alpha, b]$ la suite de la méthode de Newton est

(1) décroissante

(2) reste dans $[\alpha, b]$

(3) converge vers α.

Démonstration. Nous savons par la proposition 19.90(2) que la fonction f' est croissante, et par hypothèse $f'(\alpha) > 0$, donc sur $[\alpha, b]$ nous avons $f' > 0$. Par conséquent, nous avons aussi $f > 0$ sur $[\alpha, b]$.

Le graphe de f est au dessus de la tangente de f en $x = x_n$ (proposition 19.96). Si nous nommons t_x la fonction qui donne la tangente en x nous avons $t_{x_n}(\alpha) < 0$ parce que $f(\alpha) = 0$. Par conséquent

$$t_{x_n}(x) = 0 \tag{36.111}$$

pour $\alpha < x < x_n$. Cela prouve que $x_{n+1} \in [\alpha, b]$, et que (x_n) est une suite décroissante

Étant donné que (x_n) est une suite décroissante dans le compact $[\alpha, b]$, elle est convergente. Notons β sa limite. Nous avons la relation de récurrence

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{36.112}$$

En passant à la limite $n \to \infty$ nous avons l'équation

$$\beta = \beta - \frac{f(\beta)}{f'(\beta)}. \tag{36.113}$$

Vu que $f(x) > 0$ sur $]\alpha, b]$ nous avons automatiquement $\beta = \alpha$. □

36.6.4 Formalisation de l'algorithme

La méthode de Newton consiste a exprimer la solution x de $f(x) = 0$ avec $f \in C^1(\mathbb{R})$ comme limite d'une suite $\{x_n\}_{n\in\mathbb{N}}$ définie par récurrence par la formule

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{36.114}$$

où x_0 est arbitraire.

Si on veut exprimer cela en terme d'algorithme, nous disons que l'algorithme de Newton est donné par la suite de problèmes

$$F_n(x_{n+1}, x_n, f) = x_{n+1} - x_n + \frac{f(x_n)}{f'(x_n)}. \tag{36.115}$$

La donnée du problème est la fonction f, et rien que elle.

Plus précisément, une fois que la fonction f est donnée, il existe une infinité de problèmes : pour chaque $a \in \mathbb{R}$ nous avons le problème

$$G_a(x_n, f) = x - a + \frac{f(a)}{f'(a)}. \tag{36.116}$$

La méthode de Newton consiste à sélectionner une partie de ces problèmes de la façon suivante :

$$\begin{cases} F_0 = G_{x_0} & \text{(36.117a)} \\ F_n = G_{x_n}. & \text{(36.117b)} \end{cases}$$

Le problème F_0 fournit un nombre x_1 qui nous permet de sélectionner le problème G_{x_1} qui va fournir le nombre x_2, etc.

Au moment de calculer le conditionnement de F_n, nous ne devons pas voir x_{n-1} comme fonction de x_0 et de la donnée f. Il ne faut donc pas dériver à travers les x_n.

Proposition 36.56.
Si une racine est multiple, alors l'ordre de convergence de la méthode de Newton est 1.

Voici un algorithme possible :

```
1 def Newton(f,x0,toll,maxit):
2     fp=f.derivative()
3     n=0
```

```
    x=x0
    diff=toll+1
    while abs(diff)>toll and n<maxit:
        n=n+1
        diff=-f(x)/fp(x)
        x=x+diff
    return x,n
```

tex/frido/codeSnip_2.py

Commentaires :

(1) Notons que dans un langage vraiment numérique comme Matlab, il faut passer f' en argument.

(2) Dans le `while` il faudrait mettre $x_{n+1} - x_n$ (en valeur absolue), mais cette différence est aussi utilisée pour calculer x_{n+1} donc on la calcule une seule fois.

(3) Il faudrait faire une vérification sur $f(x_n) \neq 0$. Il n'y a pas tellement de choix que de changer le point initial.

36.6.5 Caractéristiques

L'algorithme de Newton a les caractéristiques suivantes :

(1) Pour résoudre le problème numéro n, il faut avoir résolu le problème numéro $n-1$.

(2) Aucune des solutions x_n aux problèmes intermédiaires n'est une solution au problème de départ (à moins d'un coup de chance).

(3) Étant donné que la donnée du problème F_n est la fonction f de départ, nous avons $d_m = d_n = d$ pour tout m et n.

Théorème 36.57.
Soit f continue sur un voisinage de α, racine simple. Alors il existe un voisinage de α de rayon σ tel que pour tout x_0 dans ce voisinage, la méthode converge vers α avec ordre de convergence $p = 2$.

Donc dès qu'on a continuité autour de la solution recherchée, il suffit de prendre x_0 assez proche pour que tout se passe bien. Cela se fait par localisation des racines, par exemples en traçant la fonction avec un bon niveau de zoom. Le fait est qu'on cherche disons 3 décimales à la main (travail sur ordinateur et graphique) et Newton donne les 20 décimales suivantes à la vitesse de la lumière.

36.6.6 Exemple de la racine carrée

Nous allons nous lancer dans un exemple : le cas de la racine carrée. Soit à calculer une approximation numérique de $\sqrt{2}$. Il s'agit d'une racine de la fonction $f(x) = x^2 + 2$. La fonction de la méthode de Newton associée est :

$$g(x) = x - \frac{f(x)}{f'(x)} = \frac{x^2 - 2}{2x}. \tag{36.118}$$

Cherchons un intervalle autour de $\sqrt{2}$ sur lequel nous avons convergence de la méthode de Newton. Cela s'obtient grace à la proposition 36.47 qui nous informe qu'il suffit de trouver un intervalle autour de $\sqrt{2}$ sur lequel $|g'(x)| \leqslant 1$.

Nous avons

$$g'(x) = \frac{x^2 - 2}{2x^2}, \tag{36.119}$$

et nous cherchons à résoudre $|g'(x)| \leqslant 1$. D'abord $g'(x) = 1$ n'a aucune solutions alors que $g'(\sqrt{2}) = 0$. Donc nous avons $g'(x) \leqslant 1$ pour tout $x \in \mathbb{R}^+$. Par contre l'équation $g'(x) = -1$ a des solutions : $x = \pm\sqrt{2/3}$.

Nous avons donc convergence de la méthode de Newton pour x_0 dans un intervalle de la forme

$$[\sqrt{2/3}, \sqrt{2} + \ldots] \tag{36.120}$$

où les les trois points représentent l'expression qu'il faut pour que ce soit symétrique autour de $\sqrt{2}$. La valeur précise n'a pas tellement d'importance parce, vu que nous sommes en train de chercher $\sqrt{2}$, il est peu probable que nous ayons déjà en main une bonne approximation de nombres du type $\sqrt{2/3}$.

Proposition 36.58.
La méthode de Newton pour la fonction $f(x) = x^2 - 2$ converge vers $\sqrt{2}$ pour toute valeur de départ dans $]0, +\infty[$.

Démonstration. La fonction $f(x) = x^2 - 2$ est convexe et $f'(\sqrt{2}) = 2\sqrt{2} > 0$. Donc la méthode converge vers $\sqrt{2}$ pour tout $x_0 \geqslant \sqrt{2}$ par la proposition 36.55.

Si par contre $x_0 \in]0, \sqrt{2}[$ nous avons

$$x_1 = \frac{x_0^2 + 2}{2x_0}. \tag{36.121}$$

En posant $h(x) = (x^2 + 2)/2x$ et en résolvant $h'(x) = 0$ nous trouvons $x = \sqrt{2}$. Et là, $h(\sqrt{2}) = \sqrt{2}$. Donc $h(x)$ est toujours plus grand que $\sqrt{2}$ pour tout $x \in]0, \sqrt{2}[$.

En d'autres termes, si $x_0 \in]0, \sqrt{2}[$ alors $x_1 \geqslant \sqrt{2}$ et nous retombons dans le premier cas. □

36.6.7 Si multiplicité

Supposons que α soit de multiplicité r (définition 36.49).

Cela se remarque en voyant que la méthode de Newton demande plutôt 20 itérations que 5. Le problème que cela pose est que chaque itération, les évaluations provoquent des erreurs. Donc moins d'itérations, c'est mieux.

Nous pouvons modifier la formule avec

$$x_{n+1} = x_n - r\frac{f(x_n)}{f'(x_n)}. \tag{36.122}$$

Il est possible de prouver que cette suite est à nouveau à convergence quadratique.

Ou alors on pose $F(x) = f^{(r-1)}(x)$ et α est une racine simple pour F. Donc faire Newton pour F est à nouveau quadratique, tout en donnant la même solution parce que $F(\alpha) = 0$ et $F'(\alpha) \neq 0$.

La seconde façon est bien parce que le théorème de localisation fonctionne 36.57

Et si r n'est pas connu ?

Il est toujours possible de faire $r = 2$ puis $r = 3$ et caetera jusqu'au moment où l'on remarque que le nombre d'itérations baisse un grand coup.

Mais ça demande beaucoup de calculs. Le mieux est de changer de méthode.

36.6.8 Et la dérivée ?

Un des problèmes de la méthode de Newton est que l'on doit pouvoir calculer la dérivée. Typiquement, il faut savoir f de façon analytique. Si cela n'est pas possible, nous pouvons changer de méthode et utiliser la méthode des sécantes décrite en 36.9.

36.6.9 Méthode de Newton : le cas général

Lemme 36.59.
Soient A et B deux matrices inversibles telles que la matrice $(A + \epsilon B)$ soit inversible pour tout ϵ assez petit. Alors il existe une matrice $X(\epsilon)$ telle que

$$(A + \epsilon B)^{-1} = (A^{-1} + \epsilon X) \tag{36.123}$$

et telle que $\lim_{\epsilon \to 0} X(\epsilon) = -A^{-1}BA^{-1}$.

Démonstration. Le candidat matrice X est relativement simple à trouver en écrivant

$$(A+\epsilon B)(A^{-1}+\epsilon X)=\mathbb{1}+\epsilon AX+\epsilon BA^{-1}+\epsilon^2 BX. \tag{36.124}$$

En imposant que cela soit $\mathbb{1}$, nous trouvons

$$X(\epsilon)=-(A+\epsilon B)^{-1}BA^{-1}. \tag{36.125}$$

La matrice $X(\epsilon)$ étant un inverse à droite de $(A+\epsilon B)$, son déterminant est non nul et X est inversible. Par conséquent elle est également inversible au sens usuel. Le calcul de la limite est direct :

$$\lim_{\epsilon\to 0}-(A+\epsilon B)^{-1}BA^{-1}=A^{-1}BA^{-1} \tag{36.126}$$

parce que l'inverse est une fonction continue sur $\mathbb{M}(n,\mathbb{R})$. □

Remarque 36.60.
Un calcul naïf nous permet de trouver le même résultat de façon plus heuristique. En effet un développement usuel (dans $\mathbb{R}$) est

$$\frac{1}{a+\epsilon b}=\frac{1}{a}-\frac{\epsilon b}{a^2}+\dots \tag{36.127}$$

Si nous récrivons cela avec des matrices, nous écrivons (attention : passage heuristique !) :

$$(A+\epsilon B)^{-1}=A^{-1}-\epsilon A^{-1}BA^{-1}+\dots \tag{36.128}$$

Notons le choix de généraliser b/a^2 par $a^{-1}ba^{-1}$. Dans les réels les deux écritures sont équivalentes, mais pas dans les matrices.

Étudions si $A^{-1}-\epsilon A^{-1}BA^{-1}$ est bien un inverse à ϵ^2 près de $(A+\epsilon B)$:

$$(A+\epsilon B)(A^{-1}+\epsilon A^{-1}BA^{-1})=1-\epsilon BA^{-1}+\epsilon BA^{-1}-\epsilon^2 BA^{-1}BA^{-1}=1-\epsilon^2 BA^{-1}BA^{-1}. \tag{36.129}$$

Par conséquent, à des termes en ϵ^2 près la matrice $A^{-1}-\epsilon A^{-1}BA^{-1}$ est bien un inverse de $A+\epsilon B$.

Théorème 36.61 (Méthode de Newton[404])**.**
Soit $f\colon\mathbb{R}^n\to\mathbb{R}^n$ une application de classe C^2 et un point $a\in\mathbb{R}^n$ tel que $f(a)=0$. Nous supposons que df_a est inversible.

Alors il existe un voisinage V de a tel que pour tout $x_0\in V$ la suite définie par récurrence

$$x_{n+1}=x_n-(df_a)^{-1}\big(f(x_n)\big) \tag{36.130}$$

converge vers a. De plus la vitesse est quadratique au sens où il existe $C>1$ tel que

$$\|x_n-a\|\leqslant C^{-1-2^n}. \tag{36.131}$$

Démonstration. Étant donné que df_a est inversible et que df est continue, l'application df_x est continue[9] pour tout x dans un voisinage de a. Nous prenons $r>0$ tel que df_x est inversible pour tout $x\in B(a,r)$.

Nous considérons la fonction

$$\begin{aligned}F\colon B(a,r)&\to\mathbb{R}^n\\ x&\mapsto x-(df_x)^{-1}\big(f(x)\big).\end{aligned} \tag{36.132}$$

Cela est une application C^1. La clef est de montrer que l'application de F à un point $a+h$ rapproche de a pourvu que h soit assez petit. Nous avons la formule suivante :

$$F(a+h)-F(a)=h-\big(df_{a+h}\big)^{-1}\big(f(a+h)\big). \tag{36.133}$$

9. Nous pouvons voir df comme l'application qui à x fait correspondre la matrice $df_x\in\mathbb{M}(n,\mathbb{R})$. Cette application étant continue et la non inversibilité d'une matrice étant donnée par l'annulation du déterminant, les matrices inversibles forment un ouvert dans l'ensemble des matrices.

Nous allons maintenant utiliser un développement de Taylor par rapport à h en suivant la formule (14.750). Nous avons

$$f(a+h) = f(a) + df_a(h) + \|h\|^2\xi(h) \tag{36.134}$$

où $\xi\colon \mathbb{R}^n \to \mathbb{R}^n$ est une fonction qui tend vers une constante lorsque $h \to 0$. Nous avons aussi

$$df_{a+h} = df_a + \|h\|\tau(h) \tag{36.135}$$

où $\tau\colon \mathbb{R}^n \to \mathbb{M}(n,\mathbb{R})$ est une application qui tend vers une constante lorsque $h \to 0$. En ce qui concerne l'inverse nous utilisons le lemme [10] 36.59 :

$$\big(df_a + \|h\|\tau(h)\big)^{-1} = (df_a)^{-1} + \|h\|A(h) \tag{36.136}$$

où A est une autre matrice fonction de h qui tend vers une constante lorsque h tend vers zéro. En substituant le tout dans (36.133) nous trouvons

$$F(a+h) - F(a) = \|h\|^2(df_a)^{-1}\xi(h) + \|h\|\big(A(h)\circ df_a\big)(h) + \|h\|^3A(h)\xi(h). \tag{36.137}$$

En ce qui concerne la norme nous utilisons le fait que si T est un opérateur, $\|Tx\| \leqslant \|T\|\|x\|$. Nous trouvons

$$\|F(a+h) - F(a)\| \leqslant \|h\|^2\|(df_a)^{-1}\|\|\xi(h)\| + \|h\|^2\|A(h)\circ df_a\| + \|h\|^3\|A(h)\|\|\xi(h)\| \tag{36.138a}$$

$$= \|h\|^2\alpha(h) \tag{36.138b}$$

pour une certaine fonction $\alpha\colon \mathbb{R}^n \to \mathbb{R}$ qui tend vers une constante lorsque $h \to 0$.

En posant $C = \lim_{h\to 0}\alpha(h)$ nous avons la majoration

$$\|F(x) - a\| \leqslant C\|x-a\|^2. \tag{36.139}$$

Nous pouvons également supposer que $C > 1$. Afin de prouver la vitesse de convergence (36.131), nous allons encore redéfinir r en demandant $r < 1/C^2$. De cette manière nous avons

$$\|x_0 - a\| \leqslant \frac{1}{C^2} \tag{36.140}$$

et la récurrence sur n est :

$$\|x_{n+1} - a\| = \|F(x_n) - a\| \leqslant C\|x_n - a\|^2 \leqslant C\big(C^{-1-2^n}\big)^2 = C^{-1-2^{n+1}}. \tag{36.141}$$

Note : ce dernier calcul est le lemme 36.44 appliqué à $r = (1/C^2)$. □

Remarque 36.62.
La valeur de la constante C a été fixée par l'équation (36.139). Certes nous pouvons toujours choisir C plus grand affin d'augmenter la vitesse de convergence, mais le point de départ x_0 devant être dans une boule de taille $1/C^2$ autour de a, demander C plus grand revient à demander un point de départ plus précis.

36.7 Estimation de l'ordre de convergence

Définition 36.63 ([405]).
*Nous disons que la suite (x_n) de limite x est **convergente d'ordre** q pour $q > 1$ s'il existe $\mu > 0$ tel que*

$$\lim_{k\to\infty} \frac{|x_{k+1} - x|}{|x_k - x|^q} = \mu. \tag{36.142}$$

En particulier :

10. Pour l'inversibilité de $\|h\|\tau(h)$, notons que df_a est inversible et que par hypothèse la somme $df_a + \|h\|\tau(h)$ est inversible.

— *la convergence d'ordre 2 est dite quadratique,*
— *la convergence d'ordre 3 est dite cubique,*
— *la convergence d'ordre 4 est dite quartique.*

Comment estimer numériquement l'ordre p de convergence de la méthode ? Soit une suite (x_n) convergente vers α. Considérons les 4 termes x_{n-3}, x_{n-2}, x_{n-1}, x_n. Alors nous pouvons écrire l'approximation

$$\frac{|x_n - x_{n-1}|}{|x_{n-1} - x_{n-2}|} \simeq \left(\frac{|x_{n-1} - x_{n-2}|}{|x_{n-1} - x_{n-3}|} \right)^p. \tag{36.143}$$

Cette approximation ne serait pas trop mauvaise tant que n est assez grand pour que la convergence soit bien engagée. Passons au logarithme :

$$\ln \frac{|x_n - x_{n-1}|}{|x_{n-1} - x_{n-2}|} \simeq p \ln \left(\frac{|x_{n-1} - x_{n-2}|}{|x_{n-1} - x_{n-3}|} \right). \tag{36.144}$$

et donc

$$p \simeq \frac{\ln \left(\frac{|x_n - x_{n-1}|}{|x_{n-1} - x_{n-2}|} \right)}{\ln \left(\frac{|x_{n-1} - x_{n-2}|}{|x_{n-1} - x_{n-3}|} \right)}. \tag{36.145}$$

Avec cette approximation, en réalité nous calculons une suite (p_i) qui sont les approximations de p à partir des termes i à $i+3$ de la suite (x_n). Il s'agit d'une suite d'estimations de p.

(1) Dans le cas de la bisection, nous obtenons toujours $p_i = 1$.

(2) Dans le cas de la méthode de Newton (36.6) nous avons $p = 2$. Mais les premières valeurs de p_i peuvent être aussi bien 0 que 7. Après quelques itérations pourtant les p_i se regroupent autour de 2.

En tout cas, le plus important est de savoir si $p > 1$ ou non. Rappel : nous voulons la superlinéarité parce que nous voulons utiliser le test d'arrêt de la différence entre deux termes, voir 36.46.

36.8 Autres méthodes

36.8.1 Méthode de Schröder

La formule est

$$x_{n+1} = x_n - \frac{f(x_n) f'(x_n)}{f'(x_n)^2 - f(x_n) f''(x_n)} \tag{36.146}$$

Cette méthode est d'ordre 2 pour toute racine et toute valeur de multiplicité. Le problème de cette méthode est qu'elle demande 3 évaluations de f. Son efficacité :

$$E = \sqrt[3]{2} \simeq 1.25 \tag{36.147}$$

Cela est donc moins efficace que Newton.

36.8.2 Halley

Il a $p = 3$ lorsque α est racine simple. Mais encore $p = 1$ pour les racines multiples. Plus efficace que Newton pour les racines simples, mais même problème pour les racines multiples.

$$x_{n+1} = x_n - \frac{2 f(x_n) f'(x_n)}{2 f'(x_n)^2 - f(x_n) f''(x_n)} \tag{36.148}$$

36.9 Méthode des sécantes variables

Si nous n'avons pas de formule analytique pour f, mais seulement la possibilité de calculer $f(x)$ pour tout x. Newton ne fonctionne pas, mais la bisection fonctionne.

Nous pouvons approximer

$$f'(x_n) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}. \tag{36.149}$$

En substituant dans la formule de Newton, nous obtenons

$$x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}. \tag{36.150}$$

Il s'agit de prendre la droite qui passe par $(x_{n-1}, f(x_{n-1}))$ et par $(x_n, f(x_n))$ et de prendre l'intersection de cette droite avec l'axe $y = 0$. Cela donne le x_{n+1}.

Pour cette méthode, il ne faut pas seulement x_0 mais également x_1.

L'ordre de convergence est le nombre d'or

$$p = \frac{1 - \sqrt{5}}{2} \simeq 1.618. \tag{36.151}$$

Cela est donc superlinéaire.

La nombre d'évaluations est $s = 1$ (il y a deux apparitions de f dans la formule, mais l'une des deux est récupérée dans l'itération suivante). Donc l'efficacité est

$$E = p. \tag{36.152}$$

Donc bien efficace.

Proposition 36.64.
Si α est racine simple, il existe un voisinage de α tel que pour tout choix de x_0, x_1 dans ce voisinage, la méthode converge.

Psychologiquement, on est tenté de prendre x_0 et x_1 de part et d'autre de α (pensant à la bisection), mais en réalité ce n'est pas obligatoire du tout et n'a aucune influence. Il faut seulement les prendre très proches de α.

Remarque 36.65.
La méthode de la sécante est souvent écrite sous la forme

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}. \tag{36.153}$$

C'est évidemment algébriquement équivalent.

Les formules (36.151) et (36.153) ont toutes deux des erreurs de cancellation. Laquelle est la plus grave ?

Dans la première, si la fraction est mal calculée, elle ne fait que modifier x_n. C'est à dire qu'on peut espérer qu'à la prochaine itération, ça aille mieux. En tout cas, dans ce cas si la fraction est mal calculée, ça ne détruit pas tout.

Dans la seconde, c'est la valeur elle-même qui risque d'être mal calculée. Et si la fraction est mal calculée, alors on casse complètement l'éventuel bonne approximation que nous avions déjà.

36.9.1 Aitken

La méthode du Δ^2 de Aitken est une méthode d'accélération de la convergence.

Soit (x_n) une suite qui converge. Nous voudrions une nouvelle suite (y_n) telle que

$$\lim_{n\to\infty} \frac{y_n - \alpha}{x_n - \alpha} \tag{36.154}$$

C'est la définition d'une convergence accélérée.

La façon de faire est :

$$y_n = \frac{x_{n+2}x_n - x_{n+1}^2}{x_{n+2} - 2x_{n+1} + x_n} = x_n - \frac{(x_{n+1} - x_n)}{x_{n+2} - 2x_{x+1} + x_n}. \tag{36.155}$$

La première expressions a deux cancellations (la seconde une seule) et de plus la première est y_n elle-même alors que la seconde est une correction.

Donc la seconde expression est numériquement meilleure.

L'opérateur Δ appliqué à une suite est :

$$(\Delta x)_n = x_{n+1} - x_n \tag{36.156}$$

Donc

$$(\Delta^2 x)_n = (\Delta x)_{n+1} - (\Delta x)_n = x_{n+2} - x_{n+1} - x_{n+1} + x_n = x_{n+2} - 2x_{n+1} + x_n. \tag{36.157}$$

L'accélération a alors la formule

$$y_n = \frac{(\Delta x)_n^2}{(\Delta^2 x)_n}. \tag{36.158}$$

Le problème est que ça accélère tellement que l'on arrive vite à des erreurs de cancellations, et donc à une précision en pics oscillants.

36.10 Équations algébrique

C'est une équation du type $P(x) = 0$ où P est un polynôme. Soit un polynôme de degré n. Nous en savons des choses.

(1) L'équation a exactement n solutions dans $\mathbb{C}$ en comptant les multiplicités.

(2) Les racines complexes arrivent par paire complexes conjuguée. Elles sont donc toujours en nombre pair.

Si donc nous avons $n = 3$, nous ne pouvons pas avoir 2 racine réelles. Il y en a donc 1 ou 3 réelles. Pas zéro ni deux.

Quelques méthodes : Müller, matrice compagnon, Laguerre.

36.10.1 Résoudre un système linéaire

Pour résoudre un système linéaire d'équations, nous échelonnons la matrice du système. Soit à résoudre le système $Ax = b$ où

$$A = \begin{pmatrix} 2 & 4 & -6 \\ 1 & 5 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \quad \text{et} \quad b = \begin{pmatrix} -4 \\ 10 \\ 5 \end{pmatrix}. \tag{36.159}$$

En termes de problèmes, on écrit $F\big(x, (A, b)\big) = Ax - b$. La donnée de ce problème est le couple (A, b).

En ce qui concerne l'algorithme, on pose comme premier problème

$$F_1\big(x_1, (A_1, b_1)\big) = A_1 x_1 - b_1 = 0 \tag{36.160}$$

avec $A_1 = A$ et $b_1 = b$.

Ensuite, on commence à échelonner et le second problème est

$$F_2\big(x_2, (A_2, b_2)\big) = A_2 x_2 - b_2 = 0 \tag{36.161}$$

avec

$$A = \begin{pmatrix} 2 & 4 & -6 \\ 0 & 3 & 6 \\ 0 & 1 & 5 \end{pmatrix}, \quad \text{et} \quad b = \begin{pmatrix} -4 \\ 12 \\ 13 \end{pmatrix}. \tag{36.162}$$

Le troisième problème sera

$$F_3\big(x_3, (A_3, b_3)\big) = A_3 x_3 - b_3 = 0 \tag{36.163}$$

avec

$$A = \begin{pmatrix} 2 & 4 & -6 \\ 0 & 3 & 6 \\ 0 & 0 & 3 \end{pmatrix}, \quad \text{et} \quad b = \begin{pmatrix} -4 \\ 12 \\ 3 \end{pmatrix}. \tag{36.164}$$

Ce problème est facile à résoudre « à la main ». Nous nous arrêtons donc ici avec l'algorithme, et nous trouvons le x_3 qui résous le problème F_3.

36.10.2 Caractéristiques

L'algorithme de résolution de systèmes linéaires d'équations a les propriétés suivantes, à mettre en contraste avec celles de Newton :

(1) Pour résoudre le problème numéro n, il n'a pas fallu résoudre le problème numéro $n-1$.

(2) Toutes les solutions x_n des problèmes intermédiaires sont solutions du problème de départ. Nous avons $F_n(x, d_n) = 0$ pour tout n (ici, $d_n = (A_n, b_n)$).

(3) D'un problème à l'autre, les données changent énormément : la matrice échelonnée peut être très différente de la matrice de départ.

36.10.3 Définitions

Nous allons maintenant formaliser en donnant quelques définitions pour nommer les propriétés que nous avons vues. D'abord, un algorithme est une suite de problèmes. Un **algorithme** pour résoudre un problème $F(x, d) = 0$ est une suite de problèmes $\{F_n(x_n, d_n) = 0\}_{n\in\mathbb{N}}$.

Définition 36.66.
Un tel algorithme est dit ***fortement consistant*** *si pour toutes données admissibles d_n, on a*

$$F_n(x, d_n) = 0 \quad \forall\, n, \tag{36.165}$$

où x est la solution de $F(x, d) = 0$.

L'algorithme des matrices est fortement consistant, mais pas l'algorithme de Newton.

Définition 36.67.
Un algorithme est ***consistant*** *si $\lim_{n\to\infty} F_n(x, d_n) = 0$.*

Dans le cas de l'algorithme de Newton, c'est plutôt une telle consistance qu'on attend.

L'algorithme est dit **stable** si pour tout n le problème correspondant est stable. Dans ce cas, on note K^{num} le **conditionnement relatif asymptotique** défini par

$$K^{\text{num}} = \limsup_n K_n \tag{36.166}$$

où K_n est le conditionnement relatif du problème $F_n(x_n, d_n) = 0$.

Définition 36.68.
Un algorithme est dit ***convergeant*** *(en d) si pour tout $\epsilon > 0$, il existe $N = N(\epsilon)$ et $\delta = \delta(N, \epsilon)$ tels que pour $n \geqslant 0$ et $|d - d_n| < \delta$, on ait $|x(d) - x_n(d_n)| < \epsilon$.*

Remarque 36.69.
Dans le cas de l'algorithme de Newton, nous avons vu que la donnée d_n du problème F_n était en fait la même que la donnée initiale d, donc nous avons $d_n = d$, et par conséquent nous avons toujours $|d - d_n| < \delta$. Dans ce cas, la définition de la convergence revient à demander que la suite numérique des x_n converge vers la solution x.

Remarque 36.70.
Dans le cas des matrices par contre, les données sont très différentes les unes des autres, nous avons donc en général que $|d-d_n| > \delta$. Mais en revanche nous savons que tous les problèmes intermédiaires F_n acceptent une solution unique[11] $x_n(d_n) = x(d)$. Par conséquent, $|x_n(d_n) - x(d)|$ est toujours plus petit que ϵ. L'algorithme des matrices est donc toujours un algorithme convergeant.

36.11 Équations non linéaire

Certains équations non linéaires sont résoluble explicitement, par exemples les polynômes de degré jusqu'à 4 ou des choses comme

$$\sin^2(x) + 3\sin(x) + 5 = 0. \tag{36.167}$$

Mais ces exemples sont très rares.

Nous allons étudier des équations du type $f(x) = 0$, dans $\mathbb{R}$.

(1) Un problème écrit sous la forme $x = g(x)$ peut utiliser des théorèmes de points fixes.
(2) Un problème sous la forme $f(x) = 0$ peut utiliser des méthodes de bisection, Newton ou autres.

Il y a évidemment beaucoup de façons de transformer un problème pour passer d'une forme à l'autre.

Exemple 36.71
Soit $f(x) = x^2 - a = 0$ avec $a > 0$. Nous pouvons l'écrire

$$x^2 + x - a = x \tag{36.168}$$

qui donne une forme $g(x) = x$ pour $g(x) = x^2 + x - a$.

Ou encore $x = \frac{a}{x}$ et donc $g(x) = a/x$ (si par ailleurs on sait que $x \neq 0$). Notons que $x \neq 0$ n'est pas une hypothèse très forte parce qu'on la vérifie directement sur a. △

Exemple 36.72
Soit l'équation à résoudre

$$f(x) = x^2 - 2 - \ln(x) = 0 \tag{36.169}$$

Les solutions de cette équations peuvent être vues comme les intersections avec l'axe X du graphe $y = x^2 - 2 - \ln(x)$. Tracer peut donc aider. Par ailleurs, il faut noter que

$$\lim_{x \to \pm\infty} f(x) = \infty, \tag{36.170}$$

donc les solutions sont certainement contenues dans un compact de $\mathbb{R}$.

À part tracer nous pouvons écrire

$$x^2 - 2 = \ln(x). \tag{36.171}$$

Et là, ce sont deux fonctions dont nous pouvons tracer le graphe pour trouver graphiquement les points d'intersection. Une étude de fonction montre vite qu'il y a exactement deux solutions, qu'elles sont strictement positives. Pour trouver des bornes, il faut calculer par exemple pour $x = 2$ les valeurs de $\ln(x)$ et $x^2 - 2$ pour voir si le graphe de $x^2 - 2$ est déjà plus haut. △

La majorité des méthodes numériques de résolution d'équation du type $f(x) = 0$ ou $x = g(x)$ seront sous la forme de suites. Avec questions à la clefs :

11. Nous n'envisageons que le cas où le déterminant est non nul.

(1) Quel point de départ choisir ?

(2) Convergence ?

(3) Est-ce que la limite est bien une solution ?

(4) Vu que la limite est unique, comment faire si l'équation a plusieurs solutions ? (souvent c'est le choix du point initial qui va jouer sur ce point)

36.73.
Si la fonction est très plate, il est possible d'avoir

$$|f(\tilde{\alpha})| \leqslant \epsilon \tag{36.172}$$

sans que $\tilde{\alpha}$ ne soit une bonne approximation.

Lorsqu'on fait tourner une méthode itérative résolvant $f(x) = 0$, il n'est pas suffisant de s'arrêter lorsque

$$f(x_n) \leqslant \epsilon_1. \tag{36.173}$$

Il faut aussi s'assurer que, si $\bar{x}$ est la solution exacte, $|x_n - \bar{x}| \leqslant \epsilon_2$. Ici ϵ_1 et ϵ_2 sont deux « précisions » que nous nous fixons au départ.

Évidemment, vérifier la condition $|x_n - \bar{x}| \leqslant \epsilon_2$, il faudrait savoir $\bar{x}$. Et savoir $\bar{x}$ c'est justement le problème. Nous sommes donc amenés à faire des estimations de $|x_n - \bar{x}|$.

36.74.
Lorsque nous effectuons une méthode itérative, il faut donc contrôler deux grandeurs :

$$|\bar{x} - x_n| \leqslant \epsilon_1 \tag{36.174a}$$

$$|x_{n+1} - x_n| \leqslant \epsilon_2. \tag{36.174b}$$

Proposition 36.75.
Soit p l'ordre de convergence de la suite (x_n) vers $\bar{x}$. Si $p > 1$ et $|x_{n+1} - x_n| \leqslant \epsilon_2$ alors $|\bar{x} - x_n| \leqslant \epsilon_2$.

36.11.1 Méthode de bisection

Il y a ce théorème des valeurs intermédiaires.

Théorème 36.76.
Soit f continue sur $[a, b]$ telle que $f(a)f(b) < 0$. Alors il existe au moins une solution à l'équation $f(x) = 0$ sur l'intervalle $]a, b[$.

Pour démarrer une bisection, il est toujours bon de prendre l'intervalle $[a, b]$ de façon à ne contenir qu'une seule solution.

Soit donc un premier intervalle $[a_0, b_O]$ tel que $f(a_0)f(b_0) < 0$ et ne contenant qu'une seule solution. À chaque itération nous considérons la moitié de l'intervalle précédent, mais la moitié contenant la solution.

Le test d'arrêt de la méthode de bisection se base uniquement sur la taille de l'intervalle qui reste. En effet si nous avons

$$|b_n - a_n| \leqslant \epsilon \tag{36.175}$$

nous avons certainement

$$|\bar{x} - x_n| \leqslant \frac{\epsilon}{2} \tag{36.176}$$

où x_n est le point du milieu de $[a_n, b_n]$.

36.77.
La fonction f n'intervient dans la méthode que via son signe, pas via ses valeurs exactes.

36.78.
Notons que le théorème des valeurs intermédiaires n'est pas très puissant pour choisir l'intervalle de départ ; penser à la fonction

$$f(x) = x^2 - 5 \tag{36.177}$$

sur l'intervalle $[-10, 10]$. Il y a bien deux solutions dans l'intervalle, mais elles sont invisibles du théorème des valeurs intermédiaires. La fonction $x \mapsto x^2$ a sa solution en $x = 0$, mais elle aussi n'est pas visible.

36.79.
Certes la méthodes de bisection assure la convergence vers une solution, mais elle n'assure pas la convergence monotone. Il peut arriver que $|\bar{x} - x_n| < |\bar{x} - x_{n+1}|$. C'est le cas lorsque la solution est très proche du milieu de l'intervalle choisit. Le x_0 est alors proche de $\bar{x}$ alors que x_1 sera à une distance de $\bar{x}$ d'environ un quart de l'intervalle de départ.

Supposons déjà avoir trouvé un intervalle $[a, b]$ dans lequel se trouve une unique solution à $f(x) = 0$. Voici un algorithme possible.

```
1  from __future__ import division
2
3  def bissection(f, a, b, toll, mmax):
4      """
5      f : une fonction
6      a,b : les limites de l'intervalle
7      toll : la tolérance. C'est l'amplitude de l'intervalle à ↩
           partir duquel nous nous arrêtons.
8      nmax : le nombre maximum d'itérations.
9
10     Nous supposons que \( b>a\).
11
12     Retourne un tuple (x,n) où 'x' est la solution approchée et 'n↩
           ' est le nombre d'itérations effectuées
13     """
14     n = -1
15     amp = toll + 1 # Pour s'assurer que l'on entre dans le cycle
16     while amp > toll and n < nmax :
17         n = n + 1
18         amp=abs(b - a)
19         x = a + amp / 2
20         if f(a) * f(x) < 0:
21             b = x
22         elif f(a)*f(x) > 0:
23             a = x
24         else :       # Problème ZERO
25             amp = 0
26     return (x, n)
```

tex/frido/codeSnip_1.py

Plusieurs remarques :

(1) Le fait de retourner le nombre d'itérations effectuées permet à l'utilisateur de savoir la précision et si le nombre maximum d'itérations est dépassé. Si ce `n` retourné est égal à `nmax`, l'utilisateur sait que le `x` retourné n'est pas fiable.

(2) La ligne `from __future__ import division` fait en sorte que l'opération `/` est bien la division usuelle. Sinon, le défaut en python 2 est que `/` soit la division *entière*, c'est à dire

que $1/2 = 0$ en python 2. En python 3, le symbole `/` désigne bien la division usuelle, mais Sage utilise Python 2.

(3) Même si l'intervalle $[a, b]$ contient plus d'une solution, la méthode fonctionne et donne une solution. Il est simplement éventuellement très compliqué de savoir laquelle.

(4) Nous faisons `amp=toll+1` parce que nous voulons absolument lancer le cycle au moins une fois. Sinon, le `x` à retourner ne serait pas définit au moment de sortir du cycle (si le cycle n'est pas exécuté).

(5) Calculer le point milieu d'un intervalle $[a, b]$ est par la formule $(a + b)/2$ sauf que cette opération est numériquement dangereuse parce qu'à cause de l'arithmétique en précision finie, il est possible que cela tombe *exactement* sur a ou b. D'où le fait de calculer le point milieu par

$$x = a + \frac{amp}{2}. \tag{36.178}$$

(6) Dans le cas `Problème ZERO` nous déduisons $f(x) = 0$. Attention que c'est pas que $f(x) = 0$ mais simplement que en mettant x dans f, la *machine* retourne son zéro.

Il peut cependant avoir une fonction telle que $f(1) = 10^{-50}$ et $f(2) = 0$. L'algorithme de bisection risque de s'arrêter si $x_n = 1$. Parce que la machine risque de calculer $f(x_n) = 0$.

Quoi qu'il en soit, nous y mettons `amp=0` pour être sûr de sortir de la boucle dès la prochaine vérification.

(7) Il y a moyen de sauver les valeurs de $f(a)$ et $f(x)$ pour ne pas les recalculer, et en particulier au moment de faire `b=x` nous pouvons poser `fa=fx`.

Si τ est la précision de la solution voulue, nous pouvons fixer a priori le nombre d'itérations à faire grâce à la formule

$$n \geqslant \left\lceil \log_2 \left(\frac{b-a}{\tau}\right) \right\rceil. \tag{36.179}$$

Il y a un "$\geqslant$" et non une égalité parce qu'en arithmétique numérique, le nombre obtenu à droite pourrait ne pas être le bon à 1 près.

Ici pour $\nu \in \mathbb{R}$ le nombre $\lceil \nu \rceil$ est le plus petit entier à être plus grand ou égal à ν.

36.80.
Notons l'importance de la continuité de f. Par exemple que ferait la bisection sur la fonction $f(x) = 1/x$ pour l'intervalle $[-3, 1]$?

Il y a changement de signe sans avoir de racine.

Vu que 2^{10} est déjà 1024. Donc si on veut de la précision de l'ordre de 1/1000, dix itérations suffisent. Si donc nous avons besoin de 200 itérations pour atteindre la précision voulue, c'est l'occasion de trouver un intervalle plus petit. Par exemple en traçant la fonction, en faisant un zoom et en trouvant des valeurs de a et b qui sont déjà proches.

36.81.
Dans le monde réel, il arrive souvent d'utiliser une méthode de bisection pour se donner un point de départ pour une autre méthode.

36.12 Efficacité

Définition 36.82.
*L'**efficacité** est le nombre*

$$E = \sqrt[s]{p} \tag{36.180}$$

où p est l'ordre de convergence de la méthode et s est le nombre de fois qu'il faut calculer une valeur de la fonction à chaque itération (nous ne comptons pas l'initialisation).

Que le nombre de d'évaluations de f intervienne est logique parce que chaque évaluation provoque une erreur possible.

Exemple 36.83(Bisection)
Pour la méthode de bisection, nous avons $s = 1$ parce que chercher x_{n+1}, il faut seulement calculer $f(x_n)$. △

Exemple 36.84(Newton)
Pour l'algorithme de Newton nous avons $p = 2$ et il y a deux évaluations à chaque itération (une fois f et une fois f'), donc $s = 2$ et $E = \sqrt{2}$. △

36.13 Exemples sous forme d'exercices

Exemple 36.85

Nous supposons une machine acceptant 5 chiffres significatifs. Elle retient les nombres sous la forme $\pm 0.x \cdots \times 10^{\cdots}$.
$x = 1.403$, $y = 0.4112 \times 10^{-3}$, $z = -0.4111 \times 10^{-3}$.
Soient $a = (x \oplus y) \oplus z$ et $b = (y \oplus z) \oplus x$.

(1) D'abord la calculer à la main.

(2) Quel est le calcul préférable ?

(3) Donner l'erreur relative avec 3 chiffres significatifs.

Dans le cas du calcul à la main, il faut en faire un seul parce que, algébriquement, $a = b$.
Nous avons $x = 1.403$, $y = 0.0004112$ et $z = -0.0004111$. Et la somme donne :

$$a = b = 1.4030001 = 0.14030001 \times 10^1. \tag{36.181}$$

Faisons d'abord la normalisation de x, c'est à dire $\text{fl}(x)$.

$$\text{fl}(x) = 0.1403 \times 10^1. \tag{36.182}$$

et y est déjà normalisé :

$$\text{fl}(y) = 0.4112 \times 10^{-3}. \tag{36.183}$$

Il n'y a pas d'erreurs d'assignation pour ces deux nombres.

Pour faire la somme, il faudra déjà un peu casser les nombres pour les écrire de façon à pouvoir les sommer. En effet, il faut écrire les deux nombres avec le même exposant de 10 (le plus grand), pour pouvoir les mettre en colonne :

$$0.1404 \times 10^1 \to 0.1404 \times 10^1 \tag{36.184a}$$

$$0.4112 \times 10^{-3} \to 0.00004 \times 10^1. \tag{36.184b}$$

La somme donne 0.14034×10^1. Et ça, c'est à nouveau arrondi. Le premier chiffre supprimé est un 4, donc

$$x \oplus y = 0.1403 \times 10^1. \tag{36.185}$$

Et là on remarque que nous avons la même chose que x. C'est un classique du calcul numérique.

Nous avons aussi

$$\text{fl}(z) = 0.4111 \times 10^{-3}. \tag{36.186}$$

Et pour faire la somme de cela avec $x \oplus y$ nous devons le remettre sous la forme d'un 10^1 :

$$\text{fl}(z) \to -0.00004 \times 10^1 \tag{36.187}$$

(erreur de conversion), et en sommant on trouve

$$(x \oplus y) \oplus z = 0.140216 \times 10^1, \tag{36.188}$$

qui est encore arrondi. Le premier chiffre supprimé est un 6, donc

$$\text{fl}(a) = 0.1403 \times 10^1, \tag{36.189}$$

Le nom de l'erreur qui consiste à avoir $x \oplus y = x$ est "relation annirmale".

Calculons b.

Les nombres y et z ont même ordre de grandeur, donc pas d'erreur au moment de les mettre sous forme sommable.

$$\text{fl}(x) + \text{fl}(y) = 0.00010 \times 10^{-3}. \tag{36.190a}$$

Cela est renormalisé et arrondi : $\text{fl}(x) \oplus \text{fl}(y) = 0.1000 \times 10^{-6}$.

Notons que nous avons ici commis potentiellement une erreur de cancellation parce que entre y et z, il y a 3 chiffres sur 4 qui sont identiques. Seul le chiffre 1 est significatif en réalité.

Il faut maintenant ajouter x à cela. D'abord

$$\text{fl}(x) = 0.1403 \times 10^1. \tag{36.191}$$

Pour cette somme, il faudra remettre notre 0.1000×10^{-6} avec une puissance 10^1. Et là, nous obtenons zéro parce que vraiment ce nombre est trop petit pour être écrit avec 10^1. Résultat des courses :

$$\text{fl}(b) = 0.1403 \times 10^1. \tag{36.192}$$

Dans le premier calcul nous avons deux "relations anormales" et dans le second nous en avons une plus une cancellation.

Nous préférons avoir deux relations anormales, parce que l'erreur de cancellation est plus grave : elle consiste à une perte de chiffre significatifs. Le fait est que faisant la différence à l'ordinateur nous avons obtenu 0.1 qui est certes exact, mais qui est un coup de bol : la différence aurait aussi bien pu être 0.19 avec d'autres nombres, machinement égaux.

Note : avec les données ici, il n'y a en fait pas d'erreur de cancellation. Mais il y a une erreur potentielle de cancellation, potentiellement grave.

En ce qui concerne l'erreur relative. Dans la formule

$$\epsilon_r = \frac{|a - a^*|}{|a|}, \tag{36.193}$$

la différence ne peut pas être calculée à la calculatrice justement parce qu'elle est très potentiellement sujette à erreur de cancellation.

$$\epsilon_r = \frac{|0.1030001 \times 10^1 - 0.1403 \times 10^1|}{0.14030001 \times 10^1} = \frac{0.1 \times 10^{-6}}{0.14030001 \times 10^1} \simeq 0.712758 \times 10^{-7}. \tag{36.194}$$

En passant à 3 chiffres significatifs, 0.713×10^{-7} (le premier chiffre supprimé est un 7).

△

Exemple 36.86

Soient $x = 0.1 \times 10^{21}$ et $y = 0.5 \times 10^20$ et les expressions

(1) $z_1 = \frac{x-y}{y} + \frac{x+y}{x}$

(2) $z_2 = \frac{x^2+y^2}{xy}$.

Ces deux expressions sont algébriquement équivalentes.

(1) Calculer les valeurs.

(2) On suppose une machine en précision simple. Laquelle des deux expressions est préférable ?

Pour z_1, en arithmétique exacte :

$$z_1 = \frac{0.5 \times 10^{20}}{0.5 \times 10^{20}} + \frac{1.5 \times 10^{20}}{1 \times 10^{20}} = 0.25 \times 10^1. \tag{36.195}$$

Le calcul exact de z_2 donne la même chose.

Calcul de z_1 Les deux valeurs sont mémorisables et la différence $x - y$ se fait sans erreurs de cancellation. Idem pour la somme $x + y$. Idem pour les divisions.

Calcul de z_2 Pour faire x^2, c'est pas possible parce que c'est de l'ordre de 10^{40} alors que nous sommes en précision simple. Idem pour le produit xy.

Morale : z_2 donne un `overflow` alors que z_1 fonctionne de façon exacte.

Remarque 36.87.
En réalité le z_1 n'est pas tout à fait calculable de façon exacte sur la machine parce qu'elle doit d'abord convertir en binaire, ce qui n'est pas toujours possible. Mais sur notre machine qui fonctionne en base 10, il n'y a pas de problèmes.

△

Exemple 36.88

Soit la fonction

$$f(x) = 2x^2 - 4x + 2 - e^{-x}. \tag{36.196}$$

(1) Identifier la plus grande des solutions réelles de $f(x) = 0$.

(2) Effectuer une bisection pour la savoir.

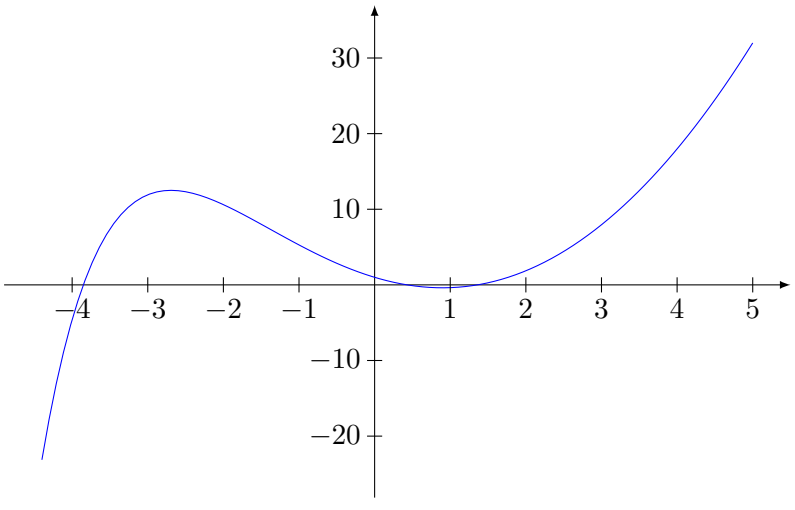
(3) Sachant que

$$\alpha \simeq 1.358500220734946, \tag{36.197}$$

quelle est l'erreur relative ?

Note : si c'est pour chercher à la main des approximations pour démarrer, il est évidemment préférable de dessiner $f_1(x) = 2x^2 - 4x + 2$ et $f_2(x) = -e^{-x}$ séparément.

Quoi qu'il en soit, voici un graphique :



Nous voyons trois racines : $\alpha_1 \in [0, 0.5,]$, $\alpha_2 \in [1, 1.5]$ et $\alpha_0 \in [-4, -3.5]$.

La plus grande solution est α_2. Nous pouvons déjà remplir le tableau des précisions :

n	a_n	b_n	x_n	$g(x_n)$	$\lvert b_n - a_n \rvert$
$< ++ >$	$< ++ >$	$< ++ >$	$< ++ >$	$< ++ >$	0.5
$< ++ >$	$< ++ >$	$< ++ >$	$< ++ >$	$< ++ >$	0.25
$< ++ >$	$< ++ >$	$< ++ >$	$< ++ >$	$< ++ >$	0.125

Et nous calculons les valeurs de f aux points d'extrémité de l'intervalle. Note que seul le signe nous importe :

$$f(1) \simeq -0.368 \tag{36.198a}$$
$$f(1.5) \simeq 0.278. \tag{36.198b}$$

Voici donc le tableau avec le signe de f indiqué :

n	a_n	b_n	x_n	$g(x_n)$	$\lvert b_n - a_n \rvert$
0	$1(+)$	$1.5(-)$	$< ++ >$	-0.162×10^0	0.5
1	$< ++ >$	$< ++ >$	$< ++ >$	$< ++ >$	0.25
2	$< ++ >$	$< ++ >$	$< ++ >$	$< ++ >$	0.125

Puis :

n	a_n	b_n	x_n	$g(x_n)$	$\lvert b_n - a_n \rvert$
0	$1(+)$	$1.5(-)$	$1.25(-)$	-0.162×10^0	0.5
1	$< ++ >$	$< ++ >$	$< ++ >$	$< ++ >$	0.25
2	$< ++ >$	$< ++ >$	$< ++ >$	$< ++ >$	0.125

Et enfin :

n	a_n	b_n	x_n	$g(x_n)$	$\lvert b_n - a_n \rvert$
0	$1(+)$	$1.5(-)$	$1.25(-)$	-0.162×10^0	0.5
1	$1.25(-)$	$1.5(+)$	$1.375(+)$	$+0.284 \times 10^{-1}$	0.25
2	$1.25(-)$	$1.375(+)$	$1.3125(-)$	$)0.738 \times 10^{-1}$	0.125

Note que les $f(x_n)$ restent toujours du même ordre de grandeur. Si un moment on voit un 1.6×10^7, c'est qu'une erreur a été commise.

En ce qui concerne le calcul de l'erreur relative, la première chose à faire est de vérifier que le α proposé est dans l'intervalle qui nous reste. Sinon c'est qu'une erreur a été commise.

De plus notre approximation est $x_n = 1.3125$, dont déjà deux chiffres sont corrects. En deux itérations de bisection en partant de 0.5, nous ne pouvons pas nous attendre à mieux.

△

36.14 Approximations de fonctions

(1) D'habitude on n'approxime pas une fonction sur tout son domaine, mais seulement sur une partie.

(2) Il y a le problème du choix de la classe des fonctions qui vont approximer. Nous allons travailler avec des polynômes.

(3) Il nous faut un critère disant si une approximation est bonne ou non.

36.14.1 Critère d'interpolation

À partir de $n+1$ abscisses points distinctes x_i, nous calculons $y_i = f(x_i)$. Il y a ce théorème qui dit qu'il existe un unique polynôme de degré (au plus) $n+1$ passant par les points (x_i, y_i).

Proposition-définition 36.89 (Base de Lagrange)**.**
Étant donnés $n+1$ valeurs distinctes x_i, l'espace des polynômes de degré n admet la base

$$L_i^{(n)}(x) = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} \tag{36.199}$$

pour $i = 0, \dots, n$.

Soit par exemple les valeurs de f données dans

x_i	$y_i = f(x_i)$
$x_0 = 5$	1
$x_1 = -7$	-23
$x_2 = 6$	-54
$x_3 = 0$	-954

Les polynômes de Lagrange pour ces données dépendent seulement des x_i, pas des y_i. En particulier,

$$L_0^{(3)}(x) = \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)}, \tag{36.200}$$

etc.

La réponse est que

$$L_0^{(3)}(x) = \frac{x^3 + 13x^2 + 42x}{660} \tag{36.201a}$$

$$L_1^{(3)}(x) = \frac{-x^3 - x^2 + 30x}{84} \tag{36.201b}$$

$$etc. \tag{36.201c}$$

Ce qu'il y a de bien avec cette base est que en posant $a_i = f(x_i)$ alors le polynôme

$$\sum_{i=0}^{n} a_i L_i^{(n)}(x) \tag{36.202}$$

passe par les points $(x_i, f(x_i))$. Du coup il suffit d'écrire

$$P_3(x) = L_0^{(3)}(x) - 23L_1^{(n)}(x) - 54L_2^{(n)}(x) - 954L_3^{(n)}(x) = 4x^3 + 35x^2 - 84x - 954. \tag{36.203}$$

Un inconvénient de cette base est qu'elle est complètement dépendante des points choisis. Si on ajoute un point ou qu'on en prend un à peine différent, tous les coefficients changent. Mais en pratique, ajouter des points est quelque chose qui arrive souvent parce que souvent, après avoir vu le résultat d'un polynôme d'interpolation, on veut ajouter un point pour avoir un meilleur résultat.

36.90.
Une habitude : le premier et le dernier nœud se choisissent aux extrémités de l'intervalle sur lequel nous voulons une approximation.

Le but d'une approximation est d'avoir des approximations de $f(x^*)$ pour des valeurs de x^* qui ne soit pas une des abscisses données (parce que sur ces points, le polynôme et la fonction sont égaux). Nous considérons donc

$$f(x^*) \simeq P_n(x^*). \tag{36.204}$$

Si x^* est dans l'intervalle $I = [x_{min}, x_{max}]$ alors nous disons que nous calculons f par **interpolation**. Si au contraire x^* est en dehors de cet intervalle nous parlons d'**extrapolation**.

Si x^* est pris à l'extérieur de I, alors l'erreur risque d'être très grande, surtout parce que les polynômes tendent tous vers $\pm\infty$ lorsque $x \to \pm\infty$.

Autant l'interpolation via polynômes est le plus souvent valable, il faut garder à l'esprit que les extrapolations sont souvent mauvaises si x^* est trop loin des extrémités de I.

36.14.2 Base de Newton

Après la base canonique et la base de Lagrange, nous voyons la base de Newton. Soient encore $n+1$ points donnés du graphe de f.

Définition 36.91.
*La **base de Newton** pour les abscisses x_i est l'ensemble des polynômes suivants :*

$$1, (x-x_0), (x-x_0)(x-x_1), \dots, (x-x_0)(x-x_1)\dots(x-x_{n-1}) \tag{36.205}$$

Notons que ces polynômes n'utilisent pas le dernier point des x_i. Le polynôme passant par les points est

$$P_n(x) = x_0 + \sum_{i=1}^{n} c_i \prod_{j=0}^{n-1}(x-x_j). \tag{36.206}$$

Le calcul des c_i n'est pas absolument évident. Mais si nous ajoutons un point d'interpolation, les polynômes déjà calculés sont encore bons ; en particulier

$$P_{n+1}(x) = P_n(x) + c_{n+1}\prod_{j=0}^{n}(x-x_j). \tag{36.207}$$

Et cela est bien, parce que ça donne une façon de les calculer par récurrence.

Il y a plusieurs façons de calculer les c_i.

Les différences divisées sont des façons d'approximer les dérivées.

Définition 36.92.
*Soient $n+1$ nœuds x_i pour la fonction f. La **différence divisée** sont :*

Ordre 0
$$f[x_i] = f(x_i) \tag{36.208}$$

Ordre 1
$$f[x_i, x_j] = \frac{f[x_i] - f[x_j]}{x_i - x_j}. \tag{36.209}$$

Ordre 2
$$f[x_i, x_j, x_k] = \frac{f[x_i, x_j] - f[x_j, x_k]}{x_i - x_k}. \tag{36.210}$$

Ordre n
$$f[x_0, \dots, x_n] = \frac{f[x_0, \dots, x_{n-1}] - f[x_1, \dots, x_n]}{x_0 - x_n}. \tag{36.211}$$

Les ordres font référence à l'ordre de dérivation qui est approximé.

Nous avons alors

$$c_i = f[x_0, \dots, x_i]. \tag{36.212}$$

Cela donne effectivement une méthode de récurrence pour trouver les coefficients c_i.

Remarque 36.93.
Pour calculer c_0, il faut seulement calculer $f[x_0] = f(x_0)$. Mais pour calculer c_1 il faut $f[x_0]$ et $f[x_1]$. Et pour c_2 il faut $f[x_0, x_1, x_2]$ qui demande $f[x_0, x_1]$ et $f[x_0, x_1]$, qui demande etc.

Il faut donc calculer en réalité tous les $f[x_i]$ pour terminer le calcul. Par contre, pour ajouter un point, il ne faut pas tout recalculer, et même pas tout conserver en mémoire. Il faut seulement garder en mémoire la dernière diagonale.

Exemple 36.94

Soit les nœuds

x	$f(x)$
3	1
1	-3
5	2
6	4

Trouver le polynôme d'interpolation via la base de Newton.

x_i	$f[x_i]$	$f[x_i, x_j]$	$f[x_i, x_j, x_k]$
3	1	$\frac{1+3}{2} = 2$	$\frac{f[x_0,x_1]-f[x_1,x_2]}{x_0-x_2} = \frac{2-\frac{5}{4}}{-2} = -\frac{3}{8}$
1	-3	$\frac{-3-2}{-4} = \frac{5}{4}$	$\frac{\frac{5}{4}-2}{-5} = \frac{3}{20}$
5	2	$\frac{2-4}{-1} = 2$	$\frac{f[x_0,x_1,x_2]-f[x_1,x_2,x_3]}{x_0-x_3} = \frac{4}{40}$
6	4	$<++>$	$<++>$

Le polynôme d'interpolation sera

$$P_3(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + c_3(x - x_0)(x - x_1)(x - x_2). \tag{36.213}$$

$\triangle$ <++>

Un exercice typique serait de donner tout pour 3 points puis de demander le polynôme qui aurait un quatrième point.

36.14.3 Méthode des minima quadratiques

Soient $m + 1$ points connus sur le graphe de la fonction f que nous devons approximer. Au lieu d'exiger que notre approximation ne passe par tous les points, nous allons chercher une approximation qui minimise la somme des carrés des erreurs sur ces points.

Soit $\mathcal{F}$ une classe de fonctions dans laquelle nous allons chercher l'approximation. Nous cherchons $g \in \mathcal{F}$ qui minimise

$$E(g) = \sum_{i=0}^{m} \big(f(x_i) - g(x_i)\big)^2 \omega_i \tag{36.214}$$

où $\omega_i > 0$ est une pondération. Souvent on prend $\omega_i = 1$, mais pas toujours. La fonction E sur $\mathcal{F}$ est la **fonction d'erreur**.

36.95.
À part dans les exercices à la main, le nombre de points est grand, du type du milliard. Il est bien entendu pas envisageable de faire passer un polynôme *exactement* par un milliard de points, parce que cela demanderait un polynôme de degré un milliard.

Plus généralement, d'un point de vue scientifique, avoir n paramètres libres pour n données expérimentales, ça ne passe pas Popper.

Afin de faire de la science qui passe Popper nous nous restreignons à une classe de fonction $\mathcal{F}$ dont la dimension n'est pas grande : $\dim(\mathcal{F}) \ll m$. Et nous notons $\dim(\mathcal{F}) = n + 1$.

Exemple 36.96
La qualité d'une expérience peut être influencée par des paramètres extérieurs comme l'humidité, le vent, etc. Donc il est normal d'avoir des expériences moins précises que d'autres. On le pèse moins. $\triangle$

Exemple 36.97
Dans un questionnaire, il se met des questions volontairement contradictoires. Si quelqu'un répond

« oui » aux deux questions, il y a une indication que la personne a répondu un peu n'importe comment, et il faut moins peser ses réponses. △

Soit $g \in \mathcal{F}$, et une base $\{g_i\}_{i=0,\dots,n}$ de $\mathcal{F}$. Nous écrivons

$$g = a_0 g_0 + a_1 g_1 + \cdots + a_n g_n \tag{36.215}$$

La fonction donnée E donnée en (36.214), est, à partir du moment où $\mathcal{F}$ et une base sont choisis, une fonction des paramètres a_i que nous nommons $F(a_0, \dots, a_i)$. Il faut minimiser F, c'est à dire poser

$$\frac{\partial F}{\partial a_j} = 0 \tag{36.216}$$

pour $j = 0, \dots, n$. Cela sont $n+1$ équations pour $n+1$ inconnues. Notons que ces équations sont linéaires parce que chacun des termes est du type

$$\left(f(x_i) - \sum_{j=0}^{m} a_j g_j(x_i) \right)^2, \tag{36.217}$$

et lors de la dérivation par rapport à a_j, nous obtenons du degré 1.

36.14.4 Notre espace de Hilbert

Nous allons maintenant formaliser un peu tout cela. Dans [406] il est expliqué que si ω est une fonction strictement positive, alors l'espace $L^2_\omega\big([a,b]\big)$ dérivé de la norme

$$\|f\|^2_{L^2_\omega} = \int_a^b |f(x)|^2 \omega(x) dx \tag{36.218}$$

est un espace de Hilbert. Nous allons tenter le coup avec $\omega = \sum_{i=0}^{m} \omega_i \delta_{x_i}$ où δ_a est la distribution de Dirac [12] centrée en a.

Sur l'ensemble des fonctions $\mathbb{R} \to \mathbb{R}$ nous considérons la relation d'équivalence $f \sim g$ si $f(x_i) = g(x_i)$ pour tout $i = 0, \dots, m$. Nous notons L^2_ω cet ensemble.

Proposition 36.98.
La formule

$$\langle f, g \rangle = \sum_{i=0}^{m} f(x_i) g(x_i) \omega_i \tag{36.219}$$

définit un produit scalaire sur L^2_ω. Ce dernier devient un espace de Hilbert.

Démonstration. Pour être un produit scalaire (définition 11.2), la forme considérée doit être symétrique et strictement définie positive. La symétrie de la formule (36.219) ne fait pas de doute. Le fait que ce soit semi-défini positif non plus. Pour le strict,

$$\langle f, f \rangle = \sum_{i=0}^{m} |f(x_i)|^2 \omega_i. \tag{36.220}$$

Étant donné que $\omega_i > 0$ pour tout i, l'annulation de $\langle f, f \rangle$ implique l'annulation de $f(x_i)$ pour tout i. Cela signifie que f est dans la classe de 0 et donc est nul dans L^2_ω.

En ce qui concerne la complétude, la proposition 10.43 répond à notre place, étant donné que L^2_ω est de dimension finie. Une base est donnée par exemple par $e_i(x) = \delta_{x,x_i}$. Ici le δ est celui de Kronecker, et non celui de Dirac. □

Lemme 36.99.
Si la classe de fonctions $\mathcal{F}$ est un sous-espace vectoriel de L^2_ω et si $f \in L^2_\omega$ il existe un unique élément g de $\mathcal{F}$ minimisant la distance à f.

12. Définition 32.31.

Démonstration. Le théorème de projection (au choix 14.239 ou 27.5) nous assure l'existence et l'unicité d'un élément de $\mathcal{F}$ minimisant la distance à $f \in L^2_\omega$. □

36.100.
Ce lemme est gentil, mais ne nous donne pas de méthodes pour trouver ce minimum. Nous allons donc écrire explicitement un système d'équations permettant de le trouver. Si $\{g_\alpha\}$ est une base (finie) de $\mathcal{F}$ alors nous cherchons le minimisant sous la forme $f = \sum_\alpha a_\alpha g_\alpha$.

Nous devons minimiser

$$E(g) = \sum_{i=0}^{m} \big(f(x_i) - g(x_i)\big)^2 \omega_i = \sum_{i=0}^{m} \big(f(x_i) - \sum_\alpha a_\alpha g_\alpha(x_i)\big)^2 \omega_i. \tag{36.221}$$

Vu que cela est maintenant plutôt une fonction des coefficients a_α que de la fonction g nous la notons $F(a_0, \dots, a_n)$. Il s'agit d'étudier le système d'équations

$$\frac{\partial F}{\partial a_\alpha} = 0. \tag{36.222}$$

Un tout petit peu de calcul mène au système

$$\sum_i \sum_\beta a_\beta \omega_i g_\alpha(x_i) g_\beta(x_i) = \sum_i \omega_i g_\alpha(x_i) f(x_i). \tag{36.223}$$

À droite nous reconnaissons $\langle f, g_\alpha \rangle$. et à gauche, $\sum_\beta a_\beta \langle g_\alpha, g_\beta \rangle$. Donc le système s'écrit

$$\sum_\beta a_\beta \langle g_\alpha, g_\beta \rangle = \langle f, g_\alpha \rangle. \tag{36.224}$$

Il y a une équation pour chaque valeur de α.

La matrice $A \in \mathbb{M}(n+1), \mathbb{R}$ donnée par $\langle g_\alpha, g_\beta \rangle$ étant strictement définie positive (c'est un produit scalaire), le système a une unique solution. Et comme cette matrice est de plus symétrique, elle est diagonalisable par le théorème spectral 11.174. Toutes ses valeurs propres sont strictement positives.

Notons pour la curiosité que si l'on considère la matrice $B \in \mathbb{M}(m \times n)$ donnée par

$$B_{ij} = \sqrt{\omega_i} g_j(x_l), \tag{36.225}$$

alors nous avons $A = B^t B$.

36.14.5 Droite de régression

La droite de régression est la cas particulier $n = 1$, c'est à dire un système 2×2. Nous cherchons $P = a_0 + a_1 x$. Et la base choisie est $g_0(x) = 1$, $g_1(x) = x$. Nous avons

$$\langle g_0, g_0 \rangle = \sum_i \omega_i g_0(x_i) g_0(x_i) = \sum_i \omega_i \tag{36.226a}$$

$$\langle g_0, g_1 \rangle = \sum_i \omega_i x_i \tag{36.226b}$$

$$\langle g_1, g_1 \rangle = \sum_i \omega_i x_i^2. \tag{36.226c}$$

Donc pour approximer une fonction f il faut résoudre le système

$$\begin{pmatrix} \sum_i \omega_i & \sum_i \omega_i x_i \\ \sum_i \omega_i x_i & \sum_i \omega_i x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \langle f, g_0 \rangle \\ \langle f, g_1 \rangle \end{pmatrix}. \tag{36.227}$$

Pour calculer les produits $\langle f, g_\alpha \rangle$ il suffit de savoir f sur les points x_i. Et encore heureux, parce que toute la méthode est basée sur le fait que nous ne connaissons pas f ailleurs. C'est pour cela que nous avons défini L^2_ω comme un ensemble quotient.

Exemple 36.101
Faisons la droite de régression pour les données avec tous les poids $\omega_i = 1$.

x_i	$f(x_i)$
-5	18
-3	7
1	0
3	7
4	16
6	50
8	67

Nous avons

$$\langle g, g_0 \rangle = \sum_i f(x_i) = 165 \tag{36.228a}$$

$$\langle f, f_1 \rangle = \sum_i f(x_i)x_i = 810. \tag{36.228b}$$

et donc le système

$$\begin{pmatrix} 7 & 14 \\ 14 & 160 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 165 \\ 810 \end{pmatrix} \tag{36.229}$$

dont résolution donne la droite de régression. △

Proposition 36.102.
Si tous les poids sont identiques, alors la droite de régression passe par le barycentre des points donnés :

$$\begin{cases} x_M = \dfrac{1}{m+1} \displaystyle\sum_{i=0}^{m} x_i & \text{(36.230a)} \\ y_M = \dfrac{1}{m+1} \displaystyle\sum_{i=0}^{m} y_i. & \text{(36.230b)} \end{cases}$$

Cela donne une vérification possible de la réponse trouvée.

Définition 36.103.
*L'**erreur quadratique** est la fonction $F(a_0, \dots, a_n)$ dont il est question plus haut. Et si une solution est connue, son erreur quadratique est la valeur de F pour cette solution.*

36.15 Conditionnement d'une matrice

Soit le système d'équations linéaires $Au = b$ avec la matrice inversible A ainsi que le système perturbé $(A + \Delta A)u' = (b + \Delta b)$. Nous notons $\Delta u = u' - u$ et nous voudrions pouvoir dire des choses de l'erreur relative $\frac{\|\Delta u\|}{\|u\|}$.

Exemple 36.104([133])
Soit la matrice

$$A = \begin{pmatrix} 10 & 7 \\ 7 & 5 \end{pmatrix} \tag{36.231}$$

et $b = \begin{pmatrix} 32 \\ 23 \end{pmatrix}$. La solution de $Au = b$ est $u = \begin{pmatrix} -1 \\ 6 \end{pmatrix}$. Si nous conservons la même matrice mais nous considérons $b = \begin{pmatrix} 32.1 \\ 22.9 \end{pmatrix}$. La solution devient $u' = \begin{pmatrix} 0.2 \\ 4.3 \end{pmatrix}$

En norme $\|.\|_\infty$ nous avons [13]

$$\frac{\|\Delta b\|}{\|b\|} = \frac{0.1}{32} = 0.003125 \tag{36.232}$$

et

$$\frac{\|\Delta u\|}{\|u\|} = \frac{1.7}{6} = 0.28. \tag{36.233}$$

Cela montre environ amplification d'un facteur 100 entre l'erreur sur b et l'erreur sur la solution. △

Définition 36.105.
*Le **conditionnement** de la matrice inversible $A \in \mathrm{GL}(n, \mathbb{C})$ est le nombre positif*

$$\mathrm{Cond}(A) = \|A\|\|A^{-1}\|. \tag{36.234}$$

Cette dénomination sera justifié par le corollaire 36.110 parce qu'il est évident que le conditionnement d'une matrice est lié au conditionnement du problème de résolution d'un système linéaire.

Remarque 36.106.
Le conditionnement dépend de la norme choisie, mais cette dependence est contrôlée par la proposition 12.4 qui nous indique que si le conditionnement d'une matrice est grand dans une norme, il sera grand dans une autre norme.

D'autre part, lorsque nous écrirons $\|A\|$ nous supposerons toujours que $\|.\|$ est une norme d'algèbre [14] et donc que nous avons toujours

$$\|AB\| \leqslant \|A\|\|B\|. \tag{36.235}$$

De plus nous supposerons toujours avoir une norme subordonnée à une norme sur l'espace $\mathbb{C}^n$, de telle sorte à avoir

$$\|Au\| \leqslant \|A\|\|u\| \tag{36.236}$$

pour tout $u \in \mathbb{C}^n$. Voir aussi le lemme 12.15.

Proposition 36.107 ([133])**.**
Si A est une matrice inversible et si $\alpha \in \mathbb{C}$ nous avons :

(1) $\mathrm{Cond}(A) \geqslant 1$

(2) $\mathrm{Cond}(A) = \mathrm{Cond}(A^{-1})$

(3) $\mathrm{Cond}(\alpha A) = \mathrm{Cond}(A)$.

Si $Q \in \mathrm{O}(n)$ alors

(1) Nous avons $\mathrm{Cond}_2(Q) = 1$ où Cond_2 est le conditionnement pour la norme $\|.\|_2$.

(2) Nous avons aussi

$$\mathrm{Cond}_2(A) = \mathrm{Cond}_2(AQ) = \mathrm{Cond}_2(QA). \tag{36.237}$$

Démonstration. Nous savons que $\mathrm{Cond}(\mathbb{1}) = 1$ et donc

$$1 = \|\mathbb{1}\| \leqslant \|A\|\|A^{-1}\| \tag{36.238}$$

parce que la norme utilisée est une norme matricielle.

Les deux autres formules sont évidentes à partit du fait que la définition du conditionnement de A est symétrique entre A et A^{-1}.

13. La proposition 12.4(3) montre que si nous voulions des estimations en norme $\|.\|_2$, il y aurait au maximum un facteur $\sqrt{2}$ par-ci par là.
14. Définition 12.12.

En ce qui concerne les formules relatives à la matrice orthogonale Q nous savons par la proposition 11.68(3) qu'une matrice orthogonale est une bijection de l'ensemble $\{x \in \mathbb{R}^n \text{ tel que } \|x\| = 1\}$. Par conséquent

$$\|AQ\| = \sup_{x \text{ tel que } \|x\|=1} \|AQx\| = \sup_{Q^{-1}x \text{ tel que } \|x\|=1} \|AQQ^{-1}x\| = \|A\|. \tag{36.239}$$

Donc $\|AQ\| = \|A\|$. Les assertions s'ensuivent immédiatement en remarquant que Q^{-1} est également orthogonale. □

Soit une matrice inversible $A \in \mathrm{GL}(n, \mathbb{C})$. La matrice A^*A est hermitienne[15] et le théorème 11.168 nous assure que ses valeurs propres sont réelles. Par la remarque 11.169, ses valeurs propres sont même positives.

Proposition 36.108 ([133]).
Soit une matrice inversible $A \in \mathrm{GL}(n, \mathbb{C})$, et $\mu_1 \leqslant \ldots \leqslant \mu_n$ les valeurs propres de A^*A. Alors nous avons la formule

$$\mathrm{Cond}_2(A) = \sqrt{\frac{\mu_n}{\mu_1}}. \tag{36.240}$$

Démonstration. Par le théorème 12.29, la norme de A est liée au au rayon spectral de A^*A par

$$\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\mu_n}. \tag{36.241}$$

Vu que le spectre de AA^* est le même que celui de A^*A (lemme 11.172) nous avons aussi

$$\|A^{-1}\|_2 = \sqrt{\rho\big((A^{-1})^*A^{-1}\big)} = \sqrt{\rho\big((A^*A)^{-1}\big)} = \frac{1}{\sqrt{\mu_1}} \tag{36.242}$$

parce que la plus grande valeur propre de $(A^*A)^{-1}$ est l'inverse de la plus petite de A^*A.

Ces deux calculs étant,

$$\mathrm{Cond}_2(A) = \|A\|_2\|A^{-1}\|_2 = \sqrt{\frac{\mu_n}{\mu_1}}. \tag{36.243}$$

□

Problèmes et choses à faire

À mon avis ce qui est dans la proposition 19.111 est le conditionnement de la matrice ou sa racine carrée ou un truc du genre. Il faut voir le lien entre les valeurs propres de A et celles de AA^*.

36.15.1 Perturbation du vecteur

Proposition 36.109 (Système linéaire : perturbation du vecteur[133]).
Soit une matrice inversible A et les systèmes d'équations linéaires

$$Au = b \tag{36.244a}$$
$$Au' = b'. \tag{36.244b}$$

En notant $\Delta u = u' - u$ et $\Delta b = b' - b$ nous avons

$$\frac{\|\Delta u\|}{\|u\|} \leqslant \mathrm{Cond}(A)\frac{\|\Delta b\|}{\|b\|}. \tag{36.245}$$

Démonstration. En soustrayant les équations (36.244) nous avons $\Delta b = A\Delta u$, et donc $\Delta u = A^{-1}\Delta b$. D'une part nous avons alors

$$\|\Delta u\| \leqslant \|A^{-1}\|\|\Delta b\|. \tag{36.246}$$

15. Définition 11.61.

Et d'autre part, $\|b\| \leqslant \|A\|\|u\|$, ce qui donne

$$\frac{\|b\|}{\|A\|} \leqslant \|u\|. \tag{36.247}$$

En mettant les deux ensemble,

$$\frac{\|\Delta u\|}{\|u\|} \leqslant \frac{\|A^{-1}\|\|\Delta b\|}{\|b\|}\|A\| = \text{Cond}(A)\frac{\|\Delta b\|}{\|b\|}. \tag{36.248}$$

□

Le corollaire suivant justifie le nom « conditionnement » au conditionnement d'une matrice.

Corollaire 36.110.
Soit $A \in \mathrm{GL}(n, \mathbb{C})$ fixée et le problème de résoudre $Au = b$, c'est à dire la fonction

$$F(u, b) = Au - b. \tag{36.249}$$

(1) Ce problème est stable pour toute valeur de b.

(2) Nous avons une majoration pour le conditionnement relatif[16] *:*

$$K_{rel}(\eta, b_0) \leqslant \text{Cond}(A). \tag{36.250}$$

Démonstration. **Stabilité** Vu que A est inversible, il existe une solution unique à tout système de la forme $Au = b'$. De plus $u(b) = A^{-1}b$, donc

$$\|u(b) - u(b_0)\| = \|A^{-1}(b - b_0)\| \leqslant \|A^{-1}\|\|b - b_0\|, \tag{36.251}$$

de telle sorte que la condition 36.25(2) fonctionne avec $K = \|A^{-1}\|$.

Conditionnement En partant de la définition 36.47, et en utilisant la majoration de la proposition 36.109 sous la forme

$$\|u(b) - u(b_0)\| \leqslant \text{Cond}(A)\|u(b_0)\|\frac{\|\Delta b\|}{\|b_0\|}, \tag{36.252}$$

nous obtenons :

$$K_{rel}(b_0, \eta) = K_{abs}(b_0, \eta)\frac{\|b_0\|}{\|u(b_0)\|} \tag{36.253a}$$

$$= \sup_{\|b-b_0\|\leqslant\eta} \frac{\|u(b) - u(b_0)\|}{\|b - b_0\|}\frac{\|b_0\|}{\|u(b_0)\|} \quad \leqslant \sup_b \text{Cond}(A)\frac{\|u(b_0)\|}{\|b_0\|}\|\Delta b\|\frac{1}{\|b - b_0\|}\frac{\|b_0\|}{\|u(b_0)\|} \tag{36.253b}$$

$$= \text{Cond}(A). \tag{36.253c}$$

□

Remarque 36.111.
La notion de conditionnement relatif dépend aussi de la norme choisie. Dans la formule (36.250) il faut prendre le conditionnement $\text{Cond}(A)$ pour la norme dans laquelle le K_{rel} est écrit. Encore une fois, toutes les normes étant équivalentes, cette majoration est à constante près bonne pour toutes les normes. Si la dimension est très grande, cette constante peut par contre être grande.

16. Si vous doutez de la norme à prendre, lisez la remarque 36.111

36.15.2 Perturbation de la matrice

Proposition 36.112 (Système linéaire : perturbation de la matrice[133]).
Soient les systèmes linéaires

$$Au = b \tag{36.254a}$$
$$A'u' = b \tag{36.254b}$$

avec A et A' inversibles. Nous notons $\Delta A = A' - A$. Alors

(1)

$$\frac{\|\Delta u\|}{\|u'\|} \leqslant \text{Cond}(A)\frac{\|\Delta A\|}{\|A\|} \tag{36.255}$$

(2)

$$\frac{\|\Delta u\|}{\|u\|} \leqslant \text{Cond}(A)\frac{\|\Delta A\|}{\|A\|}\big(1 + \alpha(\|\Delta A\|)\big) \tag{36.256}$$

où $\lim_{x\to 0}\alpha(x) = 0$.

Démonstration. D'abord nous avons

$$0 = Au' - Au \tag{36.257a}$$
$$= (A' - A)u' - Au' - Au \tag{36.257b}$$
$$= \Delta Au' + A\Delta u. \tag{36.257c}$$

Par conséquent, $\Delta u = -A^{-1}(\Delta A)u'$ et

$$\|\Delta u\| \leqslant \|A^{-1}\|\|\Delta A\|\|u'\|. \tag{36.258}$$

Donc

$$\frac{\|\Delta u\|}{\|u'\|} \leqslant \|A^{-1}\|\|A\|\frac{\|\Delta A\|}{\|A\|} = \text{Cond}(A)\frac{\|\Delta A\|}{\|A\|}. \tag{36.259}$$

Cela est (1).

Pour l'autre inégalité, nous avons $A' = A + \Delta A$ et donc

$$\|A'^{-1}\| = \|(A + \Delta A)^{-1}\| \tag{36.260}$$

Nous repartons alors de (36.258) en changeant le rôle de A et A' (et donc aussi de u et u'). Ce changement étant, $\|\Delta u\|$ et $\|\Delta A\|$ ne changent pas. Nous avons :

$$\frac{\|\Delta u\|}{\|u\|} \leqslant \|A'^{-1}\|\|\Delta A\| \tag{36.261a}$$
$$= \|(A + \Delta A)^{-1}\|\|\Delta A\|\frac{\text{Cond}(A)}{\|A\|\|A^{-1}\|} \tag{36.261b}$$
$$= \frac{\|(A + \Delta A)^{-1}\|}{\|A^{-1}\|}\frac{\|\Delta A\|}{\|A\|}\text{Cond}(A). \tag{36.261c}$$

Il reste à voir que

$$\lim_{\|\Delta A\|\to 0}\frac{\|(A + \Delta A)^{-1}\|}{\|A^{-1}\|} = 1, \tag{36.262}$$

ou autrement dit que

$$\lim_{A\to A'}\frac{\|A'^{-1}\|}{\|A^{-1}\|} = 1 \tag{36.263}$$

où la limite est celle dans $\text{GL}(n,\mathbb{C})$. Par définition de la topologie, la norme est continue (quelle qu'elle soit par l'équivalence de norme 12.5). Par le théorème 12.130, l'application $A \mapsto A^{-1}$ est également continue et commute donc avec la limite. Nous avons donc

$$\lim_{A'\to A}\|A'^{-1}\| = \|(\lim_{A'\to A}A')^{-1}\| = \|A^{-1}\|. \tag{36.264}$$

Donc la limite du quotient (36.263) est bien 1. □

36.16 Système linéaires (généralités)

Soit un système d'équations linéaires $Ax = b$ avec $A \in \mathbb{M}(n, \mathbb{R})$. Le problème est évidemment de savoir s'il existe une unique solution x et de la déterminer. Nous supposons l'existence et l'unicité. C'est à dire que les conditions équivalentes [17] sont vérifiées :

(1) A est inversible, c'est à dire qu'il existe une matrice notée A^{-1} telle que $AA^{-1} = A^{-1}A = \mathbb{1}$.

(2) $\det(A) \neq 0$.

Note : si nous avons un système pas carré du type $Bx = v$ avec $B \in \mathbb{M}(n \times m)$ alors nous pouvons nous ramener à un système carré en écrivant

$$B^tBx = B^tv. \tag{36.265}$$

Mais attention : bien que B^tB soit symétrique et semi-définie positive, certaines valeurs propres peuvent être nulles.

36.113.
Deux choses générales en calcul numérique :

(1) On ne calcule pas l'inverse d'une matrice.

(2) On ne calcule même pas son déterminant.

Par conséquent nous ne faisons pas $x = A^{-1}v$.

Il faut garder en tête le fait que dans la pratique, la matrice A possède des millions de lignes et colonnes, si pas pire. Pour une matrice de taille de l'ordre du million, il y a 1000 milliards d'entrées. Si on compte 32 bits par nombre (précision simple, définition 36.8), c'est à dire 4 octets, il faut 4000 giga-octets pour enregistrer la matrice. Même pour la mémoire actuellement disponible, ce n'est pas rien. Surtout que souvent, la précision simple n'est pas utilisée, mais la précision double, ce qui donne 8000 giga pour enregistrer la matrice.

Heureusement, dans la majorité des cas pratiques, les matrices géantes qui apparaissent sont pleines de zéros.

Définition 36.114.
Une matrice est ***creuse*** *si elle possède beaucoup de zéros. Une matrice non creuse est dite* ***dense****.*

Notons que lorsqu'on parle de matrice comprenant beaucoup de « zéros », nous pensons à des éléments très petits, et non de vrai zéros.

Les matrices creuses ne sont pas mémorisées entièrement, mais plutôt comme un dictionnaire (i, j, v) qui donne la valeur v de A_{ij}.

Définition 36.115.
Une matrice est de « grande dimension » si elle ne peut pas être mise en mémoire sur un ordinateur donné. Sur certains ordinateurs, ça commence à 5000 *inconnues. Mais sur des plus forts, on peut aller jusqu'au million ou le milliard.*

Si la matrice est de petite dimension, il est possible d'utiliser des méthodes dites « directes ». Sinon, il faudra utiliser des méthodes itératives.

36.16.1 Les méthodes directes

Une méthode directe consiste à successivement transformer un système $A^{(0)}x = b^{(0)}$ en de nouveaux systèmes $A^{(i)}x = b^{(i)}$ dont la solution est identique jusqu'à obtenir un système $A^{(n-1)}x = b^{(n-1)}$ qui est à résolution immédiate.

L'avantage d'une méthode directe est qu'elle fournit une réponse exacte, pour autant que les calculs intermédiaires soient bien faits (ce qui n'est pas le cas sur un ordinateur).

Une méthode directe fonctionne en général avec un nombre de pas fixés par la taille du système. Par exemple pour un système $n \times n$, la méthode de Gauss demande exactement n pas, et il n'y a

17. L'équivalence est la proposition 11.42(2).

pas moyen de faire mieux. Or chaque pas demande de recalculer tous les éléments de la matrice. Encore une fois, si la matrice a une taille de l'ordre du milliard, cela fait 10^{18} éléments à recalculer un milliard de fois (sans compter les éléments du vecteur b). Infaisable.

Souvent une méthode directe passe par une factorisation $A = BC$ avec $B, C \in \mathbb{M}(n \times n)$.

Quelques types de matrices dont la résolution est immédiate :

— Matrice diagonale.
— Matrice orthogonale parce que si A est orthogonale alors $Ax = v$ se résout par $x = A^t v$ qui n'est pas particulièrement lourd à faire numériquement.
— Matrice triangulaire.

Remarque 36.116.
Pour une matrice diagonale, le déterminant et l'inverse sont faciles. Mais également pour la triangulaire. Pour une matrice triangulaire, le déterminant est le produit des éléments diagonaux, et il se fait qu'il y a une algorithme facile pour calculer l'inverse.

Donc en fait les matrices à résolution immédiates sont des matrices pour lesquelles l'inverse et le déterminant sont facile à calculer.

36.16.2 Méthodes itératives

Si la matrice est trop grande, il n'est pas possible de faire des manipulations de matrices à chaque itération.

En général, les méthodes itératives ne convergent pas toujours. Mais lorsqu'une méthode converge, c'est une propriété de la matrice, et donc la convergence aura lieu pour tout vecteur de départ x_0. Cela est très différent du cas des équations non linéaires type Newton pour lesquelles la convergence peut fortement dépendre du point de départ.

36.17 Système linéaires (méthodes directes)

Les matrices que nous sommes autorisés à inverser sont les matrices
— orthogonales : l'inverse est la transposée
— diagonales : l'inverse est diagonale avec les inverses sur la diagonale
— triangulaires : nous en parlons maintenant.

36.17.1 Inversion de matrice triangulaire

Si T est une matrice triangulaire (mettons supérieure pour fixer les idées), il est possible d'en calculer l'inverse sans trop d'efforts. Notons B la matrice inverse que nous allons construire ligne par ligne. Vu que $BT = \mathbb{1}$ nous avons

$$\delta_{1j} = \sum_{k=1}^{n} B_{1k} T_{kj} = \sum_{k=1}^{j} B_{1k} T_{kl} \tag{36.266}$$

parce que $T_{kj} = 0$ pour $k > j$. Donc nous pouvons calculer les éléments B_{1j} un par un parce que chacun ne dépend que des précédents. Le même procédé fonctionne pour les autres lignes :

$$\delta_{ij} = \sum_{k=1}^{j} B_{ij} T_{kj}. \tag{36.267}$$

Et tu notes que le calcul peut être parallélisé : le calcul de la ligne numéro j ne dépend pas du résultat des autres lignes.

36.17.2 Transformation gaussienne

Définition 36.117 (Transformation gaussienne[407])**.**
*Soit $x \in \mathbb{R}^n$ avec $x_k \neq 0$. La k^e**transformation gaussienne** pour x est la matrice*

$$M_k(x) = \mathbb{1} - T_k(x) \tag{36.268}$$

où $T_k(x)$ est la matrice unité à qui on a ajouté le vecteur

$$\tau_k(x) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ x_{k+1}/x_k \\ \vdots \\ x_n/x_k \end{pmatrix} \tag{36.269}$$

à la k^ecolonne.

Autrement dit, la matrice $M_k(x)$ est la matrice

$$M_k(x) = \begin{pmatrix} 1 & & & & & \\ 0 & \ddots & & & & \\ \vdots & 0 & 1 & & & \\ \vdots & \vdots & -x_{k+1}/x_k & 1 & & \\ \vdots & \vdots & \vdots & 0 & \ddots & \\ 0 & 0 & -x_n/x_k & 0 & & 1 \end{pmatrix} \tag{36.270}$$

En coordonnées nous avons

$$\big(M_k(x)\big)_{ij} = \delta_{ij} - \tau_k(x)_i\delta_{kj}. \tag{36.271}$$

36.118.
Les matrices de transformation gaussienne sont des matrices triangulaires de diagonale unitaire (c'est à dire avec des 1 sur la diagonale).

Lemme 36.119.
Si $x \in \mathbb{R}^n$ alors nous avons

$$M_k(x)x = \begin{pmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{36.272}$$

Démonstration. Nous avons

$$\big(M_k(x)x\big)_i = \sum_l M_k(x)_{il}x_l = \sum_l \big(\delta_{il} - \tau_k(x)_i\delta_{kl}\big)x_l = x_i - \tau_k(x)_i x_k. \tag{36.273}$$

Si $i \leqslant k$ nous avons $\tau_k(x)_i = 0$ et donc $\big(M_k(x)x\big)_i = x_i$. Si par contre $i \geqslant k+1$ alors $\tau_k(x)_i = \frac{x_i}{x_k}$ et alors $\big(M_k(x)x\big)_i = 0$. □

Lemme 36.120.
Si $y \in \mathbb{R}^n$ vérifie $y_i = 0$ pour $i > k$ alors $M_{k+1}(x)y = y$.

Démonstration. C'est une simple vérification :

$$\big(M_{k+1}(x)y\big)_i = \sum_l \big(\delta_{il} - \tau_{k+1}(x)_i\delta_{k+1,l}\big)y_l = y_i - \tau_{k+1}(x)_i y_{k+1}. \tag{36.274}$$

Mais comme $y_{k+1} = 0$ il nous reste automatiquement y_i. □

Le sens de ce lemme est si un vecteur est déjà « gaussiannisé » au niveau k, alors en lui appliquant une transformation gaussienne de niveau plus élevé que k, il ne change pas. Ce fait est important parce qu'il assure que lorsque l'on avance dans le processus de Gauss, chaque étape ne détruit pas les précédentes.

Le lemme suivant nous indique que l'inverse d'une matrice de transformation gaussienne est facile à calculer [18].

Lemme 36.121.
L'inverse de la transformation gaussienne

$$M_k(x)_{ij} = \delta_{ij} - \tau_k(x)_i\delta_{kj}. \tag{36.275}$$

est la matrice donnée par

$$M_k(x)^{-1}_{ij} = \delta_{ij} + \tau_k(x)_i\delta_{kj}. \tag{36.276}$$

Autrement dit, il suffit de changer le signe de la partie non diagonale.

Démonstration. Il s'agit d'une simple vérification, utilisant le produit matriciel explicite, et en remarquant que $\tau_k(x)_k = 0$ pour tout k. □

36.17.3 Méthode de Gauss pour résoudre des systèmes d'équations linéaires

Pour résoudre un système d'équations linéaires, on procède comme suit :

(1) Écrire le système sous forme matricielle.

$$\text{p.ex. } \begin{cases} 2x + 3y &= 5 \\ x + 2y &= 4 \end{cases} \Leftrightarrow \left(\begin{array}{cc|c} 2 & 3 & 5 \\ 1 & 2 & 4 \end{array}\right)$$

(2) Se ramener à une matrice avec un maximum de 0 dans la partie de gauche en utilisant les transformations admissibles :

(a) Remplacer une ligne par elle-même + un multiple d'une autre ;

$$\text{p.ex. } \left(\begin{array}{cc|c} 2 & 3 & 5 \\ 1 & 2 & 4 \end{array}\right) \overset{L_1 - 2.L_2 \mapsto L_1'}{\Longrightarrow} \left(\begin{array}{cc|c} 0 & -1 & -3 \\ 1 & 2 & 4 \end{array}\right)$$

(b) Remplacer une ligne par un multiple d'elle-même ;

$$\text{p.ex. } \left(\begin{array}{cc|c} 0 & -1 & -3 \\ 1 & 2 & 4 \end{array}\right) \overset{-L_1 \mapsto L_1'}{\Longrightarrow} \left(\begin{array}{cc|c} 0 & 1 & 3 \\ 1 & 2 & 4 \end{array}\right)$$

(c) Permuter des lignes.

$$\text{p.ex. } \left(\begin{array}{cc|c} 0 & 1 & 3 \\ 1 & 0 & -2 \end{array}\right) \overset{L_1 \mapsto L_2' \text{ et } L_2 \mapsto L_1'}{\Longrightarrow} \left(\begin{array}{cc|c} 1 & 0 & -2 \\ 0 & 1 & 3 \end{array}\right)$$

(3) Retransformer la matrice obtenue en système d'équations.

$$\text{p.ex. } \left(\begin{array}{cc|c} 1 & 0 & -2 \\ 0 & 1 & 3 \end{array}\right) \Leftrightarrow \begin{cases} x &= -2 \\ y &= 3 \end{cases}$$

Remarques :
— Si on obtient une ligne de zéros, on peut l'enlever :

$$\text{p.ex. } \left(\begin{array}{ccc|c} 3 & 4 & -2 & 2 \\ 4 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right) \Leftrightarrow \left(\begin{array}{ccc|c} 3 & 4 & -2 & 2 \\ 4 & -1 & 3 & 0 \end{array}\right)$$

18. Elle rentre d'ailleurs dans la catégorie des matrices triangulaires dont nous avons déjà discuté l'inverse.

— Si on obtient une ligne de zéros suivie d'un nombre non-nul, le système d'équations n'a pas de solution :

$$\text{p.ex. } \left(\begin{array}{ccc|c} 3 & 4 & -2 & 2 \\ 4 & -1 & 3 & 0 \\ 0 & 0 & 0 & 7 \end{array}\right) \Leftrightarrow \begin{cases} \cdots \\ \cdots \\ 0x + 0y + 0z = 7 \end{cases} \Rightarrow \textbf{Impossible}$$

— Si on moins d'équations que d'inconnues, alors il y a une infinité de solutions qui dépendent d'un ou plusieurs paramètres :

$$\text{p.ex. } \left(\begin{array}{ccc|c} 1 & 0 & -2 & 2 \\ 0 & 1 & 3 & 0 \end{array}\right) \Leftrightarrow \begin{cases} x - 2z = 2 \\ y + 3z = 0 \end{cases} \Leftrightarrow \begin{cases} x = 2 + 2\lambda \\ y = -3\lambda \\ z = \lambda \end{cases}$$

36.17.4 Méthode de Gauss sans pivot (décomposition LU)

La méthode de Gauss est encore utilisée aujourd'hui dans les vrais problèmes.

La méthode de Gauss est souvent aussi appelée méthode « LU » qui va décomposer $A = LU$ où L est triangulaire inférieure et U est triangulaire supérieure. La décomposition est même plus précise que cela : on demande que L ait seulement des 1 sur la diagonale.

Si A est une matrice nous notons

$$\Delta_k(A) = (A_{ij})_{1\leqslant i,j\leqslant k} \tag{36.277}$$

la matrice tronquée dont nous ne gardons que le carré $k \times k$ en haut à gauche.

Lemme 36.122.
Soit S une matrice triangulaire inférieure. Soient également A et B telles que $B = SA$. Alors

$$\Delta_k(B) = \Delta_k(S)\Delta_k(A). \tag{36.278}$$

Démonstration. En effet nous avons

$$\Delta_k(B)_{ij} = \sum_{l=1}^{n} S_{il}A_{lj}. \tag{36.279}$$

Dans la somme sur l il ne reste que les termes $l \leqslant i$. Mais dans le calcul des éléments de matrice $\Delta_k(B)_{ij}$, nous avons évidemment $i, j \leqslant k$. Donc $l \leqslant i \leqslant k$. Les seuls éléments de matrice de A qui sont utilisés dans la somme (36.279) sont les éléments A_{lj} avec $l, j \leqslant k$.

Nous pouvons donc limiter la somme à $l = k$ au lieu de $l = n$ et écrire $\Delta_k(A)_{lj}$ au lieu de A_{lj}.

Même chose en ce qui concerne S. À partir du moment où l est limité à k, les éléments S_{il} et $\Delta_k(S)_{il}$ sont les mêmes. □

Théorème 36.123 (Décomposition LU[408, 1])**.**
Soit une matrice A inversible telles que $\det(\Delta_k(A)) \neq 0$ pour tout k. Alors il existe un unique couple de matrices (L, U) telles que

— *U soit triangulaire supérieure*
— *L soit triangulaire inférieure, de diagonale unité*
— *$A = LU$.*

De plus pour tout $k \leqslant n$ nous avons

$$\Delta_k(A) = \Delta_k(L)\Delta_k(U). \tag{36.280}$$

Démonstration. Nous allons prouver par récurrence le fait suivant : pour tout $1 \leqslant k \leqslant n-1$ il existe des matrices E_i $(i = 1, \ldots, k)$ telles que en posant

$$A_k = E_k \ldots E_1 A, \tag{36.281}$$

— E_j est une transformation gaussienne pour la j^{e}colonne,
— pour tout $j \leqslant k$, $A_{ij} = 0$ dès que $i > j$. Autrement dit la matrice A_k est triangulaire supérieure jusqu'à y compris la $(k+1)^{\text{e}}$colonne (laquelle est quelconque). Exemple pour fixer les idées : pour une matrice $A \in \mathbb{M}(4 \times 4)$, la matrice A_2 doit avoir la forme

$$A_2 = E_2E_1A = \begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & \circledast & * \\ 0 & 0 & * & * \end{pmatrix} \tag{36.282}$$

où les éléments notés $*$ sont a priori non nuls,
— l'élément de matrice $(A_k)_{k+1,k+1}$ est non nul (celui entouré dans l'exemple).

La chose un peu triste dans cette démonstration est que l'initialisation va être très ressemblante au pas de récurrence.

Initialisation : $k = 1$ Vu que $\Delta_1(A)$ est inversible, l'élément A_{11} est non nul. Il existe donc une transformation gaussienne E_1 telle que la première colonne de la matrice $A_1 = E_1A$ soit nul sauf la première composante. En particulier $(A_1)_{21} = 0$.

Par le lemme 36.122, nous avons $\Delta_2(A_1) = \Delta_2(E_1)\Delta_2(A)$, donc [19]

$$\det\big(\Delta_2(A_1)\big) = \det\big(\Delta_2(E_1)\big)\det\big(\Delta_2(A)\big). \tag{36.283}$$

Étant donnée la forme (36.270), toutes les matrices du type $\Delta_k(E_i)$ ont un déterminant unité, et par hypothèse $\Delta_2(A)$ est inversible, donc de déterminant non nul. Par conséquent $\det\big(\Delta_2(A_1)\big) \neq 0$. Mais comme ce déterminant est le produit des éléments diagonaux (c'est une matrice triangulaire), ces derniers ne sont pas nuls. Finalement, $(A_1)_2 2 \neq 0$.

Le pas de récurrence Nous supposons avoir $A_k = E_k \dots E_1A$ avec $(A_k)_{k+1,k+1} \neq 0$. Alors il existe une transformation gaussienne E_{k+1} de la $(k+1)^{\text{e}}$colonne telle que $A_{k+1} = E_{k+1}A_k$ soit une matrice dont la $(k+1)^{\text{e}}$colonne n'ait que des zéros en dessous de la $(k+1)^{\text{e}}$position. Vu le lemme 36.120, cette transformation n'affecte pas les colonnes précédentes.

La matrice A_{k+1} est donc triangulaire supérieure jusqu'à la $(k+2)^{\text{e}}$colonne.

Vu que le produit $E_k \dots E_1$ est une matrice triangulaire inférieure, le lemme 36.122 fonctionne encore et nous avons

$$\Delta_{k+1}(A_k) = \Delta_{k+1}(E_k \dots E_1)\Delta_{k+1}(A). \tag{36.284}$$

En ce qui concerne les déterminants, par hypothèse, nous avons $\det\big(\Delta_{k+1}(A)\big) \neq 0$ ainsi que $\det\big(\Delta_{k+1}(E_k \dots E_1)\big) = 1$. Donc

$$\det\big(\Delta_{k+1}(A_k)\big) \neq 0. \tag{36.285}$$

Cette matrice étant triangulaire, ses éléments diagonaux sont non nuls et nous avons $(A_k)_{k+1,k+1} \neq 0$.

En poussant la récurrence jusqu'au bout, la matrice

$$A_{n-1} = E_{n-1} \dots E_nA \tag{36.286}$$

est triangulaire supérieure.

Nous posons alors $L = (E_{n-1} \dots E_n)^{-1}$ et $U = A_{n-1}$. Cela prouve l'existence parce que

$$A = (E_{n-1} \dots E_1)^{-1}A_{n_1}. \tag{36.287}$$

Encore une fois, le lemme 36.122 nous donne

$$\Delta_k(A) = \Delta_k\Big((E_{n_1} \dots E_1)^{-1}\Big)\Delta_k(A_{n-1}), \tag{36.288}$$

19. Le déterminant est multiplicatif, proposition 11.42(1).

ou encore $\Delta_k(A) = \Delta_k(L)\Delta_k(U)$.

En ce qui concerne l'unicité, si $A = L_1U_1 = L_2U_2$ alors $L_2^{-1}L_1 = U_2U_1^{-1}$. Vu qu'à gauche nous avons une matrice triangulaire inférieure et que à droite nous avons une triangulaire inférieure, nous savons que les deux membres représentent une matrice diagonale. Mais à gauche, la diagonale est unitaire. Donc les deux membres représentent la matrice unité. □

36.124.
En pratique, pour résoudre $Ax = b$, il faut seulement appliquer les transformations gaussiennes à la matrice élargie $(A|b)$ pour finir sur un système du type

$$Ux = b' \tag{36.289}$$

qui est immédiatement soluble. Autrement dit, en effectuant les annulations de colonnes, la matrice U est « gratuite ».

Il n'est pas indispensable de calculer la matrice L qui, elle, demande à chaque étape de se souvenir de la matrice E_i utilisée. S'il faut résoudre plusieurs systèmes $Ax_i = b_i$, nous pouvons encore travailler avec la matrice encore plus élargie $(A|b_1 \dots b_m)$.

Si par contre nous ne connaissons pas à l'avance l'ensemble des vecteurs b avec lesquels il faudra résoudre le système, il est bon de calculer la décomposition $A = LU$ in extenso, c'est à dire de garder une trace des matrices L et U séparément. Dans ce cas, résoudre $Ax = b$ revient à résoudre $Ly = b$, et ensuite $Ux = y$. Ce sont deux systèmes de résolution directe parce que les matrices sont triangulaires.

36.125.
Le fait que

$$\Delta_k(A) = \Delta_k(L)\Delta_k(U) \tag{36.290}$$

nous dit que si après avoir calculer L et U nous remarquons que le système est un peu plus petit ou un peu plus grand que prévu, tout le travail n'est pas perdu. En particulier si le système est plus petit que prévu, l'adaptation de L et U est immédiate.

Notons que U et L sont inversibles, et que $\det(L) = 1$. Donc $\det(U) = \det(A)$.

Exemple 36.126
Pour travailler la méthode de Gauss pour le système $Ax = b$, nous introduisons la matrice un peu augmentée $(A|b)$. Nous faisons un exemple. Soit à résoudre

$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 3 & 10 \\ -2 & 1 & 73 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 11 \\ 28 \\ 3 \end{pmatrix}. \tag{36.291}$$

Nous introduisons la matrice augmentée

$$(A|b)^{(0)} = \begin{pmatrix} 2 & 1 & 3 & 11 \\ 4 & 3 & 10 & 28 \\ -2 & 1 & 7 & 3 \end{pmatrix}. \tag{36.292}$$

Le premier pas consiste à annuler tous les éléments sous la diagonale de la première colonne. Autrement dit, nous prenons le 2 comme pivot. Nous introduisons les multiplicateurs $l_{ij} = \frac{A_{ij}}{A_{i1}}$. La nouvelle matrice est :

$$(A|b)^{(1)} = \begin{pmatrix} 2 & 1 & 3 & 11 \\ 0 & 1 & 4 & 6 \\ 0 & 2 & 10 & 14 \end{pmatrix} \tag{36.293}$$

où nous avons utilisé les multiplicateurs $l_{21} = 2$, $l_{31} = -1$.

Et la matrice suivante est :

$$(A|b)^{(2)} = \begin{pmatrix} 2 & 1 & 3 & 11 \\ 0 & 1 & 4 & 6 \\ 0 & 0 & 2 & 2 \end{pmatrix} \tag{36.294}$$

où nous avons utilisé le multiplicateur $l_{32} = 2$.

Cela est un système de résolution immédiate :

$$\begin{cases} 2x + y + 3z = 11 & (36.295a) \\ y + 4z = 6 & (36.295b) \\ 2z = 2. & (36.295c) \end{cases}$$

La troisième donne $z = 1$. Ensuite $y + 4 = 6$, donc $y = 2$. Et la première donne : $2x + 2 + 3 = 11$, c'est à dire $2x = 6$, enfin : $x = 3$.

Solution : $(x, y, z) = (3, 2, 1)$.

Nous notons surtout que dans $(A|b)^{(2)}$ nous avons une matrice triangulaire supérieure. Où est la matrice triangulaire inférieure ? En réalité la matrice L est la matrice des multiplicateurs :

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 2 & 1 \end{pmatrix}. \tag{36.296}$$

△

Le problème de cette méthode est que faisant ainsi nous risquons d'avoir un zéro sur un des pivots. Par exemple tomber sur

$$(A|b) = \begin{pmatrix} 2 & 1 & 3 & 11 \\ 0 & 0 & 4 & 6 \\ 0 & 2 & 10 & 14 \end{pmatrix}. \tag{36.297}$$

Le zéro sur la deuxième ligne nous ennuie si nous voulons tout faire dans l'ordre. Mais notons qu'en échangeant les deux dernières lignes, tout va bien : le système donné par

$$(A|b) = \begin{pmatrix} 2 & 1 & 3 & 11 \\ 0 & 2 & 10 & 14 \\ 0 & 0 & 4 & 6 \end{pmatrix} \tag{36.298}$$

fonctionne très bien. Et même tellement bien qu'il est de résolution immédiate, dans ce cas.

Un autre problème est que si un des pivots est 10^{-14}, le multiplicateur sera de l'ordre 10^{14}, qui est mal représenté en mémoire. Il est donc bon de prendre les pivots le plus grand possible. Si le pivot est le plus grand nombre en valeur absolue d'une colonne, alors les nombres x_{k+i}/x_k qui entrent dans la matrice de transformation gaussienne sont des nombres dans $[-1, 1]$ qui sont bien représentés en mémoire.

Tout cela nous incite à développer une méthode de Gauss qui permet de tenir une trace des permutations.

36.17.5 Matrice de permutation élémentaires

Définition 36.127.
*Une **matrice de permutation élémentaire** est une matrice obtenue en permutant deux lignes de la matrice identité. Nous notons P_{ij} la matrice obtenue en inversant les lignes i et j de la matrice identité.*

Exemple 36.128

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = P_{12}. \tag{36.299}$$

△

Lemme 36.129.
La matrice $P_{ij}A$ est la matrice A avec ses lignes i et j inversées.

Démonstration. Il suffit d'écrire

$$(P_{ij}A)_{kl} = \sum_m (P_{ij})_{km} A_{ml} \tag{36.300}$$

et de faire trois cas selon que $k = i$, $k = j$ ou k différent de i et j. Si $k = i$ alors $(P_{ij})_{im} = \delta_{mj}$ et si k est différent de i et j alors $(P_{ij})_{mk} = \delta_{km}$ (troisième cas similaire au premier). □

Et la matrice AP_{12} est la A avec ses deux premières *colonnes* échangées.
Avec ces notations, notre matrice $(A|b)^{0'}$ est

$$P_{12}(A|b)^{(0)}. \tag{36.301}$$

Puis la matrice $(A|b)^{(1')}$ est

$$P_{23}(A|b)^{(1)}. \tag{36.302}$$

Et la matrice P qui arrive dans $PA = LU$ est la matrice $P = P_{23}P_{21}$, qui est une matrice de permutation non élémentaire. Elle vaut :

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}. \tag{36.303}$$

Lemme 36.130 ([409]).
Si $i, j > k$ alors les matrices de permutation élémentaires ont la relation de « commutation » suivante avec les transformations gaussiennes :

$$M_k(x)P_{ij} = P_{ij}M_k\big(P_{ij}(x)\big). \tag{36.304}$$

Démonstration. Il suffit de calculer les éléments de matrice :

$$\big(P_{ij}M_k(x)\big)_{st} = (P_{ij})_{st} - \sum_m (P_{ij})_{sm}\tau_k(x)_m\delta_{kt}, \tag{36.305}$$

mais

$$\sum_m (P_{ij})_{sm}\tau_k(x)_m = \big(P_{ij}\tau_k(x)\big)_s = \tau_k\big(P_{ij}(x)\big)_s \tag{36.306}$$

parce que $i, j > k$ implique que dans $P_{ij}\tau_k(x)$ nous inversons deux élément non nuls de $\tau_k(x)$, tout en laissant le k^{e}élément. Le dénominateur ne change pas et il s'agit réellement d'une inversion de ligne. Donc

$$\big(P_{ij}M_k(x)\big)_{st} = (P_{ij})_{st} - \tau_k\big(P_{ij}x\big)_s\delta_{kt}. \tag{36.307}$$

De l'autre côté,

$$\big(M_k(y)P_{ij}\big)_{st} = (P_{ij})_{st} - \tau_k(y)_s(P_{ij})_{kt}. \tag{36.308}$$

Mais comme $i, j > k$ la k^{e}ligne de P_{ij} est la même que celle de la matrice unité, donc $(P_{ij})_{kt} = \delta_{kt}$.

$$\big(M_k(y)P_{ij}\big)_{st} = (P_{ij})_{st} - \tau_k(y)_s\delta_{kt}. \tag{36.309}$$

Cela correspond bien à (36.307). □

36.18 Méthode de Gauss avec pivot partiel (décomposition PLU)

36.18.1 L'idée

À chaque pas, nous faisons une permutation de ligne. Nous permutons à chaque pas la première ligne avec celle qui a le pivot le plus grand (en valeur absolue). Donc :

$$(A|b)^{(0)} = \begin{pmatrix} 2 & 1 & 3 & 11 \\ 4 & 3 & 10 & 28 \\ -2 & 1 & 7 & 3 \end{pmatrix} \tag{36.310}$$

Nous commençons par déplacer des lignes :

$$(A|b)^{(0')} = \begin{pmatrix} 4 & 3 & 10 & 28 \\ 2 & 1 & 3 & 11 \\ -2 & 1 & 7 & 3 \end{pmatrix}. \tag{36.311}$$

Les multiplicateurs sont $l_{21} = 1/2$ et $l_{31} = -1/2$. Le fait est que les multiplicateurs ont toujours le plus grand dénominateur possible et nous avons alors toujours $0 \leqslant |l_{ij}| \leqslant 1$, qui sont des nombres relativement petits, et bien représentés en mémoire.

Nous avons la nouvelle matrice

$$(A|b)^{(1)} = \begin{pmatrix} 4 & 3 & 10 & 28 \\ 0 & -1/2 & -2 & -3 \\ 0 & 5/2 & 12 & 17 \end{pmatrix}. \tag{36.312}$$

Le pivot serait $-1/2$. Nous cherchons un pivot plus grand en dessous de ce $-1/2$ (et pas au dessus, sinon on casserait les zéros déjà trouvés). Nous trouvons le $5/2$ qui est plus grand. Nous permutons donc les deux dernières lignes :

$$(A|b)^{(1')} = \begin{pmatrix} 4 & 3 & 10 & 28 \\ 0 & 5/2 & 12 & 17 \\ 0 & -1/2 & -2 & -3 \end{pmatrix} \tag{36.313}$$

où le pivot est maintenant $l_{32} = -1/5$. La matrice suivante :

$$(A|b)^{(1)} = \begin{pmatrix} 4 & 3 & 10 & 28 \\ 0 & 5/2 & 12 & 17 \\ 0 & 0 & 2/5 & 2/5 \end{pmatrix} \tag{36.314}$$

Dans ce cas, la matrice L n'est pas aussi simple à construire parce que nous avons permuté des choses. Dans ce cas, la matrice L est encore de la forme

$$L = \begin{pmatrix} 1 & 0 & 0 \\ . & 1 & 0 \\ . & . & 1 \end{pmatrix}. \tag{36.315}$$

Mais vu que nous avons permuté les lignes 2 et 3 au deuxième pas, nous devons permuter l_{21} et l_{31} avant de remplir la matrice L avec les multiplicateurs :

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 1/2 & -1/5 & 1 \end{pmatrix}. \tag{36.316}$$

Notons que ces L et U ne sont pas les mêmes que le LU obtenu sans pivot. Où est l'unicité ? Elle est que en fait maintenant nous n'avons pas $A = LU$, mais

$$PA = LU \tag{36.317}$$

où P est une matrice de permutation.

36.18.2 Le théorème

Proposition 36.131 (Méthode de Gauss avec pivot partiel[409])**.**
Soit une matrice inversible $A \in \mathbb{M}(n,\mathbb{C})$. Il existe

— *une matrice de permutation P*
— *une matrice triangulaire inférieure de diagonale unitaire L,*
— *une matrice triangulaire supérieure inversible U*

telles que

$$PA = LU. \tag{36.318}$$

Notons que cette proposition ne demande que l'hypothèse d'inversibilité pour A. Il n'y a pas d'hypothèses sur tous les mineurs comme c'était le cas avec Gauss sans pivot.

Démonstration. Nous prouvons par récurrence qu'il existe des matrices Q_k, E_1, ..., E_k et A_k telles que

$$Q_kA = E_1 \dots E_kA_k \tag{36.319}$$

avec

(1) Q_k est une matrice de permutation
(2) E_i est une transformation gaussienne sur la i^{e}colonne
(3) A_k est triangulaire supérieure jusqu'à la k^{e}colonne.

Sachant que $\det(Q_k) = \pm 1$, et que $\det(E_i) = 1$, le passage au déterminant dans (36.319) nous donne $\det(A_k) \neq 0$ et si nous notons $\Omega_k(A)$ la matrice tronquée de A, ne gardant que les entrées plus grandes que k, nous avons

$$\det(A_k) = \prod_{i=1}^{k}(A_k)_i i \det\big(\Omega_{k+1}(A_k)\big). \tag{36.320}$$

Donc : $(A_k)_{ii} \neq 0$ pour $i \leqslant k$ et $\det\big(\Omega_{k+1}(A_k)\big) \neq 0$.

Pour fixer les idées, voici une image de $k = 2$:

$$\begin{array}{c} \overbrace{}^{\Delta_k(A_2)} \\ \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{pmatrix} \\ \underbrace{}_{\Omega_{k+1}(A_2)} \end{array} \tag{36.321}$$

Étant donné que $\det\big(\Omega_{k+1}(A_k)\big) \neq 0$, parmi les nombres $(A_k)_{i,k+1}$ $(i \geqslant k+1)$, au moins un est non nul et nous posons r_{k+1} tel que $|(A_k)_{r_{k+1},k+1}|$ soit maximum parmi ces éléments.

Le nombre r_{k+1} est enregistré parce qu'il servira à écrire la matrice P plus tard. Les matrices E_i ne sont pas enregistrées, parce que nous verrons qu'elles vont encore changer. Seule la dernière sera enregistrée.

La composante $(k+1, k+1)$ de la matrice

$$P_{r_{k+1},k+1}A_k \tag{36.322}$$

est non nulle et peut donc servir de pivot. Soit M_{k+1} la transformation gaussienne pour la $(k+1)^{\text{e}}$colonne de la matrice $P_{r_{k+1},k+1}A_k$. La matrice

$$A_{k+1} = M_{k+1}P_{r_{k+1},k+1}A_k \tag{36.323}$$

est alors une matrice triangulaire supérieure jusqu'à la $(k+1)^{\text{e}}$colonne. En posant $E_{k+1} = M_{k+1}^{-1}$ nous avons

$$P_{r_{k+1},k+1}E_{k+1}A_{k+1} = A_k, \tag{36.324}$$

et nous nous sentons en droit de récrire l'équation de départ (36.319) :

$$Q_kA = E_1 \dots E_kA_k = E_1 \dots E_kP_{r_{k+1},k+1}E_{k+1}A_{k+1}. \tag{36.325}$$

Le lemme 36.130 nous permet de ramener la matrice $P_{r_{k+1},k+1}$ en première position, quitte à modifier un peu (pas beaucoup) chacune des matrices E_i ($i = 1, \dots, k$). C'est pour cela que nous n'enregistrons pas les matrices E_i. Nous avons donc

$$P_{r_{k+1},k+1}Q_kA = E'_1 \dots E'_kE_{k+1}A_{k+1} \tag{36.326}$$

où

— Le produit $P_{r_{k+1},k+1}Q_k$ est encore une matrice de permutation, et mieux : elle vaut

$$\prod_{i=1}^{k+1} P_{r_i,i}. \tag{36.327}$$

Cela montre qu'il est suffisant d'enregistrer les nombres r_i pour reconstituer cette partie.

— La matrice E'_i est une transformation gaussienne pour la i^{e}colonne.

— La matrice A_{k+1} est triangulaire supérieure jusqu'à la $k+1^{\text{e}}$colonne.

La récurrence est maintenant finie et nous pouvons écrire avec $k = n$:

$$Q_nA = E_1 \dots E_nA_n \tag{36.328}$$

où le produit $E_1 \dots E_n$ est triangulaire inférieure et A_n est triangulaire supérieur.

Maintenant nous enregistrons la matrice $U = A_n$, le produit $L = \prod_{i=1}^{n} E_n$ et les nombres r_i qui permettent de retrouver P. □

Note : dans l'équation (36.328) nous avons bien entendu massivement renommé les E'_i en E_i. En réalité la matrice E_1 vient avec n primes sur la tête.

Dans les exemples 36.134, 36.135 et 36.136, nous allons résoudre le système

$$\begin{pmatrix} 10^{-9} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{36.329}$$

d'abord de façon exacte, et ensuite en supposant une machin ne tenant que 8 chiffres significatifs en utilisant la méthode de Gauss avec ou sans pivot.

Commençons par voir comment se passe en pratique la décomposition $PA = LU$ de Gauss avec pivot partiel.

Exemple 36.132

Décomposons la matrice

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 0 \\ 3 & 8 & 0 \end{pmatrix}. \tag{36.330}$$

Sur la première colonne, le plus grand nombre est 3. Nous commençons par permuter la première et la troisième ligne en utilisant la matrice de permutation $P_1 = P_{3,1}$ et nous enregistrons $r_1 = 3$. Nous avons alors la matrice

$$A'_0 = \begin{pmatrix} 3 & 8 & 0 \\ 2 & 5 & 0 \\ 1 & 3 & 3 \end{pmatrix}. \tag{36.331}$$

Pour trouver la matrice A_1 nous suivons l'équation (36.323). Bien que le résultat net soit des combinaisons de lignes : $L_2 \to L_2 - 2L_1/3$ et $L_3 \to L_3 - L_1/3$ (que nous pourrions savoir dès à présent), il est important de passer par la matrice gaussienne pour obtenir la matrice L_1.

La matrice de transformation gaussienne pour la première colonne de (36.331) est :

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2/3 & 1 & 0 \\ -1/3 & 0 & 1 \end{pmatrix} \tag{36.332}$$

et $L_1 = M_1^{-1}$. Le lemme 36.121 nous dit comment calculer facilement cet inverse :

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ 2/3 & 1 & 0 \\ 1/3 & 0 & 1 \end{pmatrix} \tag{36.333}$$

En suivant l'équation (36.323) nous posons $A_1 = M_1 A_0'$:

$$A_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2/3 & 1 & 0 \\ -1/3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 & 8 & 0 \\ 2 & 5 & 0 \\ 1 & 3 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 8 & 0 \\ 0 & -1/3 & 0 \\ 0 & -2/3 & 3 \end{pmatrix} \tag{36.334}$$

et nous avons

$$Q_1 A = L_1 A_1 \tag{36.335}$$

où L_1, A_1 et $r_1 = 3$ sont enregistrés. La matrice Q_1 peut être retrouvée en sachant r_1 parce que P est la matrice de permutation $P_{r_1,1}$.

Nous travaillons maintenant sur la deuxième colonne de A_1. Le plus grand élément en valeur absolue (sur ou sous la diagonale) est $-2/3$. Nous posons $r_2 = 3$ et

$$A_1' = \begin{pmatrix} 3 & 8 & 0 \\ 0 & -2/3 & 3 \\ 0 & -1/3 & 0 \end{pmatrix} \tag{36.336}$$

et la matrice gaussienne pour la deuxième colonne est

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1/2 & 1 \end{pmatrix} \tag{36.337}$$

Le $-1/2$ provient du calcul $-\big((-1/3)/(-2/3)\big)$. L'inverse de cette matrice est facile :

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/2 & 1 \end{pmatrix} \tag{36.338}$$

et la matrice suivante à enregistrer est

$$A_2 = M_2 P_{3,2} A_1 = M_2 A_1' = \begin{pmatrix} 3 & 8 & 0 \\ 0 & -2/3 & 3 \\ 0 & 0 & -3/2 \end{pmatrix}. \tag{36.339}$$

Notons toutefois que pour calculer cette matrice, seul le dernier élément demande un calcul. La première colonne ne change pas (par construction), la seconde gagne un zéro en dernière ligne (la matrice M_2 sert à ça) et sur la dernière colonne, seule la dernière ligne est sujette à changement.

Avec la matrice A_2, la trigonalisation supérieure est faite. La décomposition n'est cependant pas terminée. Nous devons encore trouver la partie triangulaire inférieure. Nous en sommes à

$$Q_1 A = L_1 A_1 = L_1 P_{3,2} L_2 A \tag{36.340}$$

où Q_1 est la première matrice de permutation.

Utilisant le lemme 36.130, il est facile de permuter L_1 avec $P_{3,2}$:

$$L_1 P_{3,2} = P_{3,2} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 0 & 1 \end{pmatrix}}_{L_1'} \tag{36.341}$$

Nous avons donc

$$P_{3,2}P_{3,1}A = L_1'L_2A \tag{36.342}$$

Deux multiplications matricielles plus tard nous terminons :

$$PA = LU \tag{36.343}$$

avec

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{pmatrix}, \quad U = A_2 = \begin{pmatrix} 3 & 8 & 0 \\ 0 & -2/3 & 3 \\ 0 & 0 & -3/2 \end{pmatrix}. \tag{36.344}$$

△

Notons que Sage utilise la méthode de Gauss avec pivots :

```
sage: A=matrix([ [1,2,3],[2,5,0],[3,8,0] ])
sage: A
[1 2 3]
[2 5 0]
[3 8 0]
sage: A.LU()
(
[0 1 0]  [  1   0   0]  [   3    8    0]
[0 0 1]  [1/3   1   0]  [   0 -2/3    3]
[1 0 0], [2/3 1/2   1], [   0    0 -3/2]
)
```

tex/sage/sageSnip006.sage

Mais attention : Sage crée une décomposition $A = PLU$ et non $PA = LU$. D'où le fait que la matrice de permutation de Sage est l'inverse de celle donnée ici.

36.18.3 D'un point de vue algorithmique

Problèmes et choses à faire

Je ne suis pas certain de l'optimalité de ce que je raconte ici. Je décris simplement ce que j'ai fait pour écrire mon programme finitediff.

Si vous êtes expert en calcul numérique, n'hésitez pas à donner votre avis.

Nous décrivons à présent la décomposition $A = PLU$ (du théorème 36.131, avec le P à droite). En suivant l'exemple 36.132 nous voyons assez bien comment créer les matrices U et P au fur et à mesure. La construction de L est peut-être moins évidente.

Écrivons un exemple très explicite pour

$$A = \begin{pmatrix} 2 & 1 & 3 \\ 4 & 3 & 10 \\ -2 & 1 & 7 \end{pmatrix}. \tag{36.345}$$

Nous commençons par permuter des lignes pour avoir un grand pivot :

$$P_{12}A = \begin{pmatrix} 4 & 3 & 10 \\ 2 & 1 & 3 \\ -2 & 1 & 7 \end{pmatrix}. \tag{36.346}$$

Et nous effectuons l'élimination avec la matrice

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix}. \tag{36.347}$$

Cela donne le premier résultat :

$$M_1P_{12}A=\begin{pmatrix}4&3&10\\0&-1/2&-2\\0&5/2&12\end{pmatrix}\tag{36.348}$$

Nous continuons avec P_{23} pour avoir un nouveau grand pivot :

$$P_{23}M_1P_{12}A=\begin{pmatrix}4&3&10\\0&5/2&12\\0&-1/2&-2\end{pmatrix}.\tag{36.349}$$

Nous utilisons la matrice

$$M_2=\begin{pmatrix}1&0&0\\0&1&0\\0&1/5&1\end{pmatrix}\tag{36.350}$$

et au final :

$$M_2P_{23}M_1P_{12}A=\begin{pmatrix}4&3&10\\0&5/2&12\\0&0&2/5\end{pmatrix}=U.\tag{36.351}$$

L'égalité obtenue est

$$M_2P_{23}M_1P_{12}A=U.\tag{36.352}$$

Pour avoir la décomposition PLU il faut écrire

$$A=P_{12}M_1^{-1}P_{23}M_2^{-1}U,\tag{36.353}$$

et permuter P_{23} avec M_1^{-1}, ce qui est facile par le lemme 36.130.

Remarque 36.133.
Nous ne devons permuter la matrice M_k avec une matrice de permutation qu'à partir de la deuxième étape. En effet l'équation (36.348) revient à

$$A=M_1^{-1}P_{12}m_U\tag{36.354}$$

qui est dans le bon ordre. Ce n'est qu'à partir de la seconde étape que des matrices de permutation apparaissent à droite des matrices gaussiennes.

Cependant dans un cas 4×4, cette méthode deviendrait fastidieuse parce que nous aurions encore des étapes à faire. En repartant de (36.353), mais avec m_U (la matrice pas encore tout à fait triangularisée) au lieu de U, nous aurons, pour un certain $k>3$:

$$M_3P_{3k}M_2P_{23}M_1P_{12}A=U,\tag{36.355}$$

ce qui fait :

$$A=P_{12}M_1^{-1}P_{23}M_2^{-1}P_{3k}M_3^{-1}U.\tag{36.356}$$

Tous les P_{ij} peuvent être mis à gauche parce que leurs indices sont toujours strictement supérieurs à ceux des M_l placés devant eux. Mais c'est fastidieux.

Nous allons donc permuter à chaque étape pour ne retenir que l'important. Si à une certaine étape nous avons

$$A=P_{1,r_1}\dots P_{k,r_k}M_1^{-1}\dots M_k^{-1}m_U\tag{36.357}$$

avec

$$m_U=P_{k+1,r_{k+1}}M_{k+1}^{-1}U\tag{36.358}$$

alors nous allons directement permuter $P_{k+1,r_{k+1}}$ avec tous les M_i^{-1}. Si nous notons P_k la permutation (pas élémentaire) à l'étape k et L_k la matrice triangulaire inférieure à de l'étape k,

$$A=P_kL_km_U=P_kL_kP_{k+1,r_{k+1}}M_{k+1}^{-1}m'_U.\tag{36.359}$$

Nous enregistrons alors $P_{k+1}=P_kP_{k+1,r_{k+1}}$ et pour L_{k+1} nous partons de L_k et nous faisons deux opérations suivantes :

— nous permutons, sur ses colonnes non triviales, les indices $k+1$ et r_{k+1},
— nous multiplions par M_{k+1}^{-1}, ce qui revient à simplement lui ajouter une colonne non triviale.

Notons que $r_{k+1} \geqslant k+1$, de telle sorte que sur les colonnes non triviales (qui sont jusqu'au numéro k), la permutation des lignes $k+1$ et r_{k+1} ne change pas l'aspect de la matrice : elle reste multi-gaussienne de dernière colonne k.

36.18.4 Exemples

Nous nous lançons dans la résolution du système

$$\begin{pmatrix} 10^{-9} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{36.360}$$

Exemple 36.134
Nous commençons de façon exacte, par la méthode de Gauss sans pivot. La première transformation gaussienne est

$$E_1 = \begin{pmatrix} 1 & 0 \\ -10^9 & 1 \end{pmatrix} \tag{36.361}$$

et nous calculons

$$E_1 A = \begin{pmatrix} 1 & 0 \\ -10^9 & 1 \end{pmatrix} \begin{pmatrix} 10^{-9} & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 10^{-9} & 1 \\ 0 & 1-10^{-9} \end{pmatrix}. \tag{36.362}$$

Vu que cette dernière est triangulaire supérieure, nous avons fini la méthode de Gauss et $U = E_1 A$. En ce qui concerne la matrice L, elle est donnée par $L = E_1^{-1}$, c'est à dire

$$L = \begin{pmatrix} 1 & 0 \\ -10^9 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 10^9 & 1 \end{pmatrix}. \tag{36.363}$$

Au final nous avons la décomposition $A = LU$ exacte suivante :

$$L = \begin{pmatrix} 1 & 0 \\ 10^9 & 1 \end{pmatrix} \quad U = \begin{pmatrix} 10^{-9} & 1 \\ 0 & 1-10^9 \end{pmatrix}. \tag{36.364}$$

Résoudre le système $Ax = b$ revient à résoudre $LUx = b$ et donc résoudre successivement les systèmes

$$\begin{cases} Ly = b & \text{(36.365a)} \\ Ux = y. & \text{(36.365b)} \end{cases}$$

D'abord le système

$$\begin{pmatrix} 1 & 0 \\ 10^9 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{36.366}$$

donne $y_1 = 1$ et $y_2 = 2 - 10^9$.

Ensuite nous résolvons

$$\begin{pmatrix} 10^{-9} & 1 \\ 0 & 1-10^9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2-10^9 \end{pmatrix}. \tag{36.367}$$

Cela donne

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -\frac{10^9}{1-10^9} \\ \frac{2-10^9}{1-10^9} \end{pmatrix} \simeq \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{36.368}$$

C'est également le résultat que trouve Sage :

```
sage: var('y')
y
sage: solve(    [10**(-9)*x+y==1,x+y==2],[x,y]    )
[[x == (1000000000/999999999), y == (999999998/999999999)]]
```

tex/sage/sageSnip007.sage

△

Exemple 36.135([409])
Nous recommençons tout le calcul avec une précision limitée à 8 chiffres significatifs, sans pivot.

Nous avons à nouveau la transformation gaussienne

$$E_1 = \begin{pmatrix} 1 & 0 \\ -10^9 & 1 \end{pmatrix}, \tag{36.369}$$

mais pour calculer U nous effectuons le produit matriciel

$$U = E_1 A = \begin{pmatrix} 1 & 0 \\ -10^9 & 1 \end{pmatrix} \begin{pmatrix} 10^{-9} & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 10^{-9} & 1 \\ 0 & * \end{pmatrix}. \tag{36.370}$$

Nous détaillons à présent le calcul de l'élément noté $*$. Le calcul de $10^9 \ominus 1)$ donne

$$999999999 = 9.99999999 \times 10^8, \tag{36.371}$$

mais la précision étant limitée à 8 chiffres, un arrondit arrive. Étant donné que le premier chiffres supprimé est un 9 nous retombons sur 10^9, et donc notre machine à précision limitée donnera

$$U = \begin{pmatrix} 10^{-9} & 1 \\ 0 & -10^9 \end{pmatrix}. \tag{36.372}$$

Ensuite le calcul de $L = E_1^{-1}$ ne cause pas de problèmes :

$$L = \begin{pmatrix} 1 & 0 \\ -10^9 & 1 \end{pmatrix}. \tag{36.373}$$

Maintenant il s'agit de résoudre les systèmes $Ly = b$ et $Ux = y$. Du système

$$\begin{pmatrix} 1 & 0 \\ 10^9 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{36.374}$$

nous tirons tout de suite $y_1 = 1$ et ensuite $10^9 + y_2 = 2$, c'est à dire $y_2 = 2 - 10^9$, qui en précision limitée donne encore $y_2 = -10^9$. À résoudre maintenant :

$$\begin{pmatrix} 10^{-9} & 1 \\ 0 & -10^9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -10^9 \end{pmatrix}. \tag{36.375}$$

Cela donne immédiatement $x_2 = 1$ et ensuite

$$10^{-9} x_1 + 1 = 1, \tag{36.376}$$

donc $x_1 = 0$. La solution trouvée est

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{36.377}$$

qui est complètement faux au niveau de la première variable. △

Exemple 36.136([409])
Nous résolvons encore le même système en précision limitée, mais en utilisant cette fois la méthode de Gauss avec pivot partiel.

Le plus grand élément de la première colonne est 1 ; nous utilisons donc la permutation $P_{1,2}$:

$$P_{1,2} A = \begin{pmatrix} 1 & 1 \\ 10^{-9} & 1 \end{pmatrix}. \tag{36.378}$$

La matrice de transformation gaussienne pour la première colonne de cette matrice est

$$M_1 = \begin{pmatrix} 1 & 0 \\ -10^{-9} & 1 \end{pmatrix} \tag{36.379}$$

et nous posons

$$A_1 = M_1 P_{1,2} A = \begin{pmatrix} 1 & 0 \\ -10^{-9} & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 10^{-9} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & -10^{-9}+1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = U \tag{36.380}$$

où un arrondi a eu lieu pour $-10^{-9} + 1 = 1$. En inversant M_1 nous avons

$$L_1 = M_1^{-1} = \begin{pmatrix} 1 & 0 \\ 10^{-9} & 1 \end{pmatrix}. \tag{36.381}$$

La décomposition est

$$A = \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}}_{P} \underbrace{\begin{pmatrix} 1 & 0 \\ 10^{-9} & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}}_{U} \tag{36.382}$$

Le moment de résoudre est venu. Vu que $PLUx = b$ nous devons résoudre les systèmes

$$\begin{cases} Pz = b & \text{(36.383a)} \\ Ly = z & \text{(36.383b)} \\ Ux = y. & \text{(36.383c)} \end{cases}$$

Pour z c'est facile :

$$z = \begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{36.384}$$

Pour y il y a un arrondi :

$$\begin{pmatrix} 1 & 0 \\ 10^{-9} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{36.385}$$

Tout de suite : $y_1 = 2$ et ensuite $2 \times 10^{-9} + y_2 = 1$, ce qui donne $y_2 = 1 \ominus 2 \times 10^{-9} = 1$. Donc

$$y = \begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{36.386}$$

Et enfin pour x c'est le système

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{36.387}$$

Nous avons $x_2 = 1$ et ensuite $x_1 + 1 = 2$ c'est à dire $x_1 = 1$. Au final la solution trouvée est

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{36.388}$$

Cette solution est considérablement meilleure que (36.377). △

36.137.
L'utilisation du pivot non seulement assure le fait que la trigonalisation va bien se passer (on évite les zéros en pivot), mais aussi et surtout, en choisissant de prendre le plus grand pivot possible, nous obtenons une meilleur stabilité numérique.

36.19 Résolution de systèmes linéaires (suite)

36.19.1 Déterminant

Pour calculer un déterminant lorsque nous avons la décomposition $A = LU$ nous pouvons faire

$$\det(A) = \det(LU) = \det(L)\det(U) = \det(U) \tag{36.389}$$

parce que L est triangulaire avec des 1 sur la diagonale.

Si par contre nous avons fait des pivots, nous avons $PA = LU$. Il nous faut le déterminant de P, qui n'est autre que ± 1. Nous avons

$$\det(P) = (-1)^s \tag{36.390}$$

où s est le nombre de permutations effectives effectuées. Nous précisons « effectives » parce qu'il ne faut pas compter le pas où nous n'avons pas permuté (les cas où le bon pivot était présent du premier coup). Nous avons alors

$$\det(A) = (-1)^s \det(U). \tag{36.391}$$

36.19.2 Plusieurs termes indépendants

Mettons un système $Ax = b$ qu'il faut résoudre pour plusieurs b différents. C'est typiquement le cas où l'on voudrait calculer l'inverse de A. Mais on va directement se calmer. Soient donc à résoudre $Ax_1 = b_1$, …, $Ax_n = b_n$.

Les opérations (avec ou sans pivot) que nous faisons ne dépendent que de la matrice A, mais aucune décisions concernant les pivots ou la matrice des multiplicateurs ne dépend de b. Autre façon de dire : si le système $(A|b_1)$ devient $(U|y_1)$, le système $(A|b_i)$ devient $(U|y_i)$ avec le même U.

Nous ne sommes donc pas obligés de faire tout le travail autant de fois qu'il n'y a de systèmes à résoudre. Donc si on a plusieurs systèmes à résoudre avec la même matrice, on fait mieux de retenir une fois pour toute la décomposition LU (avec ou sans pivots), avant de vraiment résoudre.

Ou alors on peut aussi faire que, au lieu de faire $(A|b_i)$ plein de fois, faire une seule fois

$$(A|b_1 \dots b_n). \tag{36.392}$$

Et on fait tout le travail sur tous les vecteurs d'un en même temps.

Soit e_i la base canonique. Si nous notons x_n les solutions des problèmes $Ax_i = e_i$, tous les problèmes $Ax_i = e_i$ s'écrivent d'un seul coup

$$AX = Y \tag{36.393}$$

où X est la matrice des x_i en colonnes, et Y est celle des e_i en colonnes. Oh, mais $Y = \mathbb{1}$ évidemment. Donc

$$AX = \mathbb{1}. \tag{36.394}$$

Si nous supposons que A est inversible, alors ce X est l'inverse.

Donc pour calculer l'inverse d'une matrice de dimension non trop grande, il suffit d'utiliser la méthode de Gauss sur les vecteurs de la base canonique. Cette idée est la base du calcul de l'inverse par matrice companion. En effet, si nous partons du problème

$$(A|\mathbb{1}) \tag{36.395}$$

et nous appliquons la méthode de Gauss avec pivot, nous arrivons à

$$(U|L^{-1}P). \tag{36.396}$$

Attention : le produit $L^{-1}P$ est une permutation des *colonnes* de L^{-1}. Vu que L est triangulaire inférieure avec des 1 sur la diagonale, L^{-1} est triangulaire inférieure avec des 1 sur la diagonale.

Donc si la matrice n'est pas trop grande, on peut assez facilement remettre les colonnes de $L^{-1}P$ dans l'ordre pour recomposer une matrice triangulaire inférieure avec 1 sur la diagonale.

Une autre façon de calculer l'inverse, si $A = LU$ est connue, il suffit de faire

$$A^{-1} = U^{-1}L^{-1}. \tag{36.397}$$

Et il existe un algorithme facile pour l'inverse d'une matrice triangulaire.

36.19.3 Cholesky

Le commandant Cholesky travaillait sur le tir de canon (chose éminemment liée à de nombreuses mathématiques ingénieuses). La méthode de Cholesky est encore utilisée aujourd'hui dans les vrais problèmes.

La méthode de Gauss s'applique sans hypothèses sur la matrice A, à part qu'elle doit être de petite dimension, comme pour toute méthode directe. Souvent nous savons des choses sur la matrice. Ici nous allons supposer que A est symétrique et définie positive.

Comment numériquement vérifier ces hypothèses ? En ce qui concerne la symétrique, il suffit de faire le test complet :

$$A^t = A. \tag{36.398}$$

La vérification de cela coûte au maximum n^2 comparaisons (et en fait la moitié de ça moins la diagonale).

Le fait que A soit définie positive est facile à vérifier pour utiliser Cholesky parce que il suffit de le faire, et s'il n'y a pas de nombres complexes qui arrivent, c'est que la matrice était définie positive.

Un lemme très simple à mettre en oeuvre numériquement nous permet de traiter certains cas.

Lemme 36.138.
Une matrice symétrique possédant un élément négatif sur la diagonale n'est pas définie positive.

Démonstration. Un simple calcul ou effort d'imagination montre que $\langle Me_k, e_k\rangle = M_{kk}$. Donc si M doit être définie positive, M_{kk} doit être positive par le lemme 11.181. □

Ce lemme est un moyen déjà de faire quelques vérifications. Et si les éléments diagonaux de A sont tous négatifs, on peut prendre $-A$.

Lemme 36.139.
Si A est une matrice symétrique strictement définie positive, alors pour tout k, la matrice tronquée $\Delta_k(A)$ l'est également.

Démonstration. Le fait que $\Delta_k(A)$ soit symétrique est évidemment. Le fait qu'elle soit définie positive l'est moins. Soit $y \in \mathbb{R}^k$ et le vecteur $\tau y \in \mathbb{R}^n$, qui est « complété » avec des zéros.

Nous avons $\langle \Delta_k(A)y, y\rangle_k = \langle A\tau y, \tau\rangle_n$. En effet

$$\langle \Delta_k(A)y, y\rangle = \sum_{i=1}^{k}\sum_{l=1}^{k} A_{il}y_l y_i. \tag{36.399}$$

Et à droite :

$$\langle A\tau y, \tau y\rangle = \sum_{i=1}^{n}(A\tau y)_i(\tau y)_i = \sum_{i=1}^{k}(A\tau y)_i y_i = \sum_{i=1}^{k}\sum_{l=1}^{n} A_{il}(\tau y)_l y_i = \sum_{i=1}^{k}\sum_{l=1}^{k} A_{il}y_l y_i \tag{36.400}$$

où nous avons utilisé le fait que $(\tau y)_i = 0$ dès que $i > k$ et que $(\tau y)_i = y_i$ sinon.

En conséquence de quoi $\langle \Delta_k(A)y, u\rangle > 0$ pour tout $y \in \mathbb{R}^k$ et la matrice $\Delta_k(A)$ est strictement définie positive. □

Lemme 36.140.
Si T est une matrice triangulaire, alors $(T_{ii})^{-1} = (T^{-1})_{ii}$.

Démonstration. Il suffit de se rendre compte que le coefficient ii de l'égalité $\mathbb{1} = TT^{-1}$ donne

$$1 = \sum_l T_{il}(T^{-1})li. \tag{36.401}$$

Dans la somme il ne reste que le terme $l = i$. □

Nous allons chercher une décomposition de type LU sous la forme $A = LL^t$, c'est à dire $U = L^t$. Attention : maintenant nous n'avons plus des 1 sur la diagonale. Ce n'est donc pas exactement la décomposition LU dont nous parlions plus haut. C'est pour cela que nous n'allons pas la noter LL^t mais BB^t.

Théorème 36.141 (Cholesky[408]).
Soit une matrice réelle symétrique strictement définie positive. Il existe une unique matrice réelle B telle que

- *B est triangulaire inférieure,*
- *la diagonale de B est positive,*
- *$A = BB^t$.*

Démonstration. Par la décomposition LU du théorème 36.123 nous avons des matrices L et U telles que $A = LU$. Soit D la matrice diagonale donnée par

$$D_{ii} = \sqrt{U_{ii}}. \tag{36.402}$$

Cette définition fonctionne parce que $U_{ii} > 0$. En effet nous savons que $\Delta_k(A) = \Delta_k(L)\Delta_k(U)$, et en passant au déterminant,

$$\det\big(\Delta_k(A)\big) = \det\big(\Delta_k(U)\big). \tag{36.403}$$

Vu que $\Delta_k(A)$ est strictement définie positive par le lemme 36.139, son determinant est strictement positif[20] et nous avons

$$\det\big(\Delta_k(U)\big) > 0. \tag{36.404}$$

En appliquant cela à $k = 1$ nous avons $U_{11} > 0$ puis de proche en proche, $U_i i > 0$ pour tout i.

Nous posons :

$$B = LD \qquad \text{qui est triangulaire inférieure} \tag{36.405a}$$

$$C = D^{-1}U \qquad \text{qui est triangulaire supérieure.} \tag{36.405b}$$

Nous avons bien entendu $A = BC$ et nous allons prouver que $C = B^t$. Vu que $A = A^t$ nous pouvons identifier BC et C^tB^t :

$$BC = C^tB^t. \tag{36.406}$$

En mettant les matrices triangulaires supérieures à gauche et inférieures à droite :

$$C(B^t)^{-1} = B^{-1}C^t, \tag{36.407}$$

qui sont donc deux matrices diagonales. Nous montrons que cette diagonale est en réalité l'identité.

D'abord

$$B_{ii} = \sum_{l=1}^{n} L_{il}D_{li} = L_{ii}\sqrt{U_{ii}} = \sqrt{U_{ii}} \tag{36.408}$$

parce que $L_{ii} = 1$. Notons en passant que la diagonale de B est positive. Ensuite

$$C_{ii} = \sum_{l=1}^{n}(D^{-1})_{il}U_{li} = (D^{-1})_{ii}U_{ii} = \frac{1}{\sqrt{U_i i}}U_{ii} = \sqrt{U_{ii}}. \tag{36.409}$$

20. Le théorème 11.174 donne une diagonalisation par des matrices de déterminant 1. Vu que les valeurs propres forment sur la diagonale, et qu'elles sont toutes positives, el déterminant est positif.

Donc B et C ont des diagonales égales. Calculons alors la diagonale de $B^{-1}C^t$:

$$\left(B^{-1}C^t\right)_{ii} = \sum_l (B^{-1})_{il}(C^t)_{li} = (B^{-1})_{ii}C_{ii} \tag{36.410}$$

parce que encore une fois, de la somme il ne reste que le terme $l = i$.

Mais B est une matrice triangulaire qui tombe sous le coup du lemme 36.140. Donc $(B^{-1})_{ii} = (B_{ii})^{-1} = (C_{ii})^{-1}$. Nous avons alors

$$(B^{-1}C^t)_{ii} = 1. \tag{36.411}$$

Cela conclu l'existence de la décomposition de Cholesky.

En ce qui concerne l'unicité, soient $A = BB^t = CC^t$. Nous regroupons les supérieures et les inférieures :

$$B^t(C^t)^{-1} = B^{-1}C. \tag{36.412}$$

Ces deux matrices sont donc diagonales et nous posons $D = B^{-1}C$, c'est à dire $C = BD$. Nous remplaçons donc C par BD dans (36.412) :

$$A = BB^t = BD(BD)^t = BDD^tB^t. \tag{36.413}$$

Donc $DD^t = \mathbb{1}$, ce qui signifie que les éléments diagonaux de D sont ± 1. Nous montrons qu'ils sont positifs : à partir de $C = BD$ nous déballons

$$C_{ii} = \sum_l B_{il}D_{li}, \tag{36.414}$$

et donc

$$B_{ii}D_{ii} = C_{ii}. \tag{36.415}$$

En sachant que les conditions de la décomposition de Cholesky demandent les éléments diagonaux positifs nous en déduisons que D_{ii} est positif et donc égal à 1. Finalement $D = \mathbb{1}$ et $B = C$. □

Prenons la matrice

$$A = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 10 & -7 \\ -2 & -7 & 9 \end{pmatrix} \tag{36.416}$$

Elle est symétrique et définie positive. Nous posons

$$\{ \quad l_{11} = \sqrt{a_{11}} l_{i1} = a_{i1}/l_{11} \tag{36.417a}$$

pour $i = 2, \ldots, n$. Et aussi

$$\begin{cases} l_{jj} = (a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2)^{1/2} & \text{(36.418a)} \\ l_{ij} = (a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk})/a_{jj} & \text{(36.418b)} \end{cases}$$

pour $i = j + 1, \ldots, n$.

Les formules (36.417) nous disent comment remplir la première colonne. Cela donne la matrice

$$L = \begin{pmatrix} \sqrt{4} = 2 & . & . \\ 2/2 = 1 & . & . \\ -2/2 = -1 & . & . \end{pmatrix} \tag{36.419}$$

Les formules (36.418) donnent les autres colonnes en fonction des précédentes.

Dans Sage :

```
sage: A=matrix(  [  [4,2,-2],[2,10,-7],[-2,-7,9]  ]   )
sage: A
[ 4  2 -2]
[ 2 10 -7]
[-2 -7  9]
sage: A.cholesky()
[ 2  0  0]
[ 1  3  0]
[-1 -2  2]
```

tex/sage/sageSnip005.sage

36.20 Système linéaire (méthodes itératives)

Nous trouvons des méthodes itératives lorsque les matrices sont grandes, ce qui arrive lorsque l'on discrétise une équation différentielle.

Nous allons chercher des méthodes de la forme $x_{n+1} = Bx_n + q$; ce sont des méthodes stationnaires. La convergence d'une méthode est toujours liée à la matrice B et en général, la convergence ne dépend pas du choix du vecteur initial. Nous faisons donc souvent $x_0 = 0$ et donc $x_1 = q$. Voilà donc une itération de faite gratuitement.

Nous notons e_k le **vecteur d'erreur** qui est définit par $e_k = x - x_k$. Et le **vecteur résidu** $r_k = b - Ax_k$. Attention : ici k n'est pas un indice mais un numéro de vecteur.

Notons que si x est solution, alors $b - Ax = 0$, ce qui motive le vecteur résidu.

Les conditions d'arrêt d'un algorithme seraient

$$\begin{cases} \|e_k\|_\infty \ll \epsilon_1 & \text{(36.420a)} \\ \|r_k\|_\infty < \epsilon_2 & \text{(36.420b)} \end{cases}$$

où ϵ_1 et ϵ_2 sont des précisions décidées à l'avance par l'utilisateur.

Proposition 36.142.
Si A est une matrice inversible, alors

$$\lim_{k\to\infty} e_k = \lim_{k\to\infty} r_k = 0. \tag{36.421}$$

Vu que $r_k = Ae_k$, si la matrice A est mal conditionnée, il peut arriver que r_k reste grand alors que e_k est déjà petit.

Remarque 36.143.
Dans les méthode stationnaires, nous avons $x_{n+1} = Bx_n + q$ avec B et q fixés au départ de l'algorithme. Il existe des méthodes non stationnaires pour lesquelles l'itération prend la forme $x_{n+1} = B_n x_n + q_n$ avec B_n et q_n qui changement avec les étapes.

Proposition 36.144.
Pour la méthode $x_{n+1} = Bx_n + q$ nous avons équivalence de

(1) La méthode converge pour tout x_0

(2) B est une matrice convergente[21]

(3) $\rho(B) < 1$ (rayon spectral).

De plus si $\|B\| < 1$ alors la méthode converge (quelle que soit la norme algébrique).

La norme d'une matrice (en tout cas, certaines normes) est quelque chose de facile à calculer à l'ordinateur. Typiquement $\|.\|_\infty$ est un simple maximum. Cependant si après avoir calculé $\|B\|_i$ pour des dizaines de normes i différentes, nous avons toujours $\|B\|_i \geqslant 1$, alors nous ne pouvons rien conclure.

21. C'est à dire $\lim_{k\to\infty} B^k = 0$.

36.20.1 La méthode générale

Nous décomposons la matrice A sous la forme $A = M - N$ avec M inversible. Le système $Ax = b$ devient

$$Mx - Nx = b \tag{36.422}$$

puis $Mx = Nx + b$ et finalement

$$x = M^{-1}Nx + M^{-1}b, \tag{36.423}$$

et voilà une méthode stationnaire avec $B = M^{-1}N$ et $q = M^{-1}b$.

Mais ici nous voyons que M doit être non seulement inversible, mais en plus doit être facilement calculable. En sachant que nous travaillons avec des grandes matrices, il n'est pas question d'inverser M avec une méthode de Gauss.

En bref, il faut choisir M triangulaire parce que c'est en gros la seule que nous pouvons inverser facilement [22].

Remarque 36.145.
La matrice B ne doit pas spécialement être inversible. Si elle ne l'est pas, ce n'est pas un problème.

36.20.2 Jacobi

Nous décomposons

$$A = D - E - F \tag{36.424}$$

où D est la diagonale de A, $-F$ est la partie triangulaire supérieure (sans la diagonale) et $-E$ la triangulaire inférieure (sans la diagonale). Donc D, E et F sont simplement des extractions de parties de la matrice A (et quelques changements de signes).

La méthode de Jacobi prend $M = D$ et $N = (E + F)$. L'inverse de M est facile à calculer parce que M est diagonale. Nous notons B_J la matrice B de la méthode de Jacobi.

Remarque 36.146.
Il se peut que la matrice A ait des zéros sur la diagonale, même si elle est inversible. Et cela est un problème parce qu'alors la matrice D ici construite n'est pas inversible. Dans ce cas, avant de nous lancer dans la méthode de Jacobi, il faut permuter deux lignes de A et donc de b.

Attention cependant que l'on pourrait vouloir effectuer ces permutations en mettant sur la diagonale des nombres les plus grands possibles (parce qu'ensuite, ce qui rentre dans les calculs, c'est D^{-1} qui aura alors des petits nombres). Mais il faut toutefois faire en sorte que le rayon spectral de la matrice B résultante reste plus petit que 1.

Chaque changement dans A induit des changements dans B et donc sur la convergence de la méthode.

36.20.3 Gauss-Seidel

Nous partons de la même décomposition $A = D - E - F$ que dans (36.424). La méthode de Gauss-Seidel prend $M = (D - E)$ et $N = F$.

36.20.4 Autres

Voir la méthode des gradients, et des gradients conjugués.

36.21 Indices connectés, matrice irréductible

Définition 36.147.
*Soit $A \in \mathbb{M}(n, \mathbb{R})$ et $i, j \in \{1, \dots, n\}$. Nous disons que les indices i et j sont **directement connectés** si $A_{ij} \neq 0$ ou $A_{ji} \neq 0$.*

22. Les matrices orthogonales sont aussi facilement inversibles, mais ne se prêtent pas bien à une décomposition de type somme.

Définition 36.148.
*Soit $A \in \mathbb{M}(n, \mathbb{R})$ et $i, j \in \{1, \dots, n\}$. Nous disons que les indices i et j sont **connectés** si il existe un ensemble d'indices $i_0 = i, i_1, \dots, i_{r-1}, i_r = j$ tels que $A_{i_k, i_{k+1}} \neq 0$ pour tout $0 \leqslant k \leqslant r$.*

Par exemple pour que les indices 1 et 4 soient connectés, on peut avoir les éléments A_{13}, A_{32} A_{24} non nuls.

Définition 36.149 ([410]).
*Une matrice carrée A est **réductible** si il existe une permutation σ telle que*

$$\sigma^t A \sigma = \begin{pmatrix} K & L \\ 0 & M \end{pmatrix} \tag{36.425}$$

où K et M sont carrées.

Notons que par définition de la matrice d'une application linéaire,

$$B_{ij} = \langle e_i, Be_j \rangle = \langle e_i, \sigma^t A \sigma e_j \rangle = \langle \sigma e_i, A \sigma e_j \rangle = A_{\sigma(i), \sigma(j)}. \tag{36.426}$$

Proposition 36.150 ([410]).
Soit une matrice carrée A. Les faits suivants sont équivalents :

(1) A est réductible.
(2) Il existe une partition non triviale I, J de $\{1, \dots, n\}$ telle que $I \cup J = \{1, \dots, n\}$, $I \cap J = \varnothing$ et pour tout $i \in I$, et pour tout $j \in J$, $A_{ij} = 0$.
(3) La matrice A admet des indices non connectés (définition 36.148).

Démonstration. Dans plusieurs sens...

(1) implique (2) Nous notons j^* la taille de la matrice K dans (36.425). Nous avons $B_{ij} = 0$ si

$$\begin{cases} J^* + 1 \leqslant i \leqslant n & \text{(36.427a)} \\ 1 \leqslant j \leqslant j^*. & \text{(36.427b)} \end{cases}$$

Donc en posant $I = \sigma\{j^* + 1, \dots, n\}$ et $J = \sigma\{1, \dots, j^*\}$ nous avons une partition non triviale de $\{1, \dots, n\}$ telle que si $i \in I$ et $j \in J$ alors $i = \sigma(i_0)$, $j = \sigma(j_0)$ et

$$A_{ij} = A_{\sigma(i_0), \sigma(j_0)} = B_{i_0, j_0} = 0. \tag{36.428}$$

(2) implique (1) Soit une partition I, J comme indiquée dans l'hypothèse. Soit j^* le nombre d'éléments dans J. Soit σ une permutation de $\{1, \dots, n\}$ telle que $\sigma\{j^* + 1, \dots, n\} = I$ et $\sigma\{1, \dots, j^*\} = J$. Nous posons ensuite $B = \sigma^t A \sigma$. Par construction si $i \in I$ et $j \in J$ alors $A_{ij} = 0$.
Mais si

$$\begin{cases} J^* + 1 \leqslant i \leqslant n & \text{(36.429a)} \\ 1 \leqslant j \leqslant j^*. & \text{(36.429b)} \end{cases}$$

alors $B_{ij} = A_{\sigma(i)\sigma(j)} = 0$. Donc B a la bonne forme.

(3) implique (2) Soient i et j deux indices non connectés : il n'existe pas de chaînes partant de i et arrivant à j. Nous notons I l'ensemble des indices connectés à i, et J les autres. Pat hypothèses ces ensembles sont non vides.
Si $k \in i$ et $l \in J$ alors $A_{kl} = 0$ parce que sinon on aurait une chaine de i à k puis de k à l et donc de i à l, ce qui signifierait que l est connecté à i.

(2) implique (3) Soit une partition I, J comme dans l'hypothèse. Si $j \in J$ est connecté à $i \in I$ alors il existe une chaîne

$$i = i_0, i_1, \dots, i_r = j. \tag{36.430}$$

Si i_s est le premier dans J alors $i_{s-1} \in I$ et $A_{i_{s-1}, i_s} = 0$, ce qui empêche la chaîne de connecter j à i.

□

36.22 Localisation des valeurs propres

Sur l'ensemble $\mathbb{M}(n,\mathbb{R})$ des matrices $n\times n$ à coefficients réels nous introduisons l'ordre partiel [23] donné par $A\geqslant B$ lorsque $A_{ij}\geqslant B_{ij}$ pour tout i et j. Nous définissons de façon similaire les relations $A\leqslant B$, $A<B$ et $A>B$.

Si $x\in\mathbb{R}^n$ nous notons $|x|=\big(|x_1|,\dots,|x_n|\big)$ et $x\leqslant y$ lorsque $x_i\leqslant y_i$ pour tout i.

Proposition 36.151.
Soit $A\in\mathbb{M}(n,\mathbb{R})$ et $x,y\in\mathbb{R}^n$.

(1) Si $A\geqslant 0$ et si $x\leqslant y$ alors $Ax\leqslant Ay$.

(2) Si $A\geqslant 0$ alors $Ax\leqslant|Ax|\leqslant A|x|$.

Démonstration. Pour la première inégalité, pour tout i et k nous avons $A_{ik}x_k\leqslant A_{ik}y_k$ et donc

$$(Ax)_i=\sum_k A_{ik}x_k\leqslant\sum_k A_{ik}y_k=(Ay)_k. \tag{36.431}$$

Pour la seconde, d'abord l'inégalité $Ax\leqslant|Ax|$ est évidente. Ensuite vu que $A_{ik}\geqslant 0$ nous avons

$$|Ax|_i=|\sum_k A_{ik}x_k|\leqslant\sum_k A_{ik}|x_k|=\big(A|x|\big)_i. \tag{36.432}$$

□

Soit une matrice $A\in\mathbb{M}(n,\mathbb{R})$. Nous notons

$$r_i=\sum_{j\neq i}|A_{ij}|. \tag{36.433}$$

Notons la somme sur la ligne i, pas sur la colonne : la somme est horizontale.

Définition 36.152.
Les ensembles

$$D_i=\{z\in\mathbb{C}\text{ tel que }|z-A_{ii}|\leqslant r_i\} \tag{36.434}$$

*sont les **disques de Gershgorin**. Nous allons également noter $B_i=\mathrm{Int}(D_i)$ les boules ouvertes correspondantes.*

Théorème 36.153 (Gershgorin)**.**
Soit $A\in\mathbb{M}(n,\mathbb{R})$. Si $\lambda\in\mathbb{C}$ est valeur propre de A alors $\lambda\in D_i$ pour un certain i.

Démonstration. Soit une valeur propre λ et un de ses vecteurs propres $u\in\mathbb{R}^n$: $Au=\lambda u$ avec $u\neq 0$. Soit i un indice réalisant le maximum $|u_i|=\max\{|u_k|\}_k$. Nous écrivons la i^{e}ligne de $Au=\lambda u$:

$$\sum_k A_{ik}u_k=\lambda u_i, \tag{36.435}$$

c'est à dire $A_{ii}u_i+\sum_{k\neq i}A_{ik}u_k=\lambda u_i$, ou encore

$$A_{ii}+\sum_{k\neq i}A_{ik}\frac{u_k}{u_i}=\lambda, \tag{36.436}$$

qui donne

$$|A_{ii}-\lambda|\leqslant\sum_{k\neq i}|A_{ik}|\frac{|u_k|}{|u_i|}\leqslant\sum_{k\neq i}|A_{ik}| \tag{36.437}$$

pare que $|u_i|\geqslant|u_k|$. Notons que sur la ligne précédente, $|.|$ est le module dans $\mathbb{C}$, pas la valeur absolue dans $\mathbb{R}$. □

23. Définition 2.8.

Théorème 36.154 (Gershgorin 2[410]).
Soit une matrice irréductible $A \in \mathbb{M}(n,\mathbb{R})$ et une valeur propre λ de A. Si elle est sur la frontière de l'union des disques de Gershgorin, alors elle est sur le bord de tous les disques.

Démonstration. Soit une valeur propre λ de A telle que $\lambda \in \partial\big(\bigcup_i D_i\big)$. Alors λ n'est dans l'intérieur d'aucune boule et nous avons $|\lambda - A_{ii}| \geqslant r_i$ pour tout i.

Soit un vecteur propre u de A tel que $\|u\|_\infty = 1$. Nous posons $I = \{1 \leqslant i \leqslant n \text{ tel que } |u_i| = 1\}$ et $J = \{1 \leqslant j \leqslant n \text{ tel que } |i_j| < 1\}$. Par hypothèse I n'est pas vide, et de plus $I \cap J = \varnothing$ et $I \cup J = \{1,\dots,n\}$ parce qu'aucune composante de u n'a un module[24] plus grand que 1.

La i^{e}composante de la relation $Au = \lambda u$ peut s'écrire

$$(A_{ii} - \lambda)u_i + \sum_{k\neq i} A_{ik}u_k = 0. \tag{36.438}$$

Forts de cela nous écrivons les inégalités suivantes :

$$r_i \leqslant |\lambda - A_{ii}| = \big|(\lambda - A_{ii})u_i\big| = |\sum_{k\neq i} A_{ik}u_k| \leqslant \sum_{k\neq i} |A_{ik}||u_k| \leqslant \sum_{k\neq i} |A_{ik}| = r_i. \tag{36.439}$$

Donc les inégalités sont des égalités :

$$r_i = |\lambda - A_{ii}| = \big|(\lambda - A_{ii})u_i\big| = |\sum_{k\neq i} A_{ik}u_k| = \sum_{k\neq i} |A_{ik}||u_k| = \sum_{k\neq i} |A_{ik}|. \tag{36.440}$$

En particulier l'égalité $\sum_{k\neq i} |A_{ik}||u_k| = \sum_{k\neq i} |A_{ik}|$ donne

$$\sum_{k\neq i} |A_{ik}|\big(|u_k| - 1\big) = 0. \tag{36.441}$$

Donc pour tout $k \in J$ nous avons $A_{ik} = 0$. Vu que A est irréductible, cela donnerait une partition impossible $\{1,\dots,n\} = I \cup J$. Nous en déduisons que J est vide et donc que $|u_j| = 1$ pour tout j. En repartant de (36.440) nous avons alors

$$r_i = \big|(\lambda - A_{ii})u_i\big| = |\lambda - A_{ii}||u_i| = |\lambda - A_{ii}|. \tag{36.442}$$

Cela prouve que $\lambda \in \partial D_i$ pour tout i. □

Exemple 36.155

Soit la matrice

$$B = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & -1/2 \\ -1 & 0 & 3 \end{pmatrix}. \tag{36.443}$$

D'abord nous rappelons que si vous voulez entrer cette matrice dans Sage (ou plus généralement dans Python2[25]), vous devez faire attention au 1/2 qui, tel quel, est évalué à 0. Nous vous rappelons donc que tous vos codes Sage doivent commencer par ceci :

```
1 # -*- coding: utf8 -*-
2 from __future__ import unicode_literals
3 from __future__ import division
```

tex/sage/sageSnip015.sage

24. Les composantes de u sont a priori dans $\mathbb{C}$, et non spécialement dans $\mathbb{R}$, même si A est une matrice réelle.
25. Que vous n'avez aucune raison d'utiliser autre que Sage.

Les éléments non nuls hors diagonale sont B_{13}, B_{31} et B_{23}. Elle n'est donc pas irréductible ; nous avons par exemple la partition $I = \{1, 3\}$, $J = \{2\}$ pour le critère de la proposition 36.150(2).

Les disques de Gershgorin sont

$$D_1 = \{z \in \mathbb{C} \text{ tel que } |z - 2| \leqslant 1\} \tag{36.444a}$$
$$D_2 = \{z \in \mathbb{C} \text{ tel que } |z - 1| \leqslant 1/2\} \tag{36.444b}$$
$$D_3 = \{z \in \mathbb{C} \text{ tel que } |z - 3| \leqslant 1\} \tag{36.444c}$$

Les valeurs propres de la matrice sont sur des bords de disques de Gershgorin, sans être sur tous les bords, comme ça aurait été le cas par le théorème 36.154 si la matrice avait été irréductible. Elles sont sur la figure 36.1 ; notez en particulier les valeurs propres λ_2 et λ_3 qui sont sur le bord de deux disques mais pas sur le bord des trois disques en même temps.

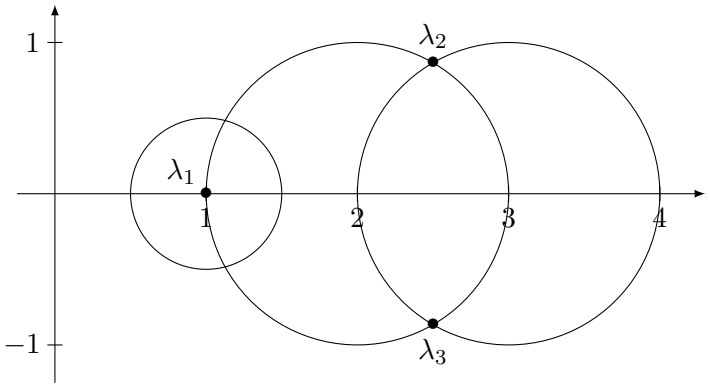


FIGURE 36.1 – Les disques de Gershgorin et les valeurs propres pour l'exemple 36.155.

△

Exemple 36.156

Soit la matrice

$$A = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 1 & 2 \\ 3 & 0 & 2 \end{pmatrix}. \tag{36.445}$$

Nous avons

$$D_1 = \{z \in \mathbb{C} \text{ tel que } |z| \leqslant 1\}, \tag{36.446a}$$
$$D_2 = \{z \in \mathbb{C} \text{ tel que } |z - 1| \leqslant 2\}, \tag{36.446b}$$
$$D_3 = \{z \in \mathbb{C} \text{ tel que } |z - 2| \leqslant 3\}, \tag{36.446c}$$

Le polynôme caractéristique est

$$\chi(\lambda) = -\lambda^3 + 3\lambda^2 - 2\lambda - 6. \tag{36.447}$$

Une fois remarqué que $\lambda_1 = -1$ est une racine, les autres sont faciles à trouver (division euclidienne de $\chi(\lambda)$ par $\lambda + 1$) : $\lambda_2 = 2 + i\sqrt{2}$ et $\lambda_3 = 2 - i\sqrt{2}$.

La matrice A est irréductible. En effet les éléments non diagonaux non nuls sont A_{12}, A_{23} et A_{31}. Ils peuvent former une chaîne reliant tous les indices entre eux.

Les contraintes sur la localisation des valeurs propres est donc qu'elles doivent être dans ou sur les disques de Gershgorin, mais que celles qui sont sur le bord d'un disque doivent être sur le bord de tous les disques en même temps. C'est cela que nous observons sur la figure 36.2. Notez en particulier la position de la valeur propre λ_1.

△

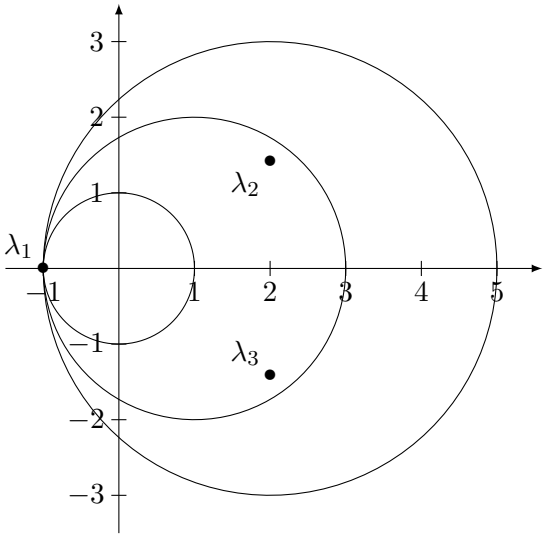


FIGURE 36.2 – Les disques de Gershgorin et les valeurs propres pour l'exemple 36.156.

36.22.1 Matrices à diagonale dominante

Définition 36.157 ([411, 410])**.**
La matrice $A \in \mathbb{M}(n, \mathbb{K})$ est ***à diagonale dominante*** *si pour tout i,*

$$\sum_{j \neq i} |A_{ij}| \leqslant |A_{ii}| \tag{36.448}$$

où $|.|$ est la module dans $\mathbb{C}$ ou la valeur absolue dans $\mathbb{R}$.

Elle est ***à diagonale fortement dominante*** *si elle est à diagonale dominante et si il existe un i tel que*

$$\sum_{j \neq i} |A_{ij}| < |A_{ii}|. \tag{36.449}$$

Elle est ***à diagonale strictement dominante*** *si*

$$\sum_{j \neq i} |A_{ij}| < |A_{ii}| \tag{36.450}$$

pour tout i (entier entre 1 et n).

Nous avons les inclusions suivantes :

$$\text{strictement dominante} \subset \text{fortement dominante} \subset \text{dominante}. \tag{36.451}$$

Lemme 36.158.
Si A est dans un des deux cas suivant :
- *diagonale strictement dominante,*
- *diagonale dominante et irréductible*[26]

alors $A_{ii} \neq 0$ pour tout i.

Démonstration. Si A est à diagonale strictement dominante, alors l'inégalité stricte (36.450) n'est pas possible.

Si A est à diagonale dominante, alors si $A_{ii} = 0$, toute la ligne est nulle. Dans ce cas, la matrice ne peut pas être irréductible. □

Proposition 36.159 ([411])**.**
Une matrice à diagonale strictement dominante est inversible.

26. Définition 36.149.

Démonstration. Soit une matrice A à diagonale strictement dominante. Soit x tel que $Ax = 0$. Le but est de montrer que $x = 0$. Soit un indice i_0 réalisant la norme maximum :

$$|x_{i_0}| = \|x\|_\infty. \tag{36.452}$$

Nous écrivons la composante i_0 de l'égalité $Ax = 0$:

$$\sum_k A_{i_0k}x_k = 0, \tag{36.453}$$

et nous séparons le terme $k = i_0$ des autres :

$$\sum_{k\neq i_0} A_{i_0k}x_K + A_{i_0i_0}x_{i_0} = 0. \tag{36.454}$$

Nous prenons le module et majorons les sommes :

$$|A_{i_0i_0}||x_{i_0}| \leqslant \sum_{k\neq i_0} |A_{i_0k}||x_k| \leqslant \sum_{k\neq i_0} |A_{i_0k}||x_{i_0}|. \tag{36.455}$$

Si $|x_{i_0}|$ est non nul nous pouvons simplifier :

$$|A_{i_0i_0}| \leqslant \sum_{k\neq i_0} |A_{i_0k}|. \tag{36.456}$$

Hélas, l'hypothèse de diagonale strictement dominante implique l'inégalité stricte dans le sens inverse. Impossible. Nous en déduisons que $|x_{i_0}| = 0$. Donc $\|x\|_\infty = 0$, ce qui signifie que $x = 0$.

Le fait que le noyau de A se réduise à $\{0\}$ implique l'inversibilité de A. □

Proposition 36.160.
Soit $A \in \mathbb{M}(n, \mathbb{C})$ une matrice qui est dans un des deux cas suivants :

- *à diagonale strictement dominante*
- *à diagonale dominante et irréductible*

Si $A = D - M$ où D est la diagonale de A (et M est « le reste ») alors D est inversible et

$$\rho(D^{-1}M) < 1 \tag{36.457}$$

où ρ est le rayon spectral (thème 39).

Démonstration. Le lemme 36.158 nous dit que les éléments diagonaux de A sont non nuls. Cela donne déjà le fait que la matrice D est inversible et que la produit $D^{-1}M$ ait un sens. Nous posons $T = D^{-1}M$. Nous avons alors

$$T_{ii} = \sum_k (D^{-1})_{ik}M_{ki}. \tag{36.458}$$

Si $k = i$ alors $M_{ki} = 0$ et si $k \neq i$ alors $D_{ik} = 0$. Donc $T_{ii} = 0$ pour tout i.

En ce qui concerne les autres éléments de T,

$$T_{ij} = \sum_k (D^{-1})_{ik}M_{kj} = \sum_k \frac{1}{A_{ik}}\delta_{ik}M_{kj} = -\frac{A_{ij}}{A_{ii}}. \tag{36.459}$$

Notes :

- Les hypothèses sur A jouent pour dire que $A_{ii} \neq 0$.
- Le signe moins est dû au fait que $M_{ij} = -A_{ij}$ lorsque $i \neq j$.

En faisant la somme des modules :

$$\sum_{j\neq i} |T_{ij}| = \sum_{j\neq i} \frac{|A_{ij}|}{|A_{ii}|} = \frac{1}{|A_{ii}|}\sum_{j\neq i} |A_{ij}| \leqslant 1. \tag{36.460}$$

La dernière inégalité est le fait que A soit à diagonale dominante.

Si A est à diagonale strictement dominante Alors nous avons l'inégalité stricte

$$\sum_{j\neq i}|T_{ij}|<1. \tag{36.461}$$

Et le théorème de Gershgorin 36.153 dit que le spectre de T est contenu dans l'union des disques

$$D_i=\{z\in\mathbb{C}\text{ tel que }|z-T_{ii}|\leqslant r_i\} \tag{36.462}$$

où

$$r_i=\sum_{j\neq i}|T_{ij}|. \tag{36.463}$$

Mais nous avons prouvé que pour tout i, $T_{ii}=0$ et $\sum_{j\neq i}|T_{ij}|<1$. Donc toutes ces boules sont contenues dans $B(0,1)$. Cela prouve que $\rho(T)<1$.

Diagonale dominante, irréductible La matrice T est alors également irréductible parce que les éléments non nuls de A et de T sont les mêmes : $T_{ij}=-A_{ij}/A_{ii}$. Nous utilisons alors le second théorème de Gershgorin 36.154. Si λ est une valeur propre de T, alors soit

$$\lambda\in\bigcup_i B\big(0,r_i\big) \tag{36.464}$$

soit

$$\lambda\in\bigcap_i\partial B(0,r_i). \tag{36.465}$$

Vu que $r_i\leqslant 1$ pour tout i, dans le premier cas λ est dans l'union des boules *ouvertes* de rayon 1. Le nombre λ est donc une la boule ouverte de rayon 1. Bref, $|\lambda|<1$.

Dans le second cas, l'intersection de deux cercles de même centre sont soit vide soit tout le cercle (auquel cas les rayons sont égaux). Dans le second cas, ledit rayon est certainement strictement plus petit que 1 parce que

$$r_i=\sum_{j\neq i}|T_{ij}|=\sum_{j\neq i}\frac{|A_{ij}|}{|A_{ii}|}<1. \tag{36.466}$$

□

36.22.2 M-matrice

Définition 36.161.
*Une matrice $A\in\mathbb{M}(n,\mathbb{R})$ est une **M-matrice** si*

(1) $A_{ii}>0$ pour tout i,

(2) $A_{ij}\leqslant 0$ si $i\neq j$

(3) A est inversible et $A^{-1}\geqslant 0$.

Proposition 36.162.
Soit $A\in\mathbb{M}(n,\mathbb{R})$ telle que $A_{ii}>0$ pour tout i et $A_{ij}\leqslant 0$ pour tout $i\neq j$. Nous posons $A=D-M$ où D est la diagonale de A.

La matrice A est une M-matrice si et seulement si $\rho(D^{-1}M)<1$.

Démonstration. En deux morceaux.

Si $\rho(D^{-1}M)<1$ Nous posons encore $T=D^{-1}M$. Par le théorème 17.118, la matrice $\mathbb{1}-T$ est inversible et

$$(\mathbb{1}-T)^{-1}=\sum_{k=0}^{\infty}T^k. \tag{36.467}$$

D'autre part, via des calculs déjà faits, et les hypothèses sur les signes des éléments de A,

$$T_{ij} = -\frac{A_{ij}}{A_{ii}} \geqslant 0. \tag{36.468}$$

Donc tous les éléments de T sont positifs (ou nuls). Par conséquent $T^k \geqslant 0$ pour tout k et $(\mathbb{1} - T)^{-1}$ est positive.
Mais $A = D - M = D(\mathbb{1} - D^{-1}M) = D(\mathbb{1} - T)$. Vu que D et $\mathbb{1} - T$ sont inversibles, nous savons que A est inversible et

$$A^{-1} = (\mathbb{1} - A)^{-1}D^{-1}, \tag{36.469}$$

qui est un produit de matrices positives. Donc $A^{-1} \geqslant 0$.

Au final, A est une M-matrice.

Si A est une M-matrice Soit une valeur propre λ de $T = D^{-1}M$ est un vecteur propre u : $Tu = \lambda u$. Vu que $T \geqslant 0$ nous avons d'une part $|\lambda u| = |\lambda|u|$ et d'autre part $|\lambda u| = |Tu| \leqslant T|u|$, ce qui donne

$$|\lambda||u| \leqslant T|u|. \tag{36.470}$$

Dans cette inégalité nous substituons T par $\mathbb{1} - (\mathbb{1} - T)$ pour avoir

$$|\mu||u| \leqslant |u| - (\mathbb{1} - T)|u| \tag{36.471}$$

ou encore

$$(\mathbb{1} - T)|u| \leqslant \big(1 - |\lambda|\big)|u|. \tag{36.472}$$

Mais $(\mathbb{1}-T)^{-1} = A^{-1}D \geqslant 0$ parce que A et D sont positives. Donc en appliquant $(\mathbb{1}-T)^{-1}$ à l'inégalité (36.472), elle est conservée (proposition 36.151(2)) :

$$|u| \leqslant (\mathbb{1} - T)^{-1}\big(1 - |\lambda|\big)|u|. \tag{36.473}$$

Si $|\lambda| \geqslant 1$ alors toutes les composantes de $\big(1 - |\mathbb{1}|\big)|u|$ sont négatives et l'inégalité n'est possible qu'avec $|u| = 0$. Dans ce cas, λ n'est pas une valeur propre (le vecteur propre soit être non nul).

Nous en déduisons que $|\lambda| < 1$ et donc que $\rho(T) = \rho(D^{-1}M) < 1$.

□

Le théorème suivant résume ce que nous avons vu en donnant une condition suffisante facile à vérifier pour être une M-matrice.

Théorème 36.163.
Soit $A \in \mathbb{M}(n, \mathbb{R})$ telle que

(1) $A_{ii} > 0$

(2) $A_{ij} \leqslant 0$ *pour* $i \neq j$

(3) vérifiant une des deux conditions suivantes :
- *à diagonale strictement dominante*
- *à diagonale dominante et irréductible.*

Alors A est une M-matrice.

Démonstration. Au vu de la proposition 36.162, il suffira de montrer que $\rho(D^{-1}M) < 1$ où D et M sont la décomposition $A = D - M$ habituelle. C'est le cas grâce à la proposition 36.160. □

Proposition 36.164.
Soit $A \in \mathbb{M}(n, \mathbb{R})$, une M-matrice irréductible. Alors $A^{-1} > 0$.

Démonstration. Nous posons $T = D^{-1}M$. En comparant la définition 36.161 de M-matrice et la caractérisation de la proposition 36.162, nous avons $\rho(D^{-1}M) < 1$. Par conséquent

$$(\mathbb{1} - T)^{-1} = \sum_{k=0}^{\infty} T^k \tag{36.474}$$

par la proposition 36.162. D'autre part, $A^{-1} = (\mathbb{1} - A)^{-1}D$ où les éléments D sont strictement positifs. Donc nous devons encore prouver que $(\mathbb{1} - T)^{-1} > 0$. Nous savons que $T \geqslant 0$, et vu que

$$\big(\sum_k T^k\big)_{ij} = \sum_k (T^k)_{ij} \tag{36.475}$$

il nous suffit de prouver que pour chaque (ij), un des $(T^k)_{ij}$ est strictement positif. Soient donc deux indices i et j. Vu que A est irréductible, ils sont connectés par une suite d'indice $i = i_0, i_1, \dots, , i_r = j$ tels que

$$T_{i_k,i_{k+1}} = -\frac{A_{i_k,i_{k+1}}}{A_{i_k,i_k}} > 0. \tag{36.476}$$

Or les indices i_k sont choisis de telle sorte que les numérateurs soient non nuls et donc strictement négatifs. Nous avons, en général :

$$(T^k)_{ij} = \sum_{l_1,\dots,l_{r-1}} T_{i,l_1}T_{l_1,l_2}\dots T_{l_{r-1},j}. \tag{36.477}$$

Chacun des termes est positif ou nul, mais pour $k = r$, il y a entre autres le terme

$$T_{i,i_1}T_{i_1,i_2}\cdots T_{i_r,j} \neq 0. \tag{36.478}$$

Donc $(T^r)_{ij} > 0$ et $\sum_{k=0}^{\infty}(T^k)_{ij} > 0$. Et par conséquent

$$A^{-1} = (\mathbb{1} - T)^{-1}D > 0. \tag{36.479}$$

□

Théorème 36.165.
Soit une M-matrice $A \in \mathbb{M}(n, \mathbb{R})$ et $g \in \mathbb{R}^n$ tel que $(Ag)_i \geqslant 1$ pour tout i. Alors $\|A^{-1}\|_\infty \leqslant \|g\|_\infty$.

Démonstration. Nous posons $u = (1, \dots, 1)$ et considérons $x \in \mathbb{R}^n$. Vu que A est une M-matrice, nous avons $A^{-1} \geqslant 0$, donc

$$|A^{-1}x| \leqslant A^{-1}|x| \leqslant \|x\|_\infty A^{-1}u \leqslant \|x\|_\infty g. \tag{36.480}$$

Justifications :

— La première inégalité est la proposition 36.151(2).

— La seconde provient de

$$\big(B|x|\big)_i = \sum_k B_{ik}|x_k| \leqslant \sum_k B_{ik}\|x\|_\infty = \|x\|_\infty \sum_k B_{ik}u_k = \|x\|_\infty Bu. \tag{36.481}$$

— Étant donné que $A^{-1} \geqslant 0$ nous conservons l'inégalité et $Ag \geqslant u$ implique $g \geqslant A^{-1}u$ (c'est la proposition 36.151(1)).

En ce qui concerne la norme de A^{-1} nous avons donc

$$\|A^{-1}\|_\infty = \sup_{|x|_\infty=1} \|A^{-1}x\|_\infty \leqslant \sup_{\|x\|_\infty=1} \|x\|_\infty\|g\|_\infty = \|g\|_\infty. \tag{36.482}$$

□

Proposition 36.166.
Une matrice de $\mathbb{M}(n, \mathbb{R})$ qui

(1) est symétrique,

(2) Vérifie une des deux conditions suivantes
- *est irréductible à diagonale fortement dominante*
- *est à diagonale strictement dominante,*

(3) vérifie $A_{ii} > 0$ pour tout i

est strictement définie positive.

Démonstration. D'après le théorème de Gershgorin 36.153, chaque valeur propre de A est dans un des disques fermés

$$D_i = \{z \in \mathbb{C} \text{ tel que } |z - A_{ii}| \leqslant r_i\}. \tag{36.483}$$

Par hypothèse, les centres de ces disques sont réels et strictement positifs. Mais le fait que A soit à diagonale dominante donne que le rayon de ces cercles sont plus petits que A_{ii}. Donc D_i n'intersecte pas $]-\infty, 0[$. Mais le fait que A soit symétrique implique que les valeurs propres soient réelles (théorème 11.174(1)). Cela montre que les valeurs propres de A sont toutes dans $[0, \infty[$.

Si la matrice A est à diagonale strictement dominante, alors les inégalités sont strictes et le théorème est prouvé.

Sinon nous somme dans le cas irréductible à diagonale fortement dominante et nous avons le théorème de Gershgorin numéro 2 36.154. Soit une valeur propre λ. Soit elle est dans un des disques ouvert (qui est inclus dans $]0, \infty[$), soit elle est dans l'intersection des bords des disques. Mais au moins un des disques n'intersecte pas 0 (parce que la diagonale est strictement dominante). Dans ce cas non plus λ ne peut pas être nul.

Nous en déduisons que dans tous les cas, les valeurs propres sont toutes réelles strictement positives. □

Chapitre 37

Méthode des différences finies

37.1 Problèmes de dimension un

Soit $u\colon \mathbb{R} \to \mathbb{R}$ une fonction et $h > 0$. Nous définissons les opérations suivantes (qui sont supposées approximer la dérivée $u'(x)$ lorsqu'elle existe).

Définition 37.1.
La ***différence progressive*** *est*

$$(D_h^+ u)(x) = \frac{u(x+h) - u(x)}{h}, \tag{37.1}$$

la ***différence régressive*** *est*

$$(D_h^- u)(x) = \frac{u(x) - u(x-h)}{h}, \tag{37.2}$$

la ***différence centrée*** *est*

$$(D_h^0 u)(x) = \frac{u(x+h) - u(x-h)}{2h}. \tag{37.3}$$

Nous ne noterons pas toujours la dépendance en h, c'est à dire que nous noterons D^+u au lieu de $D_h^+ u$ lorsque cela ne pose pas de problèmes.

Notons que u'' peut être approximé par D^+D^+u, D^0D^+, D^+D^-, et encore de nombreuses autres possibilités.

Voici un lemme qui dit que tout cela n'est pas si mal, pourvu que u soit assez régulière.

Lemme 37.2.
Soit un ouvert connexe Ω de $\mathbb{R}$, soit $x \in \Omega$ et $h > 0$ tel que $\overline{B(x,h)} \subset \Omega$.

(1) Si $u \in C^2(\Omega)$ alors

$$|u'(x) - D^+u(x)| \leqslant \frac{h}{2}\|u''\|_{\bar{\Omega}} \tag{37.4}$$

et

$$|u'(x) - D^-u(x)| \leqslant \frac{h}{2}\|u''\|_{\bar{\Omega}}. \tag{37.5}$$

(2) Si $u \in C^3(\bar{\Omega})$ alors

$$|u'(x) - D^0(x)| \leqslant \frac{h^2}{2}\|u^{(3)}\|_{\bar{\Omega}} \tag{37.6}$$

(3) Si $u \in C^4(\bar{\Omega})$ alors

$$|u''(x) - D^-D^+u(x)| \leqslant \frac{h^2}{12}\|u^{(4)}\|_{\bar{\Omega}}. \tag{37.7}$$

Démonstration. Nous prouvons le point (3). D'abord nous regardons de quoi nous avons besoin :

$$D^-D^+u(x) = \frac{(D^+u)(x) - (D^+u)(x-h)}{h} = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} \tag{37.8}$$

Nous allons y mettre les approximations de $u(x+h)$ et $u(x-h)$ par Taylor, proposition 14.288 :

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(x+\theta_1 h) \tag{37.9}$$

avec $\theta_1 \in [0,1]$. De même,

$$u(x-h) = u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(x-\theta_2 h) \tag{37.10}$$

avec $\theta_2 \in [0,1]$.

Donc

$$u(x+h) + u(x-h) - 2u(x) = h^2u''(x) + \frac{h^4}{4!}\Big(u^{(4)}(x+\theta_1 h) + u^{(4)}(x-\theta_2 h)\Big), \tag{37.11}$$

ce qui donne

$$(D^-D^+u)(x) = u''(x) + \frac{h^2}{4!}\Big(u^{(4)}(x+\theta_1 h) + u^{(4)}(x-\theta_2 h)\Big). \tag{37.12}$$

Chacun des deux termes dans la parenthèse peut être majoré par $\|u^{(4)}\|_{\bar{\Omega}}$. Notons que c'est une majoration très sauvage parce que $x+\theta_1 h$ ne prend ses valeurs que dans $[x, x+h]$ avec h supposé petit. Quoi qu'il en soit nous ne pouvons pas dire mieux que

$$|u''(x) - D^-D^+u(x)| \leqslant \frac{h^2}{12}\|u^{(4)}\|_{\bar{\Omega}}. \tag{37.13}$$

□

Remarque 37.3 ([1]).
Lorsque nous écrivons

$$|u'(x) - D^+u(x)| \leqslant \delta \tag{37.14}$$

pour tout x, nous ne pouvons pas écrire

$$\|u' - D^+u\|_\infty \leqslant \delta \tag{37.15}$$

parce que l'inégalité (37.14) n'est valable que pour les x tels que $[x-h, x+h] \subset \Omega$, de telle sorte que l'inégalité n'est pas spécialement correcte sur $\bar{\Omega}$.

37.1.1 Un schéma à cinq points

37.1.1.1 Poser le système

Soit $\Omega =]0,1[$ et l'équation différentielle

$$\begin{cases} -u''(x) + c(x)u(x) = f \quad \text{sur } \Omega \\ u(0) = \alpha \\ u(1) = \beta \end{cases} \tag{37.16}$$

où c est une fonction positive et $\alpha, \beta \in \mathbb{R}$. Nous considérons $h > 0$ assez petit pour que le reste ait un sens. Si nous cherchons des solutions dans $C^4(\bar{\Omega})$, le lemme 37.2 nous dit que

$$|u''(x) - D^-D^+u(x)| = \eta(h^2) \tag{37.17}$$

où η est une fonction telle que $\lim_{t\to 0}\eta(t) = 0$. Nous pouvons récrire l'équation différentielle sous la forme

$$-D^-D^+u(x) + c(x)u(x) = f(x) + \eta(h^2). \tag{37.18}$$

Si nous négligeons le terme $\eta(h^2)$ qui est supposé être petit nous pouvons tenter de résoudre pour la fonction u_h

$$-D^-D^+u_h(x) + c(x)u_h(x) = f(x). \tag{37.19}$$

Notons ici l'importance de la notion de problème bien posé parce que en remplaçant le paramètre (fonctionnel) f par $f+\eta(h^2)$, nous modifions les solutions. Dans la mesure où le problème est bien posé, cette petite modification ne modifiera pas trop la solution et nous pouvons espérer que $\|u-u_h\|$ soit petit pour une norme ou une autre.

Utilisant l'expression (37.8) pour D^-D^+ nous avons l'équation suivante pour u_h :

$$\frac{1}{h^2}\Big(2u_h(x)-u_h(x+h)-u_h(x-h)\Big)+c(x)u_h(x)=f(x). \tag{37.20}$$

Avons-nous gagné quelque chose ? Pas encore. L'idée de la discrétisation est de ne considérer u_h qu'en certains points, et de prendre ces points à intervalles réguliers de taille h. Soient donc N un nombre entier et $h=1/(N+1)$. Nous posons

$$x_k=kh \tag{37.21}$$

pour $i=0,\dots,N+1$. Avec cela nous avons

$$\overline{\Omega}=\bigcup_{k=0}^{N-1}[x_k,x_{k+1}] \tag{37.22a}$$

$$x_0=0 \tag{37.22b}$$

$$x_{N+1}=1. \tag{37.22c}$$

Nous posons surtout

$$\Omega_h=\{x_i\}_{i=1,\dots,N} \tag{37.23}$$

et

$$\bar{\Omega}_h=\{x_i\}_{i=0,\dots,N+1}. \tag{37.24}$$

Enfin, nous ne considérons plus u_h que comme une fonction $u_h\colon\bar{\Omega}_h\to\mathbb{R}$. C'est à dire que u_h est un vecteur à $N+2$ composantes.

L'équation (37.20) devient

$$\frac{1}{h^2}\big(2u_h(x_i)-u_h(x_{i+1})-u_h(x_{i-1})+c(x_i)u_h(x_i)\big)=f(x_i) \tag{37.25}$$

pour $i=1,\dots,N$. Sur les bords, cette équation n'est pas possible parce que x_{i-1} ou x_{i+1} n'existerait pas. Au contraire, sur les bords nous avons les conditions aux bords

$$u_h(x_0)=\alpha \tag{37.26}$$

et

$$u_h(x_{N+1})=\beta. \tag{37.27}$$

En posant $c_i=c(x_i)$ et $u_i=u_h(x_i)$, les inconnues du problème sont les nombres u_i ($i=1,\dots,N+1$). Elles sont immédiatement résolues pour u_0 et u_{N+1}. Pour les autres, il faut écrire et résoudre un système d'équation linéaire.

L'écriture du système linéaire à résoudre consiste essentiellement à écrire (37.25) en séparant les cas $i=0$ et $i=N$ parce que nous savons déjà les valeurs de u_0 et u_N. Le système que nous avons est :

$$\begin{cases}\left(\frac{2}{h^2}+c_1\right)u_1-\frac{1}{h^2}u_2=f_1+\frac{\alpha}{h^2} & i=1 \quad (37.28a)\\ \left(\frac{2}{h^2}+c_N\right)u_N-\frac{1}{h^2}u_{N-1}=f_N+\frac{\beta}{h^2} & i=N \quad (37.28b)\\ \left(\frac{2}{h^2}+c_i\right)u_i-\frac{1}{h^2}u_{i+1}-\frac{1}{h^2}u_i=f_i. & \text{autres} \quad (37.28c)\end{cases}$$

Cela se met sous la forme matricielle

$$L_hU_h=F_h \tag{37.29}$$

pour

$$F_h = \big(f_1 + \frac{\alpha}{h^2}, f_2, \dots, f_{N-1}, f_N + \frac{\beta}{h^2}\big) \tag{37.30}$$

et les éléments non nuls de L_h sont :

$$(L_h)_{i,i-1} = -\frac{1}{h^2} \qquad \text{pour } i = 2, \dots, N \tag{37.31a}$$

$$(L_h)_{i,i+1} = -\frac{1}{h^2} \qquad \text{pour } i = 1, \dots, N-1 \tag{37.31b}$$

$$(L_h)_{i,i} = \frac{2}{h^2} + c_i \qquad \text{pour } i = 1, \dots, N. \tag{37.31c}$$

Cette matrice est pleine de zéros, à part les trois diagonales centrales, et il existe des méthodes efficaces pour résoudre le système d'équation correspondant.

37.1.1.2 Propriétés du système

La matrice est la suivante :

$$L_h = \begin{pmatrix} \frac{2}{h^2} + c_1 & -1/h^2 & 0 & 0 & \cdots & 0 \\ -1/h^2 & \frac{2}{h^2} + c_2 & -1/h^2 & 0 & \cdots & 0 \\ 0 & -1/h^2 & \frac{2}{h^2} + c_3 & -1/h^2 & \ddots & \vdots \\ 0 & 0 & -1/h^2 & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & -1/h^2 \\ 0 & 0 & \cdots & 0 & -1/h^2 & \frac{2}{h^2} + c_N \end{pmatrix} \tag{37.32}$$

où nous avions déjà posé l'hypothèse $c_i \geqslant 0$ pour tout i.

Lemme 37.4.
La matrice L_h est irréductible[1] *à diagonale fortement dominante*[2].

Démonstration. Nous décomposons la preuve en plusieurs parties.

La première ligne Sur la première ligne, seuls deux éléments sont non nuls et nous avons

$$L_{11} = \frac{2}{h^2} + c_1 \geqslant \frac{2}{h^2} > \frac{1}{h^2} = L_{12}. \tag{37.33}$$

La dernière ligne Elle est semblable à la première.

Les autres lignes Sur les autres lignes nous avons trois éléments non nuls et

$$\sum_{j \neq i} |A_{ij}| = \frac{2}{h^2} \leqslant \frac{2}{h^2} + c_i = L_{ii}. \tag{37.34}$$

Diagonale fortement dominante Nous avons prouvé jusqu'à présent que L_h était une matrice à diagonale fortement dominante.

Irréductible Nous allons utiliser la caractérisation de la proposition 36.150(3). Pour cela nous considérons le chaîne d'éléments non nuls

$$A_{12}, A_{23}, \dots, A_{N-1,N} = -\frac{1}{h^2}. \tag{37.35}$$

Soient deux indices i et j avec $i < j$. Cette suite d'indice (ou un sous-suite) rend i et j connectés.

Si par contre $i > j$, il faut considérer la suite inversée grâce au fait que L_h est symétrique :

$$A_{N,N-1}, A_{N-1,N-2}, \dots, A_{32}, A_{21} = -\frac{1}{h^2}. \tag{37.36}$$

1. Caractérisation 36.150.
2. Définition 36.157.

□

Proposition 37.5.
Soit le problème

$$\begin{cases} -u''(x) + c(x)u(x) = f \quad \text{sur } \Omega \\ u(0) = \alpha \\ u(1) = \beta \end{cases} \tag{37.37}$$

où c est une fonction positive et $\alpha, \beta \in \mathbb{R}$. Nous considérons $h > 0$ assez petit pour que le reste ait un sens. Et nous approximons u'' par D^-D^+u.

La matrice L_h des différences finies associée à ce problème est

(1) une M-matrice,

(2) strictement définie positive,

(3) d'inverse $L_h^{-1} > 0$.

Démonstration. Le théorème 36.163 dit que L_h est une M-matrice. La Proposition 36.166 nous donne aussi que L_h est strictement définie positive.

Le lemme 37.4 dit que L_h est irréductible, ce qui permet à la proposition 36.164 de conclure que $L_h^{-1} > 0$. □

Cela étant rappelé, nous pouvons continuer.

Lemme 37.6.
Soit $\Omega =]0,1[$, soit $N \in \mathbb{N}$ et $h = 1/(N+1)$. La solution $w_h \colon \Omega_h \to \mathbb{R}$ du problème discrétisé

$$\begin{cases} -(D^-D^+w_h)(x_i) = 1 & \text{(37.38a)} \\ w_h(0) = 0 & \text{(37.38b)} \\ w_h(1) = 0 & \text{(37.38c)} \end{cases}$$

pour tout $x_i = ih$ $(i = 1, \dots, N)$ donne les valeurs exactes des $w(x_i)$ lorsque w est la solution de

$$\begin{cases} -w''(x) = 1 & \text{(37.39a)} \\ w(0) = 0 & \text{(37.39b)} \\ w(1) = 0. & \text{(37.39c)} \end{cases}$$

Démonstration. Un enseignement de la proposition 37.5 est que le système (37.38) peut être écrit sous la forme d'un système linéaire $L_h^0 w_h = F_h$ où L_h^0 est inversible. Il y a donc unicité de la solution.

D'autre part, la solution du système (37.39) est $w(x) = \frac{1}{2}(x - x^2)$, qui est de classe C^∞. Le lemme 37.2(3) dit que $D^-D^+w = w''$. Donc les valeurs $w(x_i)$ résolvent aussi le système (37.38). □

Lemme 37.7 (Quelques estimations)**.**
La matrice L_h du problème sus-mentionné en (37.37) vérifie[3] *:*

(1) $\|L_h\|_\infty \leqslant \frac{4}{h^2} + \|c\|_\infty$

(2) $\|L_h^{-1}\|_\infty \leqslant \frac{1}{8}$.

Démonstration. Nous nous souvenons de la formule (12.20) :

$$\|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |A_{ij}|. \tag{37.40}$$

La première ligne a pour somme : $\frac{3}{h^2} + c_1$, la dernière a pour somme $\frac{3}{h^2} + c_n$ et les autres sont pour somme $\frac{4}{h^2} + c_i$. Elles sont donc toutes majorées par $\frac{4}{h^2} + \|c\|_\infty$.

3. Dans le CTES d'analyse numérique de Marseille, l'estimation donnée est $\|L_h^{-1}\|_\infty \leqslant \frac{1}{4}$.

Pour l'estimation de $\|L_h^{-1}\|_\infty$ nous allons nous appuyer sur le théorème 36.165.

Commençons par considérer le problème

$$\begin{cases} -w'' = 1 & (37.41a)\\ w(0) = w(1) = 0. & (37.41b) \end{cases}$$

La première équation dit que w est un polynôme de degré 2. En écrivant $w(x) = ax^2 + bx + c$ et en imposante toutes les contraintes, nous trouvons l'unique solution

$$w(x) = -\frac{1}{2}(x^2 - x). \tag{37.42}$$

Le lemme 37.6 nous dit que la fonction w prise aux points $x_i = ih$ donne les valeurs de w_h.

La matrice L_h^0 est une M-matrice et le vecteur w_h vérifie $L_h^0 w_h = \mathbb{1}$. Donc le théorème 36.165 s'applique et

$$\|(L_h^0)^{-1}\| \leqslant \|w_h\|_\infty = \frac{1}{8}. \tag{37.43}$$

L'obtention de 1/8 n'est rien d'autre que la recherche du maximum (en valeur absolue) de la parabole $x \mapsto (x - x^2)/2$ pour $x \in [0, 1]$. Le maximum est atteint pour $x = 1/2$; calcul de dérivée et tout ça ...

Nous retournons maintenant à notre matrice originale L_h. Nous avons

$$L_h - L_h^0 = \operatorname{diag}(c_1, \dots, c_n) \geqslant 0, \tag{37.44}$$

et aussi

$$L_h^{-1} - (L_h^0)^{-1} = \underbrace{L_h^{-1}}_{\geqslant 0} \underbrace{(L_h^0 - L_h)}_{\leqslant 0} \underbrace{(L_h^0)^{-1}}_{\geqslant 0} \tag{37.45}$$

parce que L_h est une M-matrice. Donc tous les coefficients de $L_h^{-1} - (L_h^0)^{-1}$ sont négatifs. Cela implique

$$L_h^{-1} \leqslant (L_h^0)^{-1}. \tag{37.46}$$

Mais nous savons que les coefficients de L_h^{-1} sont positifs, donc le maximum de ses coefficients en valeur absolue est plus petit que ceux de $(L_h^0)^{-1}$, c'est à dire

$$\|L_h^{-1}\|_\infty \leqslant \|(L_h^0)^{-1}\|_\infty \leqslant \frac{1}{8}. \tag{37.47}$$

□

37.1.2 Exemple

Soit $\Omega =]0, 1[$ et une fonction $u\colon \bar{\Omega} \to \mathbb{R}$ de classe C^4 vérifiant

$$\begin{cases} -u''(x) + u(x) = \sin(x) & (37.48a)\\ u(0) = 0 & (37.48b)\\ u(1) = 0. & (37.48c) \end{cases}$$

Nous allons écrire la méthode des différences finies pour $h = 1/4$. Nous posons donc les points

$$\begin{cases} x_0 = 0 & (37.49a)\\ x_1 = 1/4 & (37.49b)\\ x_2 = 1/2 & (37.49c)\\ x_3 = 3/4 & (37.49d)\\ x_4 = 1. & (37.49e) \end{cases}$$

Vu que nous avons supposé u de classe C^4, le lemme 37.2(3) nous donne[4]

$$u''(x) = (D^-D^+u)(x) + \alpha(h) \tag{37.50}$$

avec $\lim_{h\to 0} \alpha(h)/h = 0$. L'équation discrétisée serait alors

$$\begin{cases} -(D^-D^+u)(x) + u(x) = \sin(x) & \text{(37.51a)} \\ u(0) = u(1) = 0. & \text{(37.51b)} \end{cases}$$

où nous n'avons pas précisé l'indice h au bas des opérateurs D^+ et D^-. Les équations (37.51) ne doivent être posées que pour x_1, x_2 et x_3 parce que les valeurs en x_0 et x_4 sont déjà connues.

Pour x_1

$$\frac{u_2 - 2u_1 + u_0}{h^2} + u_1 = \sin(x_1) \tag{37.52}$$

Pour x_2

$$\frac{u_3 - 2u_2 + u_1}{h^2} + u_2 = \sin(x_2) \tag{37.53}$$

Pour x_3

$$\frac{u_4 - 2u_3 + u_2}{h^2} + u_3 = \sin(x_3). \tag{37.54}$$

Nous tenons compte du fait que $u_0 = u_4 = 0$ et que $h = 1/4$ pour écrire le système

$$\begin{pmatrix} -31 & 16 & 0 \\ 16 & -31 & 16 \\ 0 & 16 & -31 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} \tag{37.55}$$

A où les s_i sont des nombres parfaitement connus : par exemple $s_1 = \sin(x_1) = \sin(1/4) \simeq 0.247403959254523$.

37.2 Problèmes de dimension deux

Nous allons considérer le système

$$\begin{cases} -\Delta u = f & \text{sur } \Omega \\ u = g & \text{sur } \partial\Omega \end{cases} \tag{37.56}$$

où $\Omega =]0, a[\times]0, b[$.

Remarque 37.8.
Pourquoi un signe moins devant le laplacien ? Pour avoir la proposition 37.5 qui dira que la matrice correspondant aux différences finies appliquées à ce système est une M-matrice. Sinon, c'est la matrice $-L_h$ qui en serait une.

37.2.1 Discrétisation en croix

Nous allons maintenant déduire une discrétisation du laplacien en discrétisant les opérations ∂_x^2 et ∂_y^2. Nous discrétisons Ω en mailles carrés de côté h : $x_k = kkh$ et $y_k = kh$. L'opération de dérivée partielle ∂_x est discrétisée par

$$(D_x^+u)(x,y) = \frac{u(x+h) - u(x,y)}{h} \tag{37.57}$$

ou

$$(D_x^-u)(x,y) = \frac{u(x,y) - u(x-h,y)}{h} \tag{37.58}$$

4. Nous ferions n'importe quoi pour ne pas écrire $u''(x) = (D^-D^+u)(x) + o(h^2)$. Notez que vous faites ce que vous voulez : écrivez avec la notation « petit o » si cela vous chante.

ou

$$(D_x^0 u)(x,y) = \frac{u(x+h,y) - u(x-h,y)}{2h} \tag{37.59}$$

où le h est sous-entendu dans les opérateurs D^0, D^+ et D^-.

La dérivée partielle seconde $\partial_x^2 u$ peut être approximée par toutes les combinaisons imaginable, par exemple

$$(D_x^- D_x^+ u)(x,y) = \frac{u(x+h,y) - 2u(x,y) + u(x-h,y)}{h^2}. \tag{37.60}$$

Pour évaluer la différence entre $(\partial_x^2 u)(x,y)$ et $(D^-D^+u)(x,y)$, il est possible de faire du Taylor en deux dimensions, mais nous pouvons également recycler ce qui a été fait. Nous posons $u_y(x) = u(x,y)$ et alors $(\partial_x^2 u)(x,y) = u_y''(x)$ et le lemme 37.2(3) donne, si u_y est de classe C^4,

$$|u_y''(x) - D^-D^+u_y(x)| \leqslant \frac{1}{12}h^2 \|u_y^{(4)}\|_\infty. \tag{37.61}$$

Là, les opérateurs D^+ et D^- sont ceux à une dimension. Mais nous avons $(D^-D^+u_y)(x) = (D^-D^+u)(x,y)$ (à droite ce sont les opérateurs à deux dimensions), donc

$$\big|(\partial_x^2 u)(x,y) - (D^-D^+u)(x,y)\big| \leqslant \frac{1}{12}h^2 \|\partial_x^4 u\|_\infty \tag{37.62}$$

et nous pouvons écrire

$$(\partial_x^2 u)(x,y) = (D^-D^+u)(x,y) + h^2 R(x,y,h) \tag{37.63}$$

où R est une fonction qui dépend de x, y et h, mais aussi de u. Le point important est que R soit majoré par une quantité indépendante de h, de telle sorte que nous ayons quelques garanties que négliger ce terme soit une bonne approximation lorsque $h \to 0$.

Au niveau de la discrétisation, nous considérons x_i avec $i = 0, \dots, N_x$ et y_j avec $j = 0, \dots, N_y$. La discrétisation de $-(\Delta u)(x,y) = f(x,y)$ donne, pour $i = 1, \dots, N_x - 1$ et $j = 1, \dots, N_y - 1$,

$$\frac{1}{h^2}(-u_{i+1,j} + 4u_{ij} - u_{i-1,j} - u_{i,j+1} + u_{i,j-1}) = f_{ij}. \tag{37.64}$$

Les équations avec i ou j valant 0 ou N_x, N_y sont les valeurs au bords.

37.9.
Nous notons pour référence ultérieure la discrétisation suivante du laplacien :

$$(\Delta_h u)(x_i, y_j) = \frac{1}{h^2}\big(-4u_{i,j} + u_{i+1,j} + u_{i-1,j} + u_{i,j-1} + u_{i,j+1}\big). \tag{37.65}$$

Elle vérifie

$$\Delta_h f = \Delta f + h^2 \alpha(h). \tag{37.66}$$

Cette discrétisation est dite « en croix » parce que les points exploités forment une croix.

37.2.2 Discrétisation en carré

L'opérateur laplacien est défini par $\Delta = \partial_x^2 + \partial_y^2$, mais il existe de nombreuses autres façons de l'écrire.

Lemme 37.10.
Le laplacien est invariant par changement de coordonnées orthogonales. Plus précisément, si A est une matrice orthogonale, en posant $u_i = \sum_k A_{ik} e_k$ nous avons

$$\sum_i \frac{\partial^2 f}{\partial u_i^2} f = \Delta f. \tag{37.67}$$

Démonstration. Nous avons :

$$\frac{\partial f}{\partial u_i} = \sum_k A_{ik} \frac{\partial f}{\partial x_k}, \tag{37.68}$$

et donc

$$\sum_i \frac{\partial^2 f}{\partial u_i^2} = \sum_i \frac{\partial}{\partial u_i} \left(\sum_k A_{ik} \frac{\partial f}{\partial x_k} \right) = \sum_{ijk} (A^t)_{ji} A_{ik} \partial^2 f = \sum_{jk} (A^t A)_{jk} \partial_{jk}^2 f. \tag{37.69}$$

En particulier si A est une matrice orthogonale, $(A^t A)_{jk} = \delta_{jk}$ et le résultat est prouvé. □

Les plus convaincus diront que $\Delta = \nabla \cdot \nabla$ et que le produit scalaire est invariant sous changement de coordonnées orthogonales.

Nous avons déjà déduit la discrétisation (37.65) du laplacien :

$$(\Delta_h u)(x_i, y_j) = \frac{1}{h^2} \big(-4u_{i,j} + u_{i+1,j} + u_{i-1,j} + u_{i,j-1} + u_{i,j+1} \big). \tag{37.70}$$

Nous allons maintenant en déduire une par l'idée de décomposer le laplacien dans la base $u = e_1 + e_2$, $v = e_1 - e_2$. Pour cela nous introduisons les opérations (le nombre h dont dépendent ces opérateurs est sous-entendu)

$$(D_u^+ f)(x, y) = \frac{f(x + h, y + h) - f(x, y)}{h} \tag{37.71a}$$

$$(D_v^+ f)(x, y) = \frac{f(x + h, y - h) - f(x, y)}{h} \tag{37.71b}$$

$$(D_u^- f)(x, y) = \frac{f(x, y) - f(x - h, y - h)}{h} \tag{37.71c}$$

$$(D_v^- f)(x, y) = \frac{f(x, y) - f(x - h, y + h)}{h}. \tag{37.71d}$$

Vu que

$$\partial_u^2 + \partial_v^2 = 2\Delta, \tag{37.72}$$

nous discrétisons le laplacien par

$$\Delta_h' = \frac{1}{2} \big(D_u^- D_u^+ + D_v^- D_v^+ \big). \tag{37.73}$$

Un peu de calcul donne :

$$(\Delta_h' f)(x, y) = \frac{1}{2h^2} \Big(-4f(x, y) + f(x + h, y + h) + f(x - h, y + h) \tag{37.74a}$$

$$+ f(x + h, y - h) + f(x - h, y - h) \Big). \tag{37.74b}$$

37.2.3 Résolution de la discrétisation en croix

Les équations (37.64) forment un système d'équations linéaires à résoudre. Certaines peuvent être simplifiées parce qu'elles « touchent » le bord. Nous verrons cela un peu plus tard.

Nous allons d'abord numéroter correctement les équations de façon à ne pas avoir deux mais un seul indice. Notre fonction de numérotation sera

$$\varphi(i, j) = (j - 1)(N_x - 1) + i \tag{37.75}$$

avec $i = 1, \dots, N_x - 1$ et $j = 1, \dots, N_y - 1$. Cela correspond à numéroter les points de l'intérieur du quadrillage ligne par ligne en bas en haut et de gauche à droite. Avec cela les équations (37.64) vont être numérotées par un seul indice I allant de $\varphi(1, 1) = 1$ à $\varphi(N_x - 1, N_y - 1) = (N_x - 1)(N_y - 1)$.

Si $I = \varphi(i,j)$ alors nous avons vite

$$\varphi(i+1,j) = I + 1 \tag{37.76a}$$
$$\varphi(i,j+1) = I + N_x - 1 \tag{37.76b}$$
$$\varphi(i-1,j) = I - 1 \tag{37.76c}$$
$$\varphi(i,j-1) = I - N_x + 1. \tag{37.76d}$$

Nous posons $U_I = u_{\varphi^{-1}(I)}$, et l'équation (37.64) devient

$$\frac{1}{h^2}(-U_{I+1} + 4U_I - U_{I-1} - U_{I+N_x-1} - U_{I-N_x+1}) = f_I. \tag{37.77}$$

Pour savoir la matrice représentant ce système, nous devons simplifier les équations qui doivent l'être. Par exemple avec $I = 1$, le terme $U_{I-1} = U_0$ vaut $u_{0,1} = f_01$. Ce n'est donc pas réellement une inconnue de notre problème.

Nous voulons mettre les équations sous la forme du système

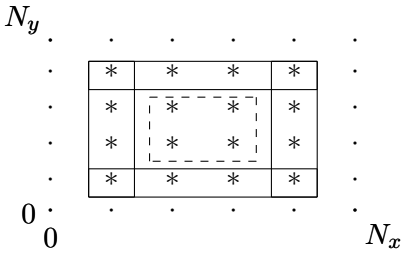
$$L_h U = F. \tag{37.78}$$

Sur la ligne numéro I de L_h, les éléments non nuls sont :

$$L_{I,I} = 4 \tag{37.79a}$$
$$L_{I,I+1} = -1 \tag{37.79b}$$
$$L_{I,I-1} = -1 \tag{37.79c}$$
$$L_{I,I+N_x-1} = -1 \tag{37.79d}$$
$$L_{I,I-N_x+1} = -1 \tag{37.79e}$$

pour peu qu'ils existent. Par exemple pour $I = 1$, il n'y a pas d'éléments $L_{I,I-1}$. Les indices I et J de $L_{I,J}$ vont de 1 à $\varphi(N_x - 1, N_y - 1) = (N_y - 1)(N_x - 1)$.

Voici un dessin de notre situation :



À chaque élément du quadrillage correspond une équation.

— Aux points simples sur le bord, correspondent des équations triviales parce que la fonction u y est directement donnée par les conditions aux bords.

— Aux points étoilés entourés en traits continus correspondent des équations « incomplètes » parce que certains termes de l'équation (37.64) sont donnés par les conditions aux bords. Elle correspondent aussi aux lignes incomplète de la matrice L_h où certains éléments donnés en (37.79) n'existent pas.

Le membre de droite de ces équations est par contre enrichi de ce qui à gauche est « donné ».

— Au points étoilés du centre entourés en traits discontinus correspondent des équations complètes.

Notons que f_{00} ne joue aucun rôle dans notre histoire parce que dans les équations (37.64), chaque point (i,j) du maillage n'est liée qu'aux quatre points situés « à côté ».

Proposition 37.11.
La matrice L_h est

(1) *irréductible et à diagonale fortement dominante*[5],

(2) *une M-matrice,*

(3) *inversible avec* $L_h > 0$,

(4) *symétrique,*

(5) *strictement définie positive.*

Démonstration. On divise la preuve.

Irréductible Une matrice $n \times n$ dont les deux premières diagonales sont entièrement composées d'éléments non nuls est toujours irréductible. En effet, la première lie l'élément $(1, 2)$ à l'élément $(n-1, n)$ et donc permet de dire que tous les $i < j$ sont connectés.

La seconde diagonale lie l'élément $(n, n-1)$ à l'élément $(2, 1)$.

Diagonale fortement dominante En ce qui concerne la dominance de la diagonale, il faut sommer sur les lignes. Or chaque ligne contient (en valeur absolue) un 4 sur la diagonale et au plus quatre éléments qui valent 1. D'où

$$|L_{II}| \geqslant \sum_{J \neq I} |L_{IJ}|. \tag{37.80}$$

La première ligne n'est jamais complète : elle contient un 4 sur l'élément $(1, 1)$ au au maximum deux 1 plus à droite. Donc la matrice L_h est à diagonale fortement dominante.

M-matrice D'après ce que nous venons de voir (proposition 37.11), le théorème 36.163 fonctionne et L_h est une M-matrice[6].

Inverse strictement positif La proposition 36.164 nous assure qu'une M-matrice irréductible est d'inverse strictement positif. Donc $L_h^{-1} > 0$.

Symétrique La ligne numéro I est

$$\big(\dots, \underbrace{-1}_{I-N_x+1}, \dots, -1, 4, -1, \dots, \underbrace{-1}_{I+N_x-1}, \dots\big) \tag{37.81}$$

Prenons par exemple l'élément $(I, I - N_x + 1)$ qui vaut -1. Son symétrique est l'élément $(I - N_x + 1, I)$ qui se trouve sur la ligne $I - N_x + 1$. Sur cette dernière ligne nous avons un -1 sur la colonne $I - N_x + 1 + N_x - 1 = I$. Donc l'élément $(I - N_x + 1, I)$ vaut bien -1 et la matrice est symétrique.

Strictement définie positive Vu que la matrice L_h est symétrique, irréductible à diagonale fortement dominante (proposition 37.11), vu que ses éléments diagonaux sont strictement positifs (ils valent 4), la proposition 36.166 nous dit que L_h est strictement définie positive.

□

37.3 Consistance, convergence

37.3.1 Définitions, mise en place

Soit un ouvert $\Omega \subset \mathbb{R}^n$ et un opérateur différentiel L sur Ω. Nous considérons le problème qui consiste à trouver une fonction u sur Ω telle que

$$Lu = f \tag{37.82}$$

pour une fonction f donnée.

Problèmes et choses à faire

La définition suivante est une invention personnelle, n'est pas précise et mérite des commentaires de la part du lecteur.

5. Définition 36.157. Le cas 1×1 est discutablement à diagonale fortement dominante, il faut avouer.

6. Notons que c'est ici que nous sommes content d'avoir posé $-\Delta u = f$ dans le système (37.56), avec un signe négatif devant le laplacien. Sinon tous le signes auraient changé et la matrice $-L_h$ aurait été une M-matrice au lieu de L_h.

Définition 37.12 ([1]).
*Un **schéma numérique** de pas h pour $Lu = f$ est la donnée de*

(1) un nombre $h > 0$ supposé petit,

(2) une quantité N de points x_i dans Ω formant l'ensemble discret Ω_h,

(3) une matrice L_h de taille $N \times N$,

(4) une solution $u_h : \Omega_h \to \mathbb{R}$ de l'équation $(L_h u_h)(x_i) = f_i$ où nous avons posé $f_i = f(x_i)$.

37.13.
Évidemment pour qu'un schéma mérite le nom de schéma de pas h pour l'équation $Lu = f$, il faut que le nombre h soit lié au choix des points x_i, et que la matrice L_h soit liée à l'opérateur L. La définition n'impose pas formellement de tels liens, parce qu'il y a de nombreuses façons d'approximer une équation différentielle en un système linéaire, sans compter que même l'équation $(L_h u_h)_i = f_i$ peut se résoudre de beaucoup de façons, exacte ou approchées.

Cela pour dire que le lien entre la solution exacte u et la solution approchée n'a rien d'évident, et va dépendre des choix faits lors de la discrétisation et lors de la résolution du système linéaire. Nous allons supposer dans un premier temps que l'équation $L_h u_h = f$ est résolue exactement (nous avons un peu parlé de ces problème dans les sections 36.15 et suivantes).

Définition 37.14.
*L'erreur **de consistance** d'un schéma numérique est la fonction $\tau_h : \Omega_h \to \mathbb{R}$ définie par*

$$\tau_h(x_i) = (L_h u)_i - (Lu)(x_i). \tag{37.83}$$

Il y a un jeu de notation pas tout à fait évident dans la définition (37.83). En effet, L_h est une matrice, et ne s'applique donc a priori pas immédiatement à une fonction. Ce que signifie la notation $(L_h u)_i$ est que l'on applique la matrice L_h au vecteur $j \mapsto u(x_j)$ et que l'on prend la composante i du résultat.

Définition 37.15.
*Nous disons que le schéma est **consistant** avec l'opérateur différentiel L lorsque*

$$\lim_{h \to 0^+} \|\tau_h\| = 0 \tag{37.84}$$

où la norme $\|.\|$ est souvent la norme uniforme, c'est à dire $\|\tau_h\| = \max_i \tau_h(x_i)$.

Notons que le lien entre h et le choix des x_i fait partie de la définition des schéma. Sur un segment de longueur L, lorsque h n'est pas un diviseur de L, le schéma devrait expliquer ce que l'on fait pour que la limite (37.84) ait un sens.

Définition 37.16.
*Le schéma (Ω_h, L_h) est **consistant à l'ordre** p avec l'opérateur différentiel L pour la norme $\|.\|$ si il existe une constante C indépendante de h telle que*

$$\|\tau_h\| \leqslant C h^p. \tag{37.85}$$

Définition 37.17.
*L'**erreur de discrétisation** entre la solution u du problème $Lu = f$ et la solution approchée u_h sur Ω_h est la fonction*

$$\begin{aligned} e_h : \Omega_h &\to \mathbb{R} \\ x_i &\mapsto u(x_i) - u_i. \end{aligned} \tag{37.86}$$

où $u_i = u_h(x_i)$ est la solution approchée.

*Le schéma discret $(L_h u_h)(x_i) = f_i$ est **convergeant** si $\lim_{h \to 0} \|e_h\| = 0$. Si de plus is existe une constante C et $p > 0$ tels que*

$$\|e_h\| \leqslant C h^p, \tag{37.87}$$

*alors nous disons que le schéma est convergence à l'**ordre** p.*

Si l'erreur de consistance est petite, le *problème* est bien approximé par la système linéaire. Cela n'implique cependant pas que la solution trouvée soit bien approximée.

Exemple 37.18(Deux opérateurs différentiels proches dont les solutions sont loin)
Soit la partie $\Omega =]0, \infty[$, et les problèmes

$$\begin{cases} L_1 u = u' = 0 & (37.88a) \\ u(0) = 1 & (37.88b) \end{cases}$$

et

$$\begin{cases} L_2 v = v' - \epsilon v = 0 & (37.89a) \\ v(0) = 1. & (37.89b) \end{cases}$$

Les solutions exactes sont $u(x) = 1$ et $v(x) = e^{\epsilon x}$.

En ce qui concerne les opérateurs, quelle que soit la norme utilisée nous avons

$$\begin{aligned} \|L_1 - L_2\| &= \sup_{\|f\|=1} \|L_1(f) - L_2(f)\| & (37.90a) \\ &= \sup_{\|f\|=1} \|\epsilon f\| & (37.90b) \\ &= \epsilon. & (37.90c) \end{aligned}$$

Donc lorsque ϵ est petit, l'opérateur L_2 approxime bien l'opérateur L_1. Pour toutes les normes. Mais

$$\big|u(x) - v(x)\big| = |1 - e^{\epsilon x}|, \tag{37.91}$$

donc quel que soit ϵ nous avons $\|u - v\|_\infty = \infty$. Et d'ailleurs, quelle que soit la norme raisonnable que nous mettons sur l'espace des fonctions, avoir $\|u - v\| = \infty$ semble inévitable.

Donc deux opérateurs différentiels proches peuvent avoir des solutions lointaines. △

37.3.2 Exemple

Soit l'opérateur différentiel L donné par

$$Lu = -u'' + cu \tag{37.92}$$

où c est une fonction. Nous considérons sur $\Omega =]0, 1[$ l'équation différentielle

$$Lu = 0. \tag{37.93}$$

En ce qui concerne la discrétisation, nous définissons le maillage $\Omega_h = \{x_i = ih\}$ avec $i = 0, \dots, N+1$. La solution approchée discrètement sera le vecteur v qui peut être vu comme fonction $v\colon \Omega_h \to \mathbb{R}$. Les nombres v_0 et v_{N+1} sont a priori donnés par les conditions aux bords. Pour les autres v_i nous avons les équations

$$(L_h v)_i = -\frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} + c(x_i)v_i. \tag{37.94}$$

Cela est la définition de l'opérateur L_h, et le vecteur v solution de $L_h v = 0$ est la solution du problème au sens de la méthode des différences finies (pour peu qu'il existe, soit unique et tout ça).

Pour calculer l'erreur de consistence, nous considérons une fonction u et nous posons $u_i = u(x_i)$. Le vecteur (u_i) ainsi construit est approximé par v (on espère). Nous avons :

$$\tau_h(x_i) = -\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - c(x_i)u_i - (Lu)(x_i). \tag{37.95}$$

Pour étudier cela nous développons $u_{i+1} = u(x_i + h)$ et $u_{i-1} = u(x_i - h)$ à l'ordre 4 : il existe $\alpha_i \in [x_i, x_i + h]$ et $\beta_i \in [x_i - h, x_i]$ tels que

$$u_{i+1} = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{3!}u^{(3)}(x_i) + \frac{h^4}{4!}u^{(4)}(\alpha_i) \tag{37.96}$$

et

$$u_{i-1} = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{3!}u^{(3)}(x_i) + \frac{h^4}{4!}u^{(4)}(\beta_i). \tag{37.97}$$

Après simplification de plusieurs termes,

$$\tau_h(x_i) = -\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - c_i u_i + u''(x_i) + c_i u_i = \frac{h^2}{4!}\big(u^{(4)}(\alpha_i) + u^{(4)}(\beta_i)\big). \tag{37.98}$$

Parler de la consistance du schéma demande d'étudier $\lim_{h\to 0^+} \|\tau_h\|$, et pour cela, il faut préciser la norme avec laquelle nous voulons travailler. L'ordre de consistance va dépendre de la norme utilisée.

Pour la norme $\|.\|_\infty$, les nombres $u^{(4)}(\alpha_i)$ et $u^{(4)}(\beta_i)$ se majorent par $\|u^{(4)}\|_\infty$ et nous avons

$$\|\tau_h\|_\infty \leqslant \frac{h^2}{12}\|u^{(4)}\|_\infty. \tag{37.99}$$

Nous avons consistance d'ordre 2.

Remarque 37.19.
La valeur de $\|\tau_h\|_\infty$ dépend de la fonction u sur laquelle nous la calculons. Cependant nous avons convergence $\|\tau_h\|_\infty \to 0$ pour toute fonction (de classe disons C^4).

La constante C pour laquelle nous avons $\|\tau_h\| \leqslant Ch^2$ et donc qui nous vaut de pouvoir dire que la consistance est d'ordre 2 ne dépend pas de h, ni des valeurs ponctuelles de u ou de ses dérivées, mais dépend des normes de u et de ses dérivées (en l'occurrence seulement de la norme de $u^{(4)}$.)

Étudions la consistance pour la norme L_1 :

$$\|\tau_h\|_1 = \sum_i |\tau_h(x_i)| \leqslant \frac{h^2}{12}\sum_i \|u^{(4)}\|_\infty \tag{37.100}$$

où nous avons majoré chacun des $u^{(4)}(\alpha_i)$ par $\|u^{(4)}\|_\infty$. Combien de termes dans la somme ? Nous avons $h = 1/(N-1)$ et donc $N = (1+h)/h$, ce qui donne

$$\|\tau_h\|_1 \leqslant N\frac{h^2}{12}\|u^{(4)}\|_\infty = (1+h)Ch. \tag{37.101}$$

La constante $1+h$ se majore par n'importe quelle constante strictement plus grande que 1. Nous pouvons donc la rentrer dans C et écrire

$$\|\tau_h\|_1 \leqslant Ch \tag{37.102}$$

et donc avoir la consistance à l'ordre 1.

37.3.3 Consistance, stabilité et convergence

Lemme 37.20.
Soit un opérateur différentiel L, soit u la solution de $Lu = f$ et un schéma numérique (L_h, Ω_h) pour cette équation. Nous notons u_h la solution de $L_h u_h = f$. Alors nous avons

$$L_h e_h = \tau_h \tag{37.103}$$

Et si de plus L_h est inversible,

$$\|e_h\| \leqslant \|L_h^{-1}\|\|\tau_h\|. \tag{37.104}$$

Démonstration. Par définition u_h est solution de $L_h u_h = f$ en tant que fonctions sur Ω_h. Nous avons donc

$$L_h e_h = L_h u_h - L_h u \tag{37.105}$$

où u doit être compris comme la restriction de u à Ω. En appliquant au point x_i,

$$(L_he_h)(x_i) = \underbrace{(L_hu_h)(x_i)}_{=f_i} - (L_hu)(x_i), \tag{37.106}$$

mais $f_i = (Lu)(x_i)$ parce que u est solution de $Lu = f$. Donc

$$(L_he_h)(x_i) = (Lu)(x_i) - (L_hu)(x_i) = \tau_h(x_i). \tag{37.107}$$

Si la matrice L_h est inversible nous avons $e_h = L_h^{-1}\tau_h$ et donc

$$\|e_h\| \leqslant \|L_h^{-1}\|\|\tau_h\| \tag{37.108}$$

par le lemme 12.15. □

Bien entendu, en tant qu'opérateur linéaire sur un espace de dimension finie, l'opérateur L_h^{-1} est borné pour chaque h. Mais si il n'y a pas une borne uniforme en h, alors le lemme 37.20 dit qu'il n'y a pas d'espoir de majorer $\|e_h\|$ de façon à passer à la limite $\lim_{h\to 0}\|L_h^{-1}\|$.

Définition 37.21.
Un schéma numérique est ***stable*** *si il existe une constante $C > 0$ indépendante de h telle que $\|L_h^{-1}\| \leqslant C$.*

Théorème 37.22.
En deux parties.

(1) Si un schéma discret est consistant et stable, alors il est convergeant.

(2) Si de plus il est consistant à l'ordre p, alors il est convergeant à l'ordre p.

Démonstration. Nous savons du lemme 37.20 (qui s'applique parce que l'inversibilité de L_h est dans la définition de la stabilité) que $\|e_h\| \leqslant \|L_h^{-1}\|\|\tau_h\|$ et que $\|L_h^{-1}\| \leqslant C$. En passant à la limite[7],

$$\lim_{h\to 0}\|e_h\| \leqslant C\lim_{h\to 0}\|\tau_h\| = 0. \tag{37.109}$$

La dernière limite est le fait que le schéma soit consistant. Le schéma est donc convergeant.

Si de plus il est consistant à l'ordre p, alors

$$\|e_h\| \leqslant C\|\tau_h\| \leqslant C'h^p \tag{37.110}$$

Il est donc également convergeant à l'ordre p. □

37.3.4 Exemple : schéma à cinq points, laplacien en croix

Nous avions développé le schéma dont l'opérateur sur Ω_h est (voir (37.64))

$$(L_hu_h)(x_i, y_j) = \frac{1}{h^2}\big(-u_{i+1,j} - u_{i,j+1} + 4u_{i,j} - u_{i-1,j} - u_{i,j-1}\big). \tag{37.111}$$

Proposition 37.23.
Le schéma est :

(1) consistant à l'ordre 2,

(2) stable pour la norme uniforme et

$$\|L_h^{-1}\|_\infty \leqslant \frac{1}{8}, \tag{37.112}$$

(3) convergeant à l'ordre 2 pour la norme $\|.\|_\infty$.

7. Toutes les limites $h \to 0$ sont en réalité des limites $h \to 0^+$, mais nous allégeons cette notation.

Démonstration. Cet opérateur avait été construit de telle sorte à avoir (voir (37.63))

$$(\Delta u)(x_i,y_j) = (L_h u)(x_i,y_i) + h^2 R(x,y,h) \tag{37.113}$$

où R peut être majoré indépendamment de h. En tant que fonctions sur Ω_h nous avons

$$\tau_h = \Delta u - L_h u = h^2 R(x,y,h), \tag{37.114}$$

et donc $\|\tau_h\|_\infty \leqslant Ch^2$, parce que le lemme 37.2(3) donne aussi

$$\|R\|_\infty \leqslant C \max\{\|\frac{\partial^4 u}{\partial x^4}\|_\infty, \|\frac{\partial^4 u}{\partial y^4}\|_\infty\}. \tag{37.115}$$

En ce qui concerne la stabilité nous allons utiliser le théorème 36.165. Nous considérons la fonction

$$g(x,y) = -\frac{1}{4}(x^2+y^2), \tag{37.116}$$

qui vérifie $-\Delta g = 1$ sur le carré $[0,1]^2$. Nous considérons le vecteur g_h d'indices $(i,j) \mapsto g_{ij} = g(x_i,y_j)$ sur lequel nous calculons L_h :

$$(L_h g_h)_{ij} = \frac{1}{h^2}(-g_{i+1,j} - g_{i,j+1} + 4g_{ij} - g_{i-1,j} - g_{i,j-1}); \tag{37.117}$$

en remplaçant les g par leurs valeurs en termes de x_i, x_{i-1}, x_{i+1}, y_j, y_{j-1} et y_{j+1}, et ne tenant compte du fait que $x_k = kh$ et $y_l = lh$ nous avons :

$$(L_h g_h)_{ij} = \frac{1}{4}\big((i+1)^2 - j^2 + i^2 + (j+1)^2 - 4i^2 - 4j^2 + (i-1)^2 + j^2 + i^2 + (j-1)^2\big) = 1. \tag{37.118}$$

Donc $L_h g_h = 1$.

Vu que L_h est une M-matrice (proposition 37.11(2)), le théorème 36.165 nous dit alors que L_h^{-1} vérifie

$$\|L_h^{-1}\|_\infty \leqslant \|g_h\|_\infty. \tag{37.119}$$

Mais

$$\|g_h\|_\infty \leqslant \|g\|_\infty = g\big(\frac{1}{2},\frac{1}{2}\big) = \frac{1}{8}. \tag{37.120}$$

Notons que cela est bien une inégalité et non une égalité parce que rien n'assure que le point $(1/2,1/2)$ soit sur le maillage ; donc rien n'assure que la valeur $g(1/2,1/2)$ ne soit parmi les valeurs du vecteur discrétisé g_h.

Notre schéma numérique est stable et consistant à l'ordre 2 pour la norme $\|.\|_\infty$. Le théorème 37.22 dit alors que le schéma est convergeant à l'ordre 2 pour la même norme. □

37.4 Autres laplaciens

Nous avons vu le laplacien en croix (37.65)

$$(\Delta_h f)(x,y) = \frac{1}{h^2}\big(-4f(x,y) + f(x+h,y) + f(x-h,y) \tag{37.121a}$$

$$+ f(x,y+h) + f(x,y-h)\big) \tag{37.121b}$$

qui vérifie

$$\Delta_h f = \Delta f + Kh^2, \tag{37.122}$$

ainsi que le laplacien en carré (37.74)

$$(\Delta'_h f)(x,y) = \frac{1}{2h^2}\Big(-4f(x,y) + f(x+h,y+h) + f(x-h,y+h) \tag{37.123a}$$

$$+ f(x+h,y-h) + f(x-h,y-h)\Big) \tag{37.123b}$$

qui vérifie également $\Delta'_h f = \Delta f + Kh^2$.

A priori toutes combinaisons de la forme

$$a\Delta_h + b\Delta'_h \tag{37.124}$$

avec $a + b = 1$ est valable comme tentative de discrétiser le laplacien. Ce sont des schémas à 9 points. Évidemment la matrice L_h correspondante va être moins creuse, mais nous pouvons espérer ajuster a et b de telle sorte à obtenir une consistance d'un ordre supérieur à 2.

Nous allons développer les $\Delta_h f$ et $\Delta'_h f$ à l'ordre 4 (reste à l'ordre 6). Quelques remarques avant de commencer.

(1) Allez relire la proposition 14.268 et les notations qui vont avec pour comprendre les différentielles.

(2) Écrivez les formules du type (14.693) pour $d^2 f$ et $d^4 f$.

(3) Allez relire le développement de Taylor du théorème 14.281.

(4) À l'ordre zéro, il n'y a rien parce que le terme $-4f(x, y)$ compense les quatre termes d'ordre zéro des autres termes.

(5) Aux ordres impairs, il n'y a rien. En effet, prenons un nombre impair l et la formule

$$(d^l f)_x(h, \dots, h) = \sum_{i_1,\dots,i_l} h_{i_1} \dots h_{i_l} \frac{\partial^l f}{\partial x_{i_1} \dots \partial x_{i_l}}(x). \tag{37.125}$$

Nous avons

$$(d^l f)_x(h, \dots, h) + (d^l f)_x(-h, \dots, -h) = 0. \tag{37.126}$$

Or dans les expressions (37.123) et (37.121), les termes arrivent par pair opposées.

Commençons par calculer $h^2(\Delta_h f)(x, y)$.

Ordre 4 . Le premier terme est :

$$(d^2 f)_{(x,y)}\big((h, 0), (h, 0)\big) = h^2 (d^2 f)_{(x,y)}\big((1, 0), (1, 0)\big). \tag{37.127}$$

La formule (14.693) à peine adaptée permet de calculer ça explicitement.

Il y a encore les termes du même type avec $(0, 1)$, $(-1, 0)$ et $(0, -1)$.

Ordre 4 Cette fois, ce sont 4 termes du type

$$h^4 (d^4 f)_{(x,y)}\big((1, 0), (1, 0), (1, 0), (1, 0)\big) \tag{37.128}$$

à calculer.

Cela fait beaucoup de termes à calculer. Je vois laisse vous persuader que le programme suivant en Sage nous donne les coefficients.

```
#! /usr/bin/env python3
# -*- coding: utf8 -*-


x,y=var("x,y")

C=[]
C.append(x)
C.append(y)

def coef4(h):
    S=0
    for i in [0,1]:
```

```
        for j in [0,1]:
            for k in [0,1]:
                for l in [0,1]:
                    S=S+h[i]*h[j]*h[k]*h[l]*C[i]*C[j]*C[k]*C[l]
    return S

def coef2(h):
    S=0
    for i in [0,1]:
        for j in [0,1]:
            S=S+h[i]*h[j]*C[i]*C[j]
    return S

cross=[]
square=[]

cross.append(  [1,0] )
cross.append(  [-1,0] )
cross.append(  [0,1] )
cross.append(  [0,-1] )

square=[]
square.append(  [-1,-1] )
square.append(  [1,-1] )
square.append(  [-1,1] )
square.append(  [1,1] )


print("Cross scheme :")

K=sum(  coef2(v) for v in cross   )
L=sum(  coef4(v) for v in cross   )
print(K)
print(L)

print("square scheme :")

K=sum(  coef2(v) for v in square   )
L=sum(  coef4(v) for v in square   )
print(K)
print(L)
```

tex/sage/coefs.sage

Le résultat est que, en utilisant la formule

$$\partial_x^4 f + 2\partial_{xxyy}^4 f + \partial_y^4 f = \Delta\Delta f, \tag{37.129}$$

nous avons

$$(\Delta_h f)(x,y) = \frac{1}{2}(2\partial_x^2 f + 2\partial_y^2 f)(x,y) + \frac{1}{4!}2h^2(\partial_x^4 f + \partial_y^4 f)(x,y) + Kh^4 \tag{37.130a}$$
$$= (\Delta f)(x,y) + \frac{1}{12}h^2(\partial_x^4 f + \partial_y^4 f)(x,y) + Kh^4 \tag{37.130b}$$
$$= \Delta f + \frac{h^2}{12}\Delta\Delta f - \frac{h^2}{12}2\partial_{xxyy}^4 f + Kh^4 \tag{37.130c}$$
$$= \Delta f + \frac{h^2}{12}\Delta\Delta f - \frac{h^2}{6}\partial_{xxyy}^4 f + Kh^4 \tag{37.130d}$$

où K est une constante qui peut être majorée en terme des dérivées quatrièmes de f. En particulier la plus grande des normes supremum de ces dérivées.

Le même genre de calculs donnent

$$(\Delta'_h f)(x,y) = \frac{1}{2}\Big[\frac{1}{2}4\Delta f + \frac{h^2}{4!}(4\partial_x^4 f + 24\partial_x^2\partial_y^2 f + 4\partial_y^2 f)\Big] + Kh^4. \tag{37.131}$$

Ça donne :

$$(\Delta'_h f) = \Delta f + \frac{h^2}{12}\Delta\Delta f + \frac{h^2}{3}\partial_{xxyy}^4 f + Kh^4 \tag{37.132}$$

avec redéfinition du K ; nous ne le préciserons plus à chaque fois.

Nous avons donc le résultat proposé dans [412] :

$$a\Delta_h f + b\Delta'_h f = (a+b)\Delta f + (a+b)\frac{h^2}{12}\Delta^2 f + h^2\frac{1}{6}(a-2b)\partial_{xxyy}^4 f + Kh^4. \tag{37.133}$$

L'idée est d'appliquer ça à une fonction u qui vérifie l'équation différentielle $-\Delta u = f$ (attention au clash de notation pour f). Le mieux est de supprimer le terme en $\partial_{xxyy} f$ en demandant $a-2b=0$. Nous avons donc à résoudre le système

$$\begin{cases} a+b=1 & \text{(37.134a)}\\ a-2b=0. & \text{(37.134b)} \end{cases}$$

Qui propose une décomposition PLU pour résoudre ce système linéaire ? Quelle que soit la manière, la solution est

$$a = \frac{2}{3}, \quad b = \frac{1}{3}. \tag{37.135}$$

Nous allons donc étudier la discrétisation à neuf points

$$L_h = \frac{2}{3}\Delta_h + \frac{1}{3}\Delta'_h. \tag{37.136}$$

En faisant quelques additions nous trouvons que l'opération

$$(L_h u)(x_i,y_j) = \frac{1}{6h^2}\Big(-20u_{ij} + 4\big(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}\big) + u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1}\Big) \tag{37.137}$$

vérifie

$$L_h f = \Delta f + \frac{h^2}{12}\Delta^2 f + Kh^4. \tag{37.138}$$

37.4.1 Travail avec le laplacien à 9 points

Nous allons écrire un schéma numérique pour l'équation différentielle $-\Delta u = f$ utilisant la discrétisation à 9 points du laplacien. Nous recopions ses propriétés fondamentales (37.136), (37.137), (37.138) :

$$L_h = \frac{2}{3}\Delta_h + \frac{1}{3}\Delta'_h, \tag{37.139}$$

et

$$(L_h u)(x_i, y_j) = \frac{1}{6h^2}\Big(-20u_{ij} + 4\big(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}\big) + u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1}\Big), \tag{37.140}$$

et

$$L_h f = \Delta f + \frac{h^2}{12}\Delta^2 f + Kh^4. \tag{37.141}$$

Nous appliquons (37.141) à u et nous isolons Δu :

$$\Delta u = L_h u - \frac{h^2}{12}\Delta^2 u + Kh^4 = \frac{1}{6h^2}T_h u + \frac{h^2}{12}\Delta f + Kh^4 \tag{37.142}$$

où nous avons utilisé $\Delta^2 u = -\Delta f$ et avons noté

$$T_h u = -20u_{ij} + 4(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}) + u_{i+1,j+1} + u_{i-1,j+1} + u_{i+1,j-1} + u_{i-1,j-1}. \tag{37.143}$$

Nous imposons maintenant $\Delta u = -f$ en écrivant

$$\frac{1}{6h^2}T_h u = -f - \frac{h^2}{12}\Delta f + \alpha(h)h^4. \tag{37.144}$$

Une idée est de remplacer Δf par son approximation en croix (37.122) :

$$T_h u = -6h^2 f - \frac{h^4}{2}(\Delta_h f + Kh^2) + \alpha(h)h^6 \tag{37.145}$$

Avec quelques calculs nous trouvons le schéma numérique suivant :

$$20u_{ij} - 4(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} - u_{i,j-1}) - u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} - u_{i-1,j-1} \tag{37.146a}$$

$$= \frac{h^2}{2}(8f_{ij} + f_{i+1,j} + f_{i-1,j} + f_{i,j+1} + f_{i,j-1}) + \alpha(h)h^6. \tag{37.146b}$$

En oubliant le terme en $\alpha(h)h^6$, nous obtenons un système d'équations linéaires.

Problèmes et choses à faire

Il me semble que ce schéma donne une convergence d'ordre 6. C'est correct ?

Chapitre 38

Variables aléatoires et théorie des probabilités

38.1 Espace de probabilité

Définition 38.1.
*Une **mesure de probabilité** sur un espace mesurable*[1] $(\Omega, \mathcal{A})$ *est une mesure positive telle que* $P(\Omega) = 1$. *Dans ce cas, le triple* $(\Omega, \mathcal{A}, P)$ *est un **espace de probabilité**.*

Un point $\omega \in \Omega$ est une **observation**, une partie mesurable $A \in \mathcal{A}$ est un **événement**. L'ensemble $A \cup B$ représente l'événement A ou B tandis que l'ensemble $A \cap B$ représente l'événement A et B.

Si les A_n sont des événements, nous avons défini en 14.309 limite supérieure et la limite inférieure de la suite A_n par

$$\limsup_{n\to\infty} A_n = \bigcap_{n\geqslant 1} \bigcup_{k\geqslant n} A_k \tag{38.1}$$

et

$$\liminf_{n\to\infty} A_n = \bigcup_{n\geqslant 1} \bigcap_{k\geqslant n} A_k \tag{38.2}$$

Si $\omega \in \liminf A_n$, alors ω réalise tous les A_n sauf un nombre fini.

Nous avons

$$\limsup A_n = \{\omega \in \Omega \text{ tel que } \omega \in A_n \text{pour une infinité de } n\}. \tag{38.3}$$

Théorème 38.2 (Borel-Cantelli)**.**
Si

$$\sum_{n=1}^{\infty} P(A_n) < \infty \tag{38.4}$$

alors $P(\limsup A_n) = 0$.

Démonstration. La condition $\sum_{n\geqslant 1} P(A_n) < \infty$ signifie que la fonction

$$\varphi = \sum_{n\geqslant 1} \mathbb{1}_{A_n} \tag{38.5}$$

est P-intégrable. Par conséquent, elle est finie presque partout (au sens de P), c'est à dire

$$P(\varphi = \infty) = 0. \tag{38.6}$$

Les points ω sur lesquels $\varphi(\omega) = \infty$ sont ceux tels que

$$\sum_{n\geqslant 1} \mathbb{1}_{A_n}(\omega) = \infty, \tag{38.7}$$

1. Espace mesurable : 16.1, mesure positive : 16.23.

c'est à dire les ω qui appartiennent à une infinité d'ensembles A_n, ou encore les $\omega \in \limsup A_n$. Nous avons donc montré que

$$\{\omega \text{ tel que } \varphi(\omega) = \infty\} = \{\omega \in \Omega \text{ tel que } \omega \in A_n \text{pour une infinité de } n\} = \limsup A_n. \tag{38.8}$$

Or l'hypothèse signifie que la probabilité du membre de gauche est nulle. □

Corollaire 38.3.
Si $\sum_{n=1}^{\infty} P(\complement A_n) < \infty$, alors presque sûrement tous les B_n sont réalisés à l'exception d'un nombre fini.

38.2 Variables aléatoires

Définition 38.4.
*Une **variable aléatoire** est une application mesurable*

$$X : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}or(\mathbb{R}^d)). \tag{38.9}$$

Nous convenons que $\mathbb{R}^1 = \bar{\mathbb{R}}$, c'est à dire que dans le cas où la variable aléatoire X est réelle, nous acceptons les valeurs $\pm\infty$.

Définition 38.5.
*Une variable aléatoire réelle X est **absolument continue** s'il existe une fonction positive et intégrable $f : \mathbb{R} \to \mathbb{R}$ telle que pour tout intervalle $I \subset \mathbb{R}$,*

$$P(X \in I) = \int_I f(t)dt. \tag{38.10}$$

*Nous disons alors que f est la **densité** de X.*

Cela ne devrait pas être sans rappeler la définition 19.29.

38.2.1 Indépendance

La définition suivante vient de l'instructive motivation de [413]. La définition d'indépendance de deux événements se généralise à n événements de la façon suivante.

Définition 38.6.
*Nous disons que les événements $A_1, \dots, A_n$ sont **indépendants** si pour tout choix $\{i_1, \dots, i_k\} \subset \{1, \dots, n\}$ nous avons*

$$P(A_{i_1} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \dots P(A_{i_k}). \tag{38.11}$$

*Les sous tribus $\mathcal{A}_1, \dots, \mathcal{A}_n$ sont **indépendantes** si pour tout choix $A_i \in \mathcal{A}_i$, les événements A_i sont indépendants.*

Exemple 38.7
Soit $\Omega = [0,1] \times [0,1]$ muni de la mesure de Lebesgue. Soient $A = [0,a] \times [0,1]$ et $B = [0,1] \times [0,b]$. Nous avons $P(A) = a$ et $P(B) = b$ ainsi que $P(A \cup B) = ab$. △

Lemme 38.8.
Les tribus $\mathcal{A}_1, \dots, \mathcal{A}_n$ sont indépendantes si et seulement si

$$P(A_1 \cap \dots \cap A_n) = P(A_1) \dots P(A_n) \tag{38.12}$$

pour tout $A_i \in \mathcal{A}_i$.

Démonstration. L'implication dans le sens direct découle immédiatement des définitions.

Nous supposons avoir un choix $(A_i)_{i=1,\dots,n}$ avec $A_i \in \mathcal{A}_i$ et nous devons montrer que ces événements sont indépendants, c'est à dire que si $J \subset \{1,\dots,n\}$ alors les événements $(A_j)_{j\in J}$ sont indépendants. Sans perte de généralité, nous pouvons supposer que si $i \notin J$, $A_i = \Omega$. Alors nous avons

$$P\big(\bigcap_{j\in J} A_j\big) = P\big(\bigcap_{i=1}^{n} A_i\big) = \prod_{i=1}^{n} P(A_i) = \prod_{j\in J} P(A_j) \tag{38.13}$$

parce que $P(A_i) = P(\Omega) = 1$ lorsque i n'est pas dans J. □

Si A est un événement, la **tribu engendrée** par A est

$$\sigma(A) = \{\varnothing, A, \complement A, \Omega\}. \tag{38.14}$$

Soit $X\colon \Omega \to \mathbb{R}^d$ une variable aléatoire. Conformément à la définition 16.68, la **tribu engendrée** est

$$\mathcal{A}_X = \{X^{-1}(B) \text{ tel que } B \in \mathcal{B}\text{or}(\mathbb{R}^d)\}. \tag{38.15}$$

Cela est la plus petite tribu sous tribu de $\mathcal{A}$ pour laquelle X est mesurable. Elle sera aussi (le plus) souvent notée $\sigma(X)$.

Définition 38.9.
*Nous disons que les variables aléatoires $X_k\colon \Omega \to \mathbb{R}^d$ sont **indépendantes** lorsque les tribus engendrées $\mathcal{A}_{X_1},\dots,\mathcal{A}_{X_n}$ le sont.*

Remarque 38.10.
Il n'a de sens de dire que X_1 et X_2 sont indépendants que si X_1 et X_2 sont des applications dont l'espace de départ est identique.

Si nous voulons modéliser le jet de deux pièce indépendantes, le mauvais choix est de faire $\Omega = \{0,1\}$, y mettre la mesure d'équiprobabilité, et de considérer les deux variables aléatoires

$$X_i(\omega) = \begin{cases} f & \text{si } \omega = 0 \\ p & \text{si } \omega = 1. \end{cases} \tag{38.16}$$

Ces deux variables sont évidemment pas indépendantes. Il faut poser $\Omega = \{0,1\}\times\{0,1\}$, y mettre la mesure d'équiprobabilité et poser

$$X_1(x,y) = \begin{cases} f & \text{si } x = 0 \\ p & \text{si } x = 1 \end{cases}, \tag{38.17}$$

$$X_2(x,y) = \begin{cases} f & \text{si } y = 0 \\ p & \text{si } y = 1 \end{cases}, \tag{38.18}$$

Ces variables aléatoires sont indépendantes. Par exemple

$$X_1^{-1}\{p\} = \{(1,0),(1,1)\} \tag{38.19a}$$
$$X_2^{-1}\{p\} = \{(0,1),(1,1)\} \tag{38.19b}$$

et on a bien

$$P\big(X_1^{-1}\{p\} \cap X_2^{-1}\{p\}\big) = P\{(1,1)\} = \frac{1}{4} \tag{38.20}$$

ainsi que

$$P\{X_p^{-1}(p)\} = \frac{1}{2} \tag{38.21a}$$

pour $i = 1$ et $i = 2$.

Proposition 38.11.
Soient $(X_k : \Omega \to \mathbb{R}^{d_k})$ des variables aléatoires indépendantes.

(1) Si $B_k \in \mathcal{B}or(\mathbb{R}^{d_k})$. Alors

$$P(X_k \in B_k \forall k \leqslant n) = P(X_1 \in B_1) \dots P(X_n \in B_n). \tag{38.22}$$

(2) Les événements $\{X_i \in B_i\}$ sont indépendants.

(3) Les tribus engendrées par des X_i et d'autres sont indépendantes. Plus précisément, si I et J sont deux ensembles disjoints de $\mathbb{N}$ alors les tribus

$$\sigma(\{X_i, i \in I\}) \tag{38.23}$$

et

$$\sigma(\{X_i, i \in J\}) \tag{38.24}$$

sont indépendantes.

Démonstration. Lorsque nous écrivons $X_i \in B_i$, nous parlons de l'événement

$$(X_i \in B_i) = \{\omega \in \Omega \text{ tel que } X_i(\omega) \in B_i\} = X_i^{-1}(B_i) \in \mathcal{A}_{X_i}. \tag{38.25}$$

Vu que par hypothèse les tribus $(\mathcal{A}_i)$ sont indépendantes, le lemme 38.8 nous montre que

$$P\Big(\bigcap_{i=1}^{n} X_i \in B_i\Big) = \prod_i P(X_i \in B_i). \tag{38.26}$$

Il reste à voir que l'ensemble $X_i^{-1}(B_i)$ fait partie de la tribu $\mathcal{A}$ de départ. Cela est la définition du fait que l'application X_i soit une variable aléatoire : elle doit être mesurable en tant qu'application

$$X_i : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}\text{or}(\mathbb{R}^d)). \tag{38.27}$$

Les affirmations (2) et (3) ne sont que des façons alternatives d'exprimer la même chose. □

Lemme 38.12.
Les événements $(A_i)_{i=0,\dots,n}$ sont indépendants si et seulement si les événements que nous obtenons en remplaçant certains des A_i par $\complement A_i$ le sont.

Démonstration. Sans perte de généralité, nous pouvons nous contenter de prouver que les événements $\complement A_0, A_1, \dots, A_n$ sont indépendants sous l'hypothèse que les événements $A_0, A_1, \dots, A_n$ sont indépendants. Soit I un sous-ensemble de $\{1, \dots, n\}$. Nous avons

$$\begin{aligned}
P\big(\complement A_0 \bigcap_{i\in I} A_i\big) &= P\Big(\bigcap_{i\in I} A_i \backslash \bigcap_{i\in I} A_i \cap A_0\Big) && (38.28a)\\
&= P\Big(\bigcap_{i\in I} A_i\Big) - P\Big(\bigcap_{i\in I} A_i \cap A_0\Big) && (38.28b)\\
&= P\Big(\bigcap_{i\in I} A_i\Big)\big(1 - P(\complement A_0)\big) && (38.28c)\\
&= P\Big(\bigcap_{i\in I} A_i\Big) P(\complement A_0). && (38.28d)
\end{aligned}$$

□

Proposition 38.13.
Les événements $(A_i)_{i=1,\dots,n}$ sont indépendants si et seulement si les variables aléatoires associées $\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_n}$ le sont.

Démonstration. La tribu engendrée par la variable aléatoire $\mathbb{1}_{A_k}$ est

$$\mathcal{A}_{\mathbb{1}_{A_k}} = \{\varnothing, A_k, \complement A_k, \Omega\}. \tag{38.29}$$

En effet si $1 \in B$, alors $A_i \subset \mathbb{1}_{A_i}^{-1}(B)$, et si $0 \in B$, alors $\complement A_i \subset \mathbb{1}_{A_i}^{-1}(B)$. Les éléments 0 et 1 sont tous deux soit dans B, soit hors de B. Cela donne les 4 possibilités énumérées dans (38.29).

Supposons que les événements (A_i) sont indépendants. Nous devons vérifier que les tribus le soient, c'est à dire que les événements A_i et $\complement A_j$ sont indépendants. Cela est une conséquence du lemme 38.12. □

Théorème 38.14 (Doob[195])**.**
Soit $X\colon \Omega \to \mathbb{R}^d$ une variable aléatoire. Une fonction $Y\colon \Omega \to \mathbb{R}^p$ est une variable aléatoire $\mathcal{A}_X$-mesurable si et seulement s'il existe une fonction borélienne $f\colon \mathbb{R}^d \to \mathbb{R}^p$ telle que $Y = f(X)$.

Proposition 38.15.
Soient des variables aléatoires $X_k\colon \Omega \to \mathbb{R}^{d_k}$ des variables aléatoires indépendantes et des fonctions boréliennes $f_k\colon \mathbb{R}^{d_k} \to \mathbb{R}^{p_k}$. Alors les variables aléatoires $f_k(X_k)$ sont indépendantes.

Démonstration. Le théorème 38.14 assure que les applications

$$f_k \circ X_k\colon \Omega \to \mathbb{R}^{d_k} \tag{38.30}$$

sont $\mathcal{A}_{X_k}$-mesurables. En particulier pour tout borélien $B \subset \mathbb{R}^{p_k}$, nous avons $X_k^{-1} \circ f_k^{-1}(B) \in \mathcal{A}_{X_k}$. Nous avons donc

$$\sigma(f_k \circ X_k) \subset \sigma(X_k), \tag{38.31}$$

et par conséquent les tribus $\sigma(f_k \circ X_k)$ sont indépendantes étant donné que les tribus $\sigma(X_k)$ le sont. □

Lemme 38.16 (Lemme de regroupement)**.**
Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité et $(\mathcal{A})_{i\in I}$ une famille de tribus indépendantes dans $\mathcal{A}$. Si $(M_j)_{j\in J}$ est une partition de I, alors les tribus

$$\mathcal{B}_j = \sigma\Big(\bigcup_{i\in M_j} \mathcal{A}_i\Big) \tag{38.32}$$

sont indépendantes.

Si les variables aléatoires $\{X_1, X_2, X_3, X_4, X_5\}$ sont indépendantes, et si f et g sont des fonctions mesurables, alors les variables aléatoires $f(X_2, x_3, X_5)$ et $g(X_1, X_4)$ sont indépendantes.

Une preuve a l'air d'être donnée dans [414].

38.2.2 Lois conjointes et indépendance

Définition 38.17.
*Deux événements A et B sont dits **indépendants** si*

$$P(A \cap B) = P(A)P(B). \tag{38.33}$$

Si nous considérons n variables aléatoires réelles $X_1, \dots, X_n\colon \Omega \to \mathbb{R}$, la loi du n-uplet $X = (X_1, \dots, X_n)$ est une variable aléatoire $X\colon \Omega \to \mathbb{R}^n$ appelée la **loi conjointe** des lois X_i. Dans ce cas, les variables aléatoires X_i elles-mêmes sont dites lois **marginales** de X.

Proposition 38.18.
Les variables aléatoires $\{X_i\}$ sont indépendantes si et seulement si

$$P_{(X_1,\dots,X_n)} = P_{X_1} \otimes \dots \otimes P_{X_n}. \tag{38.34}$$

Définition 38.19.
*Soient $\{X_i\}_{1\leqslant i\leqslant n}$ des variables aléatoires réelles (pas spécialement indépendantes). La **densité conjointe** de $X_1,\dots,X_n$ est la fonction $f\colon \mathbb{R}^n \to \mathbb{R}$ qui satisfait*

(1) $f(x_1,\dots,x_n) \geqslant 0$ pour tout $(x_1,\dots,x_n) \in \mathbb{R}^n$,

(2) $\int_{\mathbb{R}^n} f = 1$,

(3) pour tout $A_i \subset \mathbb{R}$ nous avons

$$P(\bigcap_{i=1}^{n} X_i \in A_i) = \int_{\prod_i A_i} f(x_1,\dots,x_n)dx_1\dots dx_n. \tag{38.35}$$

Proposition 38.20.
Si les variables aléatoires $X_1,\dots X_n$ sont indépendantes et ont des densités $f_{X_1},\dots,f_{X_n}$, alors la variable aléatoire conjointe $X = (X_1,\dots,X_n)$ a pour densité conjointe la fonction

$$f_X(x_1,\dots,x_n) = f_{X_1}(x_1)\dots f_{X_n}(x_n). \tag{38.36}$$

Démonstration. En partant de la définition de l'indépendance et de la fonction de densité conjointe, ainsi qu'en utilisant le théorème de Fubini,

$$\begin{aligned}\int_{A_1\times\dots\times A_n} f_X(x_1,\dots,x_n)dx_1\dots dx_n &= P(X_1\in A_1,\dots,X_n\in A_n)\\ &= P(X_1\in A_1)\dots P(X_n\in A_n)\\ &= \left(\int_{A_1} f_{X_1}(x_1)dx_1\right)\dots\left(\int_{A_n} f_{X_n}(x_n)dx_n\right)\\ &= \int_{A_1\times\dots\times A_n} f_{X_1}(x_1)\dots f_{X_n}(x_n)dx_1\dots dx_n.\end{aligned} \tag{38.37}$$

La fonction $(x_1,\dots,x_n)\mapsto f_{X_1}(x_1)\dots f_{X_n}(x_n)$ vérifie donc la condition (3) de la définition 38.19. La vérification des autres conditions est immédiate. □

La proposition suivante provient du fait que la mesure d'une loi conjointe est le produit des mesures lorsque les variables aléatoires sont indépendantes (proposition 38.18).

Proposition 38.21 ([195])**.**
Si les variables aléatoires réelles $X_1,\dots,X_n$ sont intégrables et indépendantes, alors leur produit est intégrable et l'espérance du produit est égal au produit des espérances :

$$E(X_1\cdots X_n) = E(X_1)\dots E(X_n). \tag{38.38}$$

38.2.3 Somme et produit de variables aléatoires indépendantes

Soient X et Y, deux variables aléatoires réelles indépendantes. Nous voudrions étudier la loi de la variable aléatoire $S = X + Y$. Nous commençons par calculer la fonction de répartition en utilisant le résultat de la proposition 38.20 :

$$F_{X+Y}(z) = P(X+Y\leqslant z) = \int_{x+y\leqslant z} f_{X,Y}(x,y)dx\,dy \tag{38.39a}$$

$$= \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} dy f_X(x)f_Y(y) \tag{38.39b}$$

$$= \int_{\mathbb{R}} \left(\int_{-\infty}^{z-x} f_Y(y)dy\right) f_X(x)dx \tag{38.39c}$$

$$= \int_{\mathbb{R}} F_Y(z-x)f_X(x)dx. \tag{38.39d}$$

Pour calculer la fonction de densité de S, nous dérivons la fonction de répartition :

$$f_{X+Y}(z) = \frac{dF_{X+Y}}{dz}(z) \tag{38.40a}$$
$$= \int_{\mathbb{R}} f_Y(z-x) f_X(x) dx, \tag{38.40b}$$

ce qui nous amène à dire que la densité de la somme est le produit de convolution [2] des densités :

$$f_{X+Y}(x) = \int_{\mathbb{R}} f_Y(x-t) f_X(t) dt, \tag{38.41}$$

ou encore $f_{X+Y} = f_X * f_Y$.

Notez que nous avons passé sous le silence la difficulté d'inverser la dérivée et l'intégrale. Un exemple sera donné au point 38.5.8.

Lemme 38.22.
Soient X et Y, deux variables aléatoires indépendantes. Alors

$$E(XY) = E(X)E(Y). \tag{38.42}$$

Démonstration. Par indépendance et par proposition 38.20, la fonction de densité conjointe de X et Y vaut $f_{X,Y} = f_X f_Y$. Par conséquent l'utilisation de Fubini sous la forme (16.712) entraine

$$E(XY) = \int_{\mathbb{R}\times\mathbb{R}} xy f_{X,Y}(x,y) dxdy = E(X)E(Y). \tag{38.43}$$

□

Nous dirons tout un tas de chose sur l'indépendance et la variance en 38.5.13, mais pour l'instant nous allons mentionner et démontrer déjà ceci :

Lemme 38.23.
Soient X et Y deux variables aléatoires indépendantes et identiquement distribuées. Alors

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y). \tag{38.44}$$

Démonstration. Par définition, $\text{Var}(X+Y) = E\big([X+Y-E(X)-E(Y)]^2\big)$. En développant le carré et en utilisant le lemme 38.22,

$$\text{Var}(X+Y) = E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 = \text{Var}(X) + \text{Var}(Y). \tag{38.45}$$

□

Exemple 38.24
Deux variables aléatoires non indépendantes dont la covariance est nulle. Nous considérons la variable aléatoire

$$Z\colon \Omega \to \{(1,0),(-1,0),(0,1),(0,-1)\} \tag{38.46}$$

de loi uniforme. C'est à dire que $P\big(Z=z\big) = \frac{1}{4}$ pour tout z. Ensuite nous considérons les variables aléatoires $X = \texttt{proj}_1 \circ Z$ et $Y = \texttt{proj}_2 \circ Z$. Toute personne étant capable de compter jusqu'à 4 voit que

$$P(X=1) = P(X=-1) = \frac{1}{4} \tag{38.47a}$$
$$P(X=0) = \frac{1}{2}, \tag{38.47b}$$

2. Définition 28.39.

et les mêmes probabilités pour Y. De même $E(X) = E(Y) = 0$. Par conséquent

$$\text{Cov}(X, Y) = E(XY) = 0 \tag{38.48}$$

parce que pour tout $\omega \in \Omega$ nous avons soit $X(\omega) = 0$ soit $Y(\omega) = 0$. Ces variables aléatoires X et Y ne sont donc pas corrélées.

Mais elles ne sont pas indépendantes pour autant, comme nous allons le voir pas plus tard qu'immédiatement. Nous avons

$$P(X = 0|Y = 0) = \frac{P(X = 0, Y = 0)}{P(Y = 0)} = 0 \tag{38.49}$$

parce que X et Y ne peuvent pas être simultanément nulles, tandis que

$$P(X = 0)P(Y = 0) = \frac{1}{4}. \tag{38.50}$$

△

38.2.4 Espérance

Nous dirons que la variable aléatoire X a un **moment d'ordre** p si $X \in L^p(\Omega, \mathcal{A}, P)$ $(1 \leqslant p < \infty)$. Si X est **intégrable** (c'est à dire si $X \in L^1$), alors nous définissons l'**espérance** de X par

$$E(X) = \int_\Omega X dP \in \mathbb{R}^d. \tag{38.51}$$

Si $E(X) = 0$ nous disons que la variable aléatoire est **centrée**. La variable aléatoire $X - E(X)$ est la variable aléatoire centrée associée à X.

Le **moment** d'ordre p de la variable aléatoire X est l'espérance

$$m_n(X) = E(X^n). \tag{38.52}$$

Proposition 38.25.
Si X et Y sont deux variables aléatoires (pas spécialement indépendantes), nous avons

$$E(X + Y) = E(X) + E(Y). \tag{38.53}$$

Nous donnons la preuve dans le cas de variables aléatoires indépendantes. Le cas plus général de variable aléatoires non indépendantes peut être trouvé dans [415].

Démonstration. Nous avons le calcul suivant :

$$\begin{aligned}
E(X + Y) &= \int_{\mathbb{R}} x f_{X+Y}(x) dx && \text{(38.54a)}\\
&= \int_{\mathbb{R}} x \int_{\mathbb{R}} f_Y(x - t) f_X(t) dt dx && \text{(38.54b)}\\
&= \int_{\mathbb{R}} f_X(t) \underbrace{\int_{\mathbb{R}} x f_Y(x - t) dx}_{=E(Y)+t} dt && \text{(38.54c)}\\
&= \int_{\mathbb{R}} f_X(t)\big(E(Y) + t\big) dt && \text{(38.54d)}\\
&= E(Y) + \int_{\mathbb{R}} t f_X(t) dt && \text{(38.54e)}\\
&= E(Y) + E(X) && \text{(38.54f)}
\end{aligned}$$

où nous avons utilisé la proposition 38.41 et le fait que l'intégrale sur $\mathbb{R}$ d'une densité vaut 1. □

Une application de l'inégalité de Hölder (proposition 28.23) est la suivante. Si X et Y sont des variables aléatoires intégrables alors

$$E(XY) \leqslant E(X^2)^{1/2}E(Y^2)^{1/2}. \tag{38.55}$$

En effet

$$E(XY) \leqslant \|XY\|_{L^1(\Omega)} \leqslant \|X\|_{L^2(\Omega)}\|Y\|_{L^2(\Omega)}. \tag{38.56}$$

38.2.5 Variance

Si $X \in L^2(\Omega, \mathcal{A}, P)$ alors nous définissons la **variance** de X par

$$\operatorname{Var}(X) = E\big([X - E(X)]^2\big). \tag{38.57}$$

Proposition 38.26.
La variance de la variable aléatoire X peut être exprimée par la formule

$$\operatorname{Var}(X) = E(X^2) - [E(X)]^2 \tag{38.58}$$

où $X^2 = X \cdot X$ et $E(X)^2 = E(X) \cdot E(X)$ sont des produits scalaires dans $\mathbb{R}^d$.

Démonstration. De façon explicite, nous avons

$$E\big([X - E(X)]^2\big) = \int_\Omega \big(X(\omega) - E(X)\big) \cdot \big(X(\omega) - E(X)\big)dP(\omega) \tag{38.59}$$

où $E(X) \in \mathbb{R}^d$ est une constante. En développant le produit scalaire nous avons

$$\begin{aligned} E\big([X - E(X)]^2\big) &= E\big(X^2 - 2X \cdot E(X) + E(X)^2\big) && \text{(38.60a)}\\ &= E(X^2) - 2E(X)^2 + E(X)^2 && \text{(38.60b)}\\ &= E(X^2) - E(X)^2. && \text{(38.60c)} \end{aligned}$$

□

Nous définissons l'**écart-type** de X par

$$\sigma_X = \sqrt{\operatorname{Var}(X)}. \tag{38.61}$$

En d'autres termes,

$$\sigma_X = \|X - E(X)\|_{L^2}. \tag{38.62}$$

On définit encore la **moyenne quadratique** de X par

$$\|X\|_{L^2} = \big[E(X^2)\big]^{1/2}. \tag{38.63}$$

La variable aléatoire

$$\bar{V}_n = \frac{1}{n}\sum_i (X_i - \bar{X}_n)^2 \tag{38.64}$$

est la **variance empirique** de l'échantillon (X_i).

Lemme 38.27.
Si X est une variable aléatoire,

(1) $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$ *pour tout $a \in \mathbb{R}$;*

(2) Si de plus Y est une variable aléatoire indépendante de X, alors $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.

Démonstration. Nous avons

$$\begin{aligned}\operatorname{Var}(X+Y) &= E(X^2+Y^2+2XY)-\big(E(X)+E(Y)\big)^2 &(38.65a)\\ &= E(X^2)+E(Y^2)+2E(XY)-E(X)^2-E(Y)^2+2E(X)E(Y). &(38.65b)\end{aligned}$$

Étant donné que X et Y sont indépendantes nous avons $E(XY)=E(X)E(Y)$ par le lemme 38.22. □

Si les $X_1,\dots,X_n$ sont des variables aléatoires on considère la **moyenne empirique**

$$\bar{X}_n=\frac{X_1+\cdots+X_n}{n}. \tag{38.66}$$

38.2.6 Covariance

Soient X et Y, deux variables aléatoires réelles. Leur **covariance** est définie par

$$\operatorname{Cov}(X,Y)=E\Big[\big(X-E(X)\big)\big(Y-E(Y)\big)\Big] \tag{38.67}$$

L'idée est que la covariance devient grande si X et Y s'écartent de leurs moyennes dans le même sens. Il existe une formule alternative :

$$\operatorname{Cov}(X,Y)=E(XY)-E(X)E(Y) \tag{38.68}$$

En ce qui concerne les dimensions plus hautes, si $X\colon\Omega\to\mathbb{R}^d$ est un vecteur aléatoire de carré intégrable, nous définissons

$$\operatorname{Cov}(X)=E\Big[\big(X-E(X)\big)\otimes\big(X-E(X)\big)\Big] \tag{38.69}$$

où par $a\otimes b$ nous entendons la matrice $(a\otimes b)_{ij}=a_ib_j$. Cela peut aussi être noté a^tb si l'on fait bien attention à qui est un vecteur colonne et qui est un vecteur ligne.

Proposition 38.28.
Si X et Y sont deux variables aléatoires non spécialement indépendantes, nous avons

$$\operatorname{Var}(X+Y)=\operatorname{Var}(X)+\operatorname{Var}(Y)+2\operatorname{Cov}(X,Y). \tag{38.70}$$

Démonstration. Il s'agit d'un calcul en partant de

$$\begin{aligned}\operatorname{Var}(X+Y)&=E\big((X+Y)^2\big)-E(X+Y)^2\\ &=E(X^2)+E(Y^2)+2E(XY)\\ &\quad+\big(E(X)+E(Y)\big)^2-2E(X)^2-2E(X)E(Y)\\ &\quad-2E(Y)E(X)-2E(Y)^2.\end{aligned} \tag{38.71}$$

À partir d'ici il s'agit de recombiner tous les termes pour former la formule annoncée. □

Plus généralement nous avons la formule

$$\operatorname{Var}(\sum_i X_i)=\sum_i\operatorname{Var}(X_i)+2\sum_{1\leqslant i<j\leqslant n}\operatorname{Cov}(X_i,X_j). \tag{38.72}$$

38.2.7 Probabilité conditionnelle : événements

Proposition-définition 38.29.
Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité et $B \in \mathcal{A}$ avec $P(B) > 0$. Alors avec la formule

$$P_B(A) = \frac{P(A \cap B)}{P(B)}, \tag{38.73}$$

*l'espace $(\Omega, \mathcal{A}, P_B)$ est un espace probabilisé. Nous notons $P_B(A)$ le nombre $P(A|B)$ et nous le nommons **probabilité conditionnelle** de A sachant B.*

Démonstration. On vérifie que $(\Omega, \mathcal{A}, P_B)$ est un espace de probabilité parce que $P_B(\Omega) = 1$ et

$$P_B(\bigcup_i A_i) = \sum_i P_B(A_i) \tag{38.74}$$

si les A_i sont deux à deux disjoints. □

Une conséquence immédiate de (38.73) est que si A et B sont des événements indépendants alors

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A). \tag{38.75}$$

La probabilité conditionnelle à B est quelque chose qui ne tient compte que de ce qui se passe dans B. Si K est un événement tel que $A \cap B = K \cap B$, alors

$$P(A|B) = P(K|B). \tag{38.76}$$

Théorème 38.30.
Soient $(B_n)_{n\geqslant 1}$ une partition finie de Ω telle que $P(B_i) > 0$. Soit $A \in \mathcal{A}$ tel que $P(A) > 0$.

(1) Si A, B et C sont des événements, alors

$$P(A \cap B|C) = P(A|B \cap C)P(B|C). \tag{38.77}$$

(2) Si $P(B) > 0$, alors $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.

*(3) On a la **formule des probabilités totales** :*

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) = \sum_i P(A \cap B_i). \tag{38.78}$$

*(4) On a la **formule de Bayes** :*

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_i P(A|B_i)P(B_i)}. \tag{38.79}$$

Démonstration. (1) En développant le membre de droite,

$$\begin{aligned} P(A \cap B|C) &= \frac{P(A \cap B \cap C)}{P(B \cap C)} \frac{P(B \cap C)}{P(C)} \\ &= P(A \cap B|C). \end{aligned} \tag{38.80}$$

(2) C'est la définition de $P(A|B)$ et $P(B|A)$.

(3) Vu que les B_i forment une partition, nous avons

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i). \tag{38.81}$$

(4) En utilisant les deux premiers points, nous trouvons

$$\begin{aligned} P(A|B_k)P(B_k) &= P(A \cap B_k) \\ &= P(B_k|A)P(A) \\ &= P(B_k|A) \sum_i P(A|B_i)P(B_i). \end{aligned} \tag{38.82}$$

□

38.2.8 Espérance conditionnelle

Théorème-définition 38.31 (Définition de l'espérance conditionnelle[327]).
Soit un espace de probabilité $(\Omega, \mathcal{A}, P)$ et une variable aléatoire intégrable $X : \Omega \to \mathbb{R}$. Pour chaque sous tribu $\mathcal{F}$ de $\mathcal{A}$, il existe une (presque partout) unique variable aléatoire $Y : \Omega \to \mathbb{R}$ telle que

(1) Y est $\mathcal{F}$-mesurable

(2) Y est P-intégrable

(3) pour tout $B \in \mathcal{F}$,

$$\int_B X dP = \int_B Y dP. \tag{38.83}$$

Cette variable aléatoire sera notée $E(X|\mathcal{F})$ pour des raisons qui apparaîtront plus tard.

Démonstration. Remarquons que prendre $Y = X$ ne fonctionne pas parce qu'en général si $\mathcal{O}$ est mesurable dans $\mathbb{R}$, alors $X^{-1}(\mathcal{O})$ est dans la tribu $\mathcal{A}$, mais n'est pas automatiquement dans la tribu $\mathcal{F}$. Il faudra donc un peu plus travailler.

Unicité Si Y_1 et Y_2 vérifient tous les deux les conditions, l'ensemble $\{Y_1 < Y_2\}$ est un élément de $\mathcal{F}$ et nous avons

$$\int_{\{Y_1<Y_2\}} X = \int_{Y_1<Y_2} Y_1 = \int_{Y_1<Y_2} Y_2. \tag{38.84}$$

En particulier nous avons $\int_{\{Y_1<Y_2\}}(Y_1 - Y_2) = 0$ et donc

$$(Y_1 - Y_2)\mathbb{1}_{Y_1-Y_2} = 0 \tag{38.85}$$

presque partout. Le corollaire 16.136 montre alors que $Y_1 - Y_2 \geqslant 0$ presque partout. De la même manière, l'ensemble $\{Y_2 < Y_1\}$ est dans $\mathcal{F}$ et nous trouvons que $Y_2 - Y_1 \geqslant 0$ presque partout. Par conséquent $Y_1 = Y_2$ presque partout.

Existence dans le cas de carré intégrable Nous supposons que $X \in L^2(\Omega, \mathcal{A}, P)$ et nous considérons K, le sous-ensemble de $L^2(\Omega, \mathcal{A}, P)$ des fonctions $\mathcal{F}$-mesurables. Le théorème des projections 27.5 nous indique que

$$L^2(\Omega, \mathcal{A}, P) = K \oplus K^\perp \tag{38.86}$$

par la décomposition $X = \mathtt{proj}_K X + (X - \mathtt{proj}_K X)$. La variable aléatoire $Y = \mathtt{proj}_K X$ a les propriétés d'être $\mathcal{F}$-mesurable et $\langle Y - X, Z\rangle = 0$ pour tout $Z \in K$. Soit $A \in \mathcal{F}$, si nous considérons $Z = \mathbb{1}_A$, la dernière condition signifie que

$$\int_\Omega X\mathbb{1}_A = \int_\Omega Y\mathbb{1}_A, \tag{38.87}$$

ou encore

$$\int_A Y = \int_A X. \tag{38.88}$$

La variable aléatoire $Y = \mathtt{proj}_K(X)$ répond donc à la question lorsque $X \in L^2(\Omega, \mathcal{F}, P)$.

Existence en général Nous considérons maintenant que $X \in L^1(\Omega, \mathcal{A}, P)$. Quitte à décomposer X en deux fonctions positives X_+ et X_- telles que $X = X_+ + X_-$, nous pouvons supposer que X est positive. Par hypothèse $X \in L^1(\Omega, \mathcal{A}, P)$; pour chaque $n \in \mathbb{N}$ nous posons

$$X_n(\omega) = \min\{X(\omega), n\}. \tag{38.89}$$

Étant donné que la mesure P est une mesure de probabilité, les constantes sont intégrables et $X_n \in L^2(\Omega, \mathcal{A}, P)$. De plus la suite (X_n) est croissante et

$$\lim_{n\to\infty} X_n(\omega) = X(\omega). \tag{38.90}$$

Si nous notons encore K l'ensemble des variables aléatoires dans $L^2(\Omega, \mathcal{A}, P)$ qui sont $\mathcal{F}$-mesurables, pour chaque n nous avons donc la variable aléatoire

$$Y_n = \texttt{proj}_K X_n = E(X_n|\mathcal{F}) \tag{38.91}$$

qui est $\mathcal{F}$-mesurable et telle que

$$\int_A X_n = \int_A Y_n \tag{38.92}$$

pour tout $A \in \mathcal{F}$. Nous voudrions prouver que la variable aléatoire $Y = \lim_n Y_n$ existe et est la solution au problème, c'est à dire est $E(X|\mathcal{F})$.
Commençons par prouver que $Y_n \geqslant 0$ presque partout. Pour cela nous remarquons que l'ensemble $\{Y_n < 0\}$ est mesurable et

$$0 \geqslant \int_{Y_n<0} Y_n = \int_{Y_n<0} X_n \geqslant 0. \tag{38.93}$$

La première inégalité est évidente et la dernière est due au fait que X_n est positive. Par conséquent

$$\int_{Y_n<0} Y_n = 0 \tag{38.94}$$

et le lemme 16.136 conclut que $P(Y_n < 0) = 0$.
Soit $Z : \Omega \to \mathbb{R}$ une variable aléatoire positive dans $L^2(\Omega, \mathcal{A}, P)$. Montrons que $\texttt{proj}_K Z$ est encore positive. Pour cela nous considérons l'ensemble $A = \{\texttt{proj}_K Z < 0\}$ et les inégalités

$$0 \leqslant \int_A Z = \int_A \texttt{proj}_K Z \leqslant 0, \tag{38.95}$$

ce qui montre que $\int_A \texttt{proj}_K Z = 0$ et par conséquent que $P\{\texttt{proj}_K(Z) < 0\} = 0$. Cela nous montre que la projection depuis L^2 conserve la positivité.
Étant donné que $X_{n-1} - X_n \geqslant 0$ nous avons aussi

$$Y_{n-1} - Y_n \geqslant 0 \tag{38.96}$$

La suite de fonctions

$$n \mapsto Y_n = E(X_n|\mathcal{F}) \tag{38.97}$$

est croissante et vérifie le théorème de la convergence monotone :

$$\int_A X = \lim_{n\to\infty} \int_A X_n = \lim_{n\to\infty} \int_A E(X_n|\mathcal{F}) = \int_A \lim_{n\to\infty} E(X_n|\mathcal{F}) = \int_A Y. \tag{38.98}$$

Par conséquent $E(X|\mathcal{F})$ existe et

$$Y = \lim_{n\to\infty} E(X_n|\mathcal{F}) = E(X|\mathcal{F}). \tag{38.99}$$

□

38.32.
Vu la définition 38.31 nous pourrions croire que la variable aléatoire $E(X|\mathcal{F}) = X$ fait l'affaire. Il n'en est rien parce que la variable aléatoire X n'est pas spécialement $\mathcal{F}$-mesurable alors qu'il est requis que $E(X|\mathcal{F})$ le soit. Avec la tribu $\mathcal{F} = \{\varnothing, \Omega\}$, nous n'avons en général pas que $X^{-1}(B) \in \mathcal{F}$ pour tout borélien B.

Par contre si $\sigma(X)$ est le tribu engendrée par la variable aléatoire X, alors $E\big(X|\sigma(X)\big) = X$.

Définition 38.33.
*Soit Z une variable aléatoire. L'**espérance conditionnelle** « X sachant Z » est la variable aléatoire*

$$E(X|Z) = E(X|\sigma(Z)) \tag{38.100}$$

où $\sigma(Z)$ est la tribu engendrée par Z. Le membre de droite est une variable aléatoire définie en 38.31.

Définition 38.34.
Soient $A \in \mathcal{A}$ un événement et $\mathcal{F}$ une sous-tribu de $\mathcal{A}$. Nous définissons $P(A|\mathcal{F})$ par

$$P(A|\mathcal{F}) = E(\mathbb{1}_A|\mathcal{F}). \tag{38.101}$$

Notons que cela est une variable aléatoire et non un réel. Le membre de droite est l'espérance conditionnelle de la variable aléatoire $\mathbb{1}_A$ par rapport à $\mathcal{F}$ définie en 38.31.

Et l'espérance conditionnelle d'un événement par rapport à une variable aléatoire est :

$$E(A|X) = E\big(A|\sigma(A)\big). \tag{38.102}$$

Proposition 38.35.
Soit une espace probabilisé $(\Omega, \mathcal{A}, P)$ ainsi qu'une variable aléatoire X à valeurs dans $\mathbb{R}^d$, et un événement A. Alors

$$E\big(P(A|X)\big) = P(A). \tag{38.103}$$

Démonstration. Tout le point de la preuve est de remarquer que $E(\mathbb{1}_A) = E(\mathbb{1}_A|X)$.

La formule $E(\mathbb{1}_A) = E(\mathbb{1}_A|X)$ La notation $E(\mathbb{1}_A|X)$ est un raccourcis pour écrire la variable aléatoire $E\big(\mathbb{1}_A|\sigma(X)\big)$. Cette dernière est l'application $\Omega \to \mathbb{R}^d$ telle que

$$\int_B E\big(\mathbb{1}_A|\sigma(X)\big) = \int_B \mathbb{1}_A \tag{38.104}$$

pour tout borélien B de $\mathbb{R}^d$ tout en étant $\sigma(X)$-mesurable. Comme expliqué en 38.32, il est tentant de dire $E\big(\mathbb{1}_A|\sigma(X)\big) = \mathbb{1}_A$, mais ce n'est pas le cas parce qu'il n'y a aucune raisons que $\mathbb{1}_A$ soit une application $\sigma(X)$-mesurable. Au niveau des espérances, par contre, l'égalité tient :

$$E\big(E(\mathbb{1}_A|X)\big) = \int_\Omega E(\mathbb{1}_A|X) = \int_\Omega \mathbb{1}_A = E(\mathbb{1}_A) \tag{38.105}$$

où nous avons utilisé le fait que Ω lui-même soit $\sigma(X)$-mesurable.

La preuve Nous avons alors

$$P(A) = E(\mathbb{1}_A) = E\big(E(\mathbb{1}_A|X)\big), \tag{38.106}$$

alors que $E(\mathbb{1}_A|X) = P(A|X)$. En mettant l'un dans l'autre :

$$P(A) = E\big(P(A|X)\big). \tag{38.107}$$

□

Proposition 38.36 (Transitivité de l'espérance conditionnelle)**.**
Si $\mathcal{B}_2 \subseteq \mathcal{B}_1 \subset \mathcal{A}$ alors

$$E\Big(E(X|\mathcal{B}_1)|\mathcal{B}_2\Big) = E(X|\mathcal{B}_2). \tag{38.108}$$

Démonstration. Si $B \in \mathcal{B}_2$, nous avons

$$\int_B E\big(E(X|\mathcal{B}_1)|\mathcal{B}_2\big)dP = \int_B E(X|\mathcal{B}_1)dP = \int_B dP. \tag{38.109}$$

La première égalité est la définition de l'espérance conditionnelle par rapport à $\mathcal{B}_2$. La seconde égalité est celle de l'espérance conditionnelle par rapport à $\mathcal{B}_1$ et le fait que $B \in \mathcal{B}_2 \subset \mathcal{B}_1$. Ce que nous avons prouvé est que

$$E\big(E(X|\mathcal{B}_1)|\mathcal{B}_2\big) \tag{38.110}$$

est une variable aléatoire $\mathcal{B}_2$-mesurable vérifiant la condition

$$\int_B E\big(E(X|\mathcal{B}_1)|\mathcal{B}_2\big) = \int_B E(X|\mathcal{B}_2) \tag{38.111}$$

pour tout $B \in \mathcal{B}_2$. C'est donc $E(X|\mathcal{B}_2)$ par la partie unicité du théorème 38.31. □

Proposition 38.37.
Soit $(\Omega, \mathcal{F}, P)$ un espace de probabilité, soit $\mathcal{A}$ une sous tribu de $\mathcal{F}$ et X, une variable aléatoire $\mathcal{F}$-mesurable et intégrable. Alors la variable aléatoire $E(X|\mathcal{A})$ du théorème 38.31 est l'unique (presque partout) variable aléatoire à être $\mathcal{A}$-mesurable telle que nous ayons

$$E\big(E(X|\mathcal{A})Y\big) = E(XY). \tag{38.112}$$

pour toute variable aléatoire Y $\mathcal{A}$-mesurable.

Démonstration. Supposons pour commencer que Y soit une fonction simple positive, alors $Y = \sum_{i=1}^{n} a_i \mathbb{1}_{E_i}$ et nous avons

$$\int_\Omega E(X|Y) = \sum_i a_i \int_{E_i} E(X|\mathcal{A}) \tag{38.113a}$$

$$= \sum_i a_i \int_{E_i} X \tag{38.113b}$$

$$= \int_\Omega XY. \tag{38.113c}$$

Maintenant si Y est mesurable et bornée, elle est limite croissante de fonctions étagées bornées (proposition 16.103) et le résultat tient par la convergence monotone, théorème 16.116.

Si Y n'est pas positive, nous séparons $Y = Y_+ - Y_-$.

Pour l'unicité, soit Z et Z' deux variables aléatoires telles que pour toute variable aléatoire Y,

$$\int_\Omega ZY = \int_\Omega XY = \int_\Omega Z'Y. \tag{38.114}$$

Si nous prenons $Y = \mathbb{1}_{\{Z \neq Z'\}}$, nous avons

$$0 = \int_\Omega (Z - Z')\mathbb{1}_{Z\neq Z'} = \int_{Z\neq Z'} Z - Z', \tag{38.115}$$

d'où le fait que $P(Z \neq Z') = 0$. □

Si X est une variable aléatoire dont la tribu engendrée est indépendante de la tribu $\mathcal{F}$, nous voudrions que la connaissance de $\mathcal{F}$ n'influence pas la connaissance de X, c'est à dire que

$$E(X|\mathcal{F}) = E(X). \tag{38.116}$$

Ce que nous avons est même mieux. Nous avons le lemme suivant.

Lemme 38.38 ([195]).
Les tribus $\mathcal{F}_1$ et $\mathcal{F}_2$ sont indépendantes si et seulement si

$$E(U|\mathcal{F}_1) = E(U) \tag{38.117}$$

pour toute variable aléatoire U étant $\mathcal{F}_1$-mesurable.

Ici, par $E(U)$ nous entendons la variable aléatoire constante prenant la valeur numérique $E(U)$ en tout point de Ω.

Démonstration. Si $\mathcal{F}_1$ et $\mathcal{F}_2$ sont indépendantes, alors pour tout $B \in \mathcal{F}_2$ nous avons

$$\int_B U dP = E(U\mathbb{1}_B) \tag{38.118a}$$

$$= E(U)E(\mathbb{1}_B) \tag{38.118b}$$

$$= E(U)\int_\Omega \mathbb{1}_B dP \tag{38.118c}$$

$$= \int_B E(U) dP. \tag{38.118d}$$

Justifications.

— L'intégrale $\int_B UdP$ a un sens même si $B \in \mathcal{F}_2$ alors que U est $\mathcal{F}_1$-mesurable. Le supremum (16.289) définissant l'intégrale est tout de même bien défini, en particulier, l'ensemble sur lequel on prend le supremum est non vide.
— Pour (38.118b), la variable aléatoire U est $\mathcal{F}_1$-mesurable (donc la tribu engendrée par U est dans $\mathcal{F}_1$) alors que $\mathbb{1}_B$ est $\mathcal{F}_2$-mesurable. Les tribus engendrées étant indépendantes, les variables aléatoires le sont et nous pouvons décomposer l'espérance.

Ce que montre le calcul (38.118) est que $E(U)$ est une variable aléatoire $\mathcal{F}_2$-mesurable (parce que constante) dont l'intégrale sur chaque élément de $\mathcal{F}_2$ vaut l'intégrale de U. Par la partie unicité du théorème 38.31, nous déduisons que $E(U) = E(U|\mathcal{F}_2)$. □

Corollaire 38.39.
Si X est une variable aléatoire et si $\mathcal{F}$ est une tribu, alors

$$E\big(E(X|\mathcal{F})\big) = E(X). \tag{38.119}$$

Démonstration. Il suffit d'appliquer la définition (38.83) à $B = \Omega$:

$$E\big(E(X|\mathcal{F})\big) = \int_\Omega E(X|\mathcal{F})(\omega)dP(\omega) = \int_\Omega X(\omega)dP(\omega) = E(X). \tag{38.120}$$

□

Exemple 38.40
Soient X_1, X_2 deux variables aléatoires à valeurs dans $\{0,1\}$ avec probabilité $1/2$ et indépendantes. Nous considérons $S = X_1 + X_2$. La situation est modélisée par l'espace

$$\Omega = \{(0,0),(0,1),(1,0),(1,1)\} \tag{38.121}$$

et les variables aléatoires

$$X_i(\omega_1,\omega_2) = \omega_i \tag{38.122a}$$
$$S(\omega_1,\omega_2) = \omega_1 + \omega_2. \tag{38.122b}$$

Pour vérifier que de cette manière nous avons bien que X_1 est indépendante de X_2, nous commençons par voir les tribus associées. Un ouvert de $\mathbb{R}$ soit contient 0 et 1, soit contient un seul des deux soit n'en contient aucun des deux. En appliquant X_1^{-1} à chacune de ces quatre situations nous voyons que la tribu $\sigma(X_1)$ est

$$\mathcal{F}_1 = \sigma(X_1) = \big\{\{(0,0),(0,1)\},\{(1,0),(1,1)\},\Omega,\varnothing\big\}. \tag{38.123}$$

De la même façon nous avons

$$\mathcal{F}_2 = \sigma(X_1) = \big\{\{(0,0),(1,0)\},\{(0,1),(1,1)\},\Omega,\varnothing\big\}. \tag{38.124}$$

Nous posons

$$A_0 = \{(0,0),(0,1)\} \tag{38.125a}$$
$$A_1 = \{(1,0),(1,1)\} \tag{38.125b}$$
$$B_0 = \{(0,0),(1,0)\} \tag{38.125c}$$
$$B_1 = \{(0,1),(1,1)\}. \tag{38.125d}$$

Étant donné que $A_i \cap B_j = (i,j)$, nous avons toujours que $P(A_i \cap B_j) = \frac{1}{4} = P(A_i)P(B_j)$. L'indépendance est donc assurée.

Calculons l'espérance conditionnelle $E(S|\mathcal{F}_1)$. Une fonction $\mathcal{F}_1$-mesurable doit être constante sur A_0 et A_1, donc l'espérance conditionnelle est une fonction constante sur A_0 et A_1 dont l'intégrale sur ces ensembles est égale à l'intégrale de S. Nous avons en particulier

$$\int_{A_0} E(S|\mathcal{F}_1) = \int_{A_0} S, \tag{38.126}$$

c'est à dire

$$E(S|\mathcal{F}_1)(0,0) + E(S|\mathcal{F}_1)(0,1) = S(0,0) + S(0,1) = 1. \tag{38.127}$$

Nous en concluons que $E(S|\mathcal{F}_1)(0,0) = E(S|\mathcal{F}_1)(0,1) = \frac{1}{2}$. Cela correspond à l'intuition que si on est au point $(0,1)$ ou au point $(0,0)$ en ne sachant que X_1, nous ne savons que le premier zéro, et donc l'espérance de la somme est $\frac{1}{2}$.

Un calcul très similaire montre que

$$E(S|\mathcal{F}_1)(1,0) = E(S|\mathcal{F}_1)(1,1) = \frac{3}{2}. \tag{38.128}$$

Cela correspond au fait qu'en ces points, nous ne savons que le fait que le premier tirage a donné 1, et donc que l'espérance est $\frac{3}{2}$.

Complétons ce tour d'horizon en mentionnant que la tribu engendrée par X_1 et X_2 est la tribu des parties de Ω, de telle façon que l'espérance conditionnelle de S sachant X_1 et X_2 est égale à S. △

Proposition 38.41 ([195])**.**
Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé et X, Y deux variables aléatoires sur Ω réelles. Soit $\mathcal{B}$ une sous-tribu de $\mathcal{A}$. Nous supposons que $X \in L^1(\Omega, \mathcal{A}, P)$, que $Z \in L^\infty(\Omega, \mathcal{B}, P)$ et que $XZ \in L^1(\Omega, P)$. Alors

$$E(ZX|\mathcal{B}) = ZE(X|\mathcal{B}) \tag{38.129}$$

presque sûrement.

Démonstration. Nous commençons par prouver que

$$\int_\Omega ZE(X|\mathcal{B}) = \int_\Omega ZX. \tag{38.130}$$

Si $Z = \mathbb{1}_B$ pour un ensemble $B \in \mathcal{B}$, alors cette égalité est vraie par définition de l'espérance conditionnelle[3]. Donc cette égalité est correcte tant que Z est une variable aléatoire $\mathcal{B}$-mesurable et étagée. Nous considérons alors, grâce au lemme 16.100, une suite Z_n de variables aléatoires étagées et $\mathcal{B}$-mesurables avec $|Z_n| < Z$. Pour chaque n nous avons donc

$$\int_\Omega Z_n X = \int_\Omega ZE(X|\mathcal{B}). \tag{38.131}$$

Notre idée est de passer à la limite. Vu que Z et Z_n sont bornées (et donc intégrables sur Ω), pour chaque n nous avons $|Z_n X| \leqslant M|X|$ où M majore Z et donc tous les Z_n de façon uniforme vis-à-vis de n. Tout cela pour dire que le théorème de la convergence dominée fonctionne et que

$$\lim_{n\to\infty} \int_\Omega Z_n X = \int_\Omega ZX. \tag{38.132}$$

D'autre part vu que $X \in L^1$ et que $\Omega \in \mathcal{B}$ nous avons l'égalité $\int_\Omega E(X|\mathcal{B}) = \int_\Omega X$, ce qui prouve que $|E(X|\mathcal{B})|$ est intégrable. Cela nous permet d'utiliser encore la convergence dominée avec l'inégalité $|Z_n E(X|\mathcal{B})| \leqslant |E(X|\mathcal{B})|$ pour écrire

$$\lim_{n\to\infty} \int_\Omega Z_n E(X|\mathcal{B}) = \int_\Omega ZE(X|\mathcal{B}). \tag{38.133}$$

En passant à la limite des deux côtés de (38.131) nous avons donc

$$\int_\Omega ZE(X|\mathcal{B}) = \int_\Omega ZX. \tag{38.134}$$

L'égalité (38.130) est prouvée.

3. Théorème 38.31.

Nous passons maintenant à la preuve de l'égalité demandée : $E(EX|\mathcal{B}) = ZE(X|\mathcal{B})$. Pour cela il faut montrer que pour tout $B \in \mathcal{B}$ nous avons

$$\int_B ZE(X|\mathcal{B}) = \int_B ZX. \tag{38.135}$$

Cela n'est rien d'autre que l'égalité (38.130) que nous venons de prouver avec $Z\mathbb{1}_B$ au lieu de Z. □

Proposition 38.42.
Soit une variable aléatoire réelle $X \in L^1(\Omega, \mathcal{A}, P)$. Pour toute variable aléatoire $Y : \Omega \to \mathbb{R}^d$, il existe une fonction borélienne $\mathcal{A}_Y$-mesurable $h : \mathbb{R}^d \to \mathbb{R}$ telle que

$$E(X|Y) = h \circ Y. \tag{38.136}$$

Démonstration. Nous utilisons le résultat de Doob (théorème 38.14). Par définition $E(X|Y)$ est une variable aléatoire réelle $\mathcal{A}_Y$-mesurable, et il existe une fonction borélienne $h : \mathbb{R}^d \to \mathbb{R}$ telle que $E(X|Y) = f \circ Y$. □

Cette fonction $h : \mathbb{R}^d \to \mathbb{R}$ nous permet de définir

$$E(X|Z = z) = h(z). \tag{38.137}$$

Cela est l'espérance conditionnelle d'une variable aléatoire par rapport à une valeur donnée d'une autre variable aléatoire.

38.2.9 Probabilité conditionnelle : tribu

Soit un espace probabilisé $(\Omega, \mathcal{A}, P)$.

Lemme 38.43.
Soit $(B_i)_{i\in\mathbb{N}}$ une partition de Ω en éléments de $\mathcal{A}$ deux à deux disjoints tels que $P(B_i) \neq 0$. Soit $\mathcal{F}$ la tribu engendrée par les B_i. Une variable aléatoire réelle est $\mathcal{F}$-mesurable si et seulement si elle est constante sur chaque B_i.

Proposition 38.44.
Soit $(B_i)_{i\in\mathbb{N}}$ une partition de Ω en éléments de $\mathcal{A}$ deux à deux disjoints tels que $P(B_i) \neq 0$. Soit $\mathcal{F}$ la tribu engendrée par les B_i. Soit une variable aléatoire X. Alors nous avons :

$$E(X|\mathcal{F}) = \sum_{i\in\mathbb{N}} \left(\frac{1}{P(B_i)} \int_{B_i} X dP \right) \mathbb{1}_{B_i}. \tag{38.138}$$

Démonstration. Si X est une variable aléatoire, alors la variable aléatoire $E(X|\mathcal{F})$ définie en 38.31 est une variable aléatoire $\mathcal{F}$-mesurable et elle est donc constante sur les ensembles B_i par le lemme 38.43 :

$$E(X|\mathcal{F}) = \sum_{i\in\mathbb{N}} a_i \mathbb{1}_{B_i}. \tag{38.139}$$

Étant donné que, par construction, B_i est $\mathcal{F}$-mesurable, nous avons

$$\int_{B_i} X dP = \int_{B_i} E(X|\mathcal{F}) = \sum_j a_j \int_{B_i} \mathbb{1}_{B_j} = \sum_j a_j \delta_{ij} P(B_j) = a_i P(B_i). \tag{38.140}$$

Par conséquent

$$a_i = \frac{1}{P(B_i)} \int_{B_i} X dP \tag{38.141}$$

et

$$E(X|\mathcal{F}) = \sum_{i\in\mathbb{N}} \left(\frac{1}{P(B_i)} \int_{B_i} X dP \right) \mathbb{1}_{B_i}, \tag{38.142}$$

ce qu'il fallait. □

Notons que si $B \in \mathcal{A}$ alors la tribu engendrée par B est aussi celle engendrée par la partition $\{B, \complement B\}$ de Ω. Cette circonstance nous permet d'aller plus loin.

Proposition 38.45.
Soit un espace probabilisé $(\Omega, \mathcal{A}, P)$ et un événement $B \in \mathcal{A}$ avec sa tribu engendrée $\mathcal{F} = \sigma(B)$. Alors

$$E(\mathbb{1}_A|\mathcal{F}) = P(A|B)\mathbb{1}_B + P(A|\complement B)\mathbb{1}_{\complement B}. \tag{38.143}$$

Démonstration. Nous allons particulariser la formule (38.138). Si $B \in \mathcal{A}$ nous considérons la partition $\{B, \complement B\}$ de Ω et la tribu engendrée

$$\mathcal{F} = \{\varnothing, B, \complement B, \Omega\}. \tag{38.144}$$

La formule (38.138) devient

$$E(X|\mathcal{F}) = \left(\frac{1}{P(B)}\int_B X dP\right)\mathbb{1}_B + \left(\frac{1}{P(\complement B)}\int_{\complement B} X dP\right)\mathbb{1}_{\complement B}. \tag{38.145}$$

Si nous considérons $A \in \mathcal{A}$, nous écrivons cette égalité avec $X = \mathbb{1}_A$ pour obtenir

$$E(\mathbb{1}_A|\mathcal{F}) = \frac{P(A \cap B)}{P(B)}\mathbb{1}_B + \frac{P(A \cap \complement B)}{P(\complement B)}\mathbb{1}_{\complement B} = P(A|B)\mathbb{1}_B + P(A|\complement B)\mathbb{1}_{\complement B} \tag{38.146}$$

parce que nous avons reconnu la probabilité conditionnelle $P(A|B)$ de la définition 38.29. □

Remarque 38.46.
Les nombres $P\big(A|\sigma(B)\big) = P\big(\mathbb{1}_A|\sigma(B)\big)$ n'est pas la probabilité conditionnelle de A sachant B.

Il nous reste à définir la probabilité conditionnelle d'un événement relativement à une variable aléatoire.

Définition 38.47.
Si la variable aléatoire X est à valeurs discrètes, nous disons que $P(A|X)$ est la variable aléatoire de valeur

$$P(A|X)(\omega) = P(A|X = X(\omega)). \tag{38.147}$$

Dans le cas d'une variable aléatoire à valeurs continues, cette définition ne fonctionne pas parce que la condition $X = X(\omega)$ est souvent de probabilité nulle, tandis que c'est toujours une mauvaise idée de conditionner par rapport à un événement de probabilité nulle. C'est la base du paradoxe de Borel. La bonne définition du conditionnement de l'événement A par rapport à la variable aléatoire X est

Définition 38.48.
Si A est un événement et X une variable aléatoire à valeurs continues dans $\mathbb{R}$, nous définissons

$$P(A|X) = P(A|\sigma(X)) = E\big(\mathbb{1}_A|\sigma(X)\big). \tag{38.148}$$

La première égalité est une notation. La seconde est la définition.

Cette définition s'appuie également sur la définition 38.31.

Proposition 38.49.
Si X est une variable aléatoire et si A est un événement, alors

$$E\big(P(A|X)\big) = P(A). \tag{38.149}$$

Démonstration. Nous commençons par le cas discret, c'est à dire $X\colon \Omega \to \mathbb{N}$. Nous notons $p_k = P(X = k)$. En décomposant l'intégrale sur Ω par rapport à l'union disjointe

$$\Omega = \bigcup_{k\in\mathbb{N}} A_k = \bigcup_{k\in\mathbb{N}} \{\omega \in \Omega \text{ tel que } X(\omega) = k\}, \tag{38.150}$$

nous obtenons

$$E\big(P(A|X)\big) = \int_\Omega P(A|X)(\omega)dP(\omega) \tag{38.151a}$$
$$= \sum_{k=0}^{\infty} \int_{A_k} P(A|X = X(\omega))dP(\omega) \tag{38.151b}$$
$$= \sum_k \int_{A_k} \frac{P(A \cap X = k)}{P(X = k)} dP(\omega) \qquad \text{dans } A_k,\ X(\omega) = k \tag{38.151c}$$
$$= \sum_k \frac{1}{p_k} P(A \cap X = k) \underbrace{\int_{A_k} 1 dP(\omega)}_{=P(A_k)=p_k} \tag{38.151d}$$
$$= \sum_k P(A \cap X = k) \tag{38.151e}$$
$$= P(A). \tag{38.151f}$$

Nous devons maintenant prouver la propriété dans le cas où X prend des valeurs continues. Pour cela il suffit d'appliquer le corollaire 38.39 :

$$E\big(E(\mathbb{1}_A|\sigma(A))\big) = E(\mathbb{1}_A) = P(A). \tag{38.152}$$

□

38.2.10 Variables de Rademacher indépendantes

Une variable aléatoire de Rademacher est une variable aléatoire qui prend les valeurs 1 et -1 avec probabilité $\frac{1}{2}$. Nous pouvons en décrire une explicitement de la façon suivante. L'espace probabilité est à deux éléments : $\Omega = \{a, b\}$ avec la mesure $P(\{a\}) = P(\{b\}) = \frac{1}{2}$. La variable aléatoire est alors l'application $X : \Omega \to \mathbb{R}$ donnée par $X(a) = 1$ et $X(b) = -1$.

Soient X et Y deux variables aléatoires de Rademacher indépendantes. Cela donne $\Omega = \{a, b\}^2$ et

$$\begin{array}{cccc} X(a,a) = 1 & X(a,b) = 1 & X(b,a) = -1 & X(b,b) = -1 \\ Y(a,a) = 1 & Y(a,b) = -1 & Y(b,a) = 1 & Y(b,b) = -1 \end{array} \tag{38.153}$$

Remarque 38.50.
Si une variable aléatoire d'un certain type est donnée par une application $X : \Omega \to \mathbb{R}$, pour construire des variables aléatoires indépendantes identiquement distribuées, il faut considérer les variables aléatoires sur (au moins) le produit $\Omega \times \Omega$ munie de la mesure produit.

Tribu du produit XY Quelle est la tribu de la variable aléatoire produit XY ? Le produit XY peut prendre les valeurs 1 et -1. Nous avons

$$(XY)^{-1}(1) = \{(a,a),(b,b)\} (XY)^{-1}(-1) \quad = \{(a,b),(b,a)\} \tag{38.154}$$

La tribu est donc

$$\sigma(XY) = \{\Omega, \varnothing, A, B\} \tag{38.155}$$

avec $A = \{(a,a),(b,b)\}$ et $B = \{(a,b),(b,a)\}$.

Calcul de $E(X|XY)$ La définition de l'espérance à calculer est le théorème 38.31. Pour chaque élément B de $\sigma(XY)$ nous avons besoin de $\int_B X = \int_B E(X|XY)$. Nous notons $V = E(X|XY)$ pour alléger la notation. Nous avons

$$4 \int_A V = V(a,a) + V(b,b) \tag{38.156}$$

et

$$4 \int_A X = X(a,a) + X(1,1) = 0. \tag{38.157}$$

Pourquoi le facteur 4 ? Parce que sur Ω nous avons la mesure produit de celle que dont nous avions parlé sur $\{a,b\}$. C'est la mesure d'équiprobabilité et donc chaque singleton a mesure $1/4$. Pour plus de détails, il y a le théorème 16.166.

Nous en déduisons $V(a,a)+V(b,b)=0$. Mais pour tout $t\in\mathbb{R}$ nous avons $V^{-1}(t)\in\sigma(XY)$ parce que la contrainte est que V soit XY-mesurable. En particulier

$$V^{-1}\big(V(a,a)\big) \tag{38.158}$$

est un mesurable qui contient (a,a). C'est donc soit Ω, soit $\{(a,a),(b,b)\}$. Dans les deux cas nous avons $V(a,a)=V(b,b)$ et nous en déduisons $V(a,a)=V(b,b)=0$.

En faisant de même avec $\int_B V=V(a,b)+V(b,a)$ nous déduisons $V(a,b)=V(b,a)=0$ et au final nous avons

$$E(X|XY)=0. \tag{38.159}$$

Cette égalité signifie $E(X|XY)(\omega)=0$ pour tout $\omega\in\Omega$.

Calcul de $E(X|X+Y)$ Il ne faudrait pas croire que, seulement parce que X a une espérance nulle, nous trouverons une espérance nulle quel que soit le conditionnement. Juste pour le plaisir, nous calculons $E(X|X+Y)$.

La variable aléatoire $X+Y$ peut prendre trois valeurs : -2, 0 et 2. La tribu engendrée par $X+Y$ doit en particulier contenir $A=\{(a,a)\}$, $B=\{(b,b)\}$ et $C=\{(a,b),(b,a)\}$.

Nous notons $V=E\big(X|\sigma(X+Y)\big)$. Vu que

$$\int_A V=\int_A V, \tag{38.160}$$

nous avons $V(a,a)=X(a,a)=1$. Même chose pour B qui donne $V(b,b)=X(b,b)=-1$. En ce qui concerne l'intégrale sur C nous avons

$$V(a,b)+V(b,a)=X(a,b)+X(b,a)=0. \tag{38.161}$$

Par ailleurs l'ensemble $V^{-1}\big(V(a,b)\big)$ est un ensemble mesurable qui doit au moins contenir (a,b). Vu la tribu que nous avons, cela doit également contenir (b,a), de telle sorte que $V(a,b)=V(b,a)$. La relation (38.161) nous permet alors de conclure que $V(a,b)=V(b,a)=0$.

Quoi qu'il en soit, l'espérance conditionnelle $E(XY|X+Y)$ n'est pas nulle.

Calcul de $E(XY|\sigma(XY))$. Celle-là, elle est facile par 38.32 : c'est XY.

Nous aurions pu croire que si X et Y sont indépendantes, alors

$$E(XY|\mathcal{A})=E(X|\mathcal{A})E(Y|\mathcal{A}). \tag{38.162}$$

L'exemple que nous venons de faire montre qu'il n'en est rien.

Exemple 38.51([416])
Un autre exemple, peut-être plus simple, pour contredire l'équation (38.162). Soient X et Y des expériences indépendantes de pile ou face non truquées. Les résultats sont représentés par 0 et 1. Nous notons $\mathcal{A}$ la tribu engendrée par l'événement « les résultats des deux lancers sont différents » ; c'est à dire la tribu engendrée par l'événement $A=(1,0),(1,0)$. La variable aléatoire X et la tribu $\mathcal{A}$ sont indépendants (définition 38.6), donc, donc $E(X|\mathcal{A})=E(X)=1/2$. Pareil pour Y. En revanche, le produit XY est nul sur A donc $E(XY|\mathcal{A})$ aussi. Ça ne peut donc être égal à la constante $1/4=(1/2)^2$. △

38.2.11 Un petit paradoxe

Attention : ce qui est écrit ici est ma réflexion personnelle sur le sujet. Merci de me dire si je me trompe.

Soit une famille dont vous savez seulement qu'il y a exactement deux enfants. Trois questions :

(1) Vous frappez, une fille ouvre la porte et dit « Bonjour, je suis l'aînée ». Quelle est la probabilité que l'autre enfant soit une fille ?

(2) Vous frappez, une fille ouvre la porte et dit « Bonjour ». Quelle est la probabilité que l'autre enfant soit une fille ?

(3) Vous demandez aux parents s'il y a au moins une fille, ils répondent « oui ». Vous demandez si l'autre est une fille, quelle est la probabilité qu'ils répondent « oui » ?

Dans les trois cas l'intuition dit que la probabilité est 1/2. Il semble que de plus la (2) et la (3) soient les mêmes parce que l'on sait qu'il y a une fille et on se demande quelle est la probabilité qu'il y ait deux filles.

Nous allons voir ça de plus près.

38.2.11.1 « Bonjour, je suis l'aînée »

Résolution Si nous notons X_0 et X_1 les variables aléatoires donnant le sexe des deux enfants, ce sont des variables aléatoires indépendantes et identiquement distribuées, avec $P(X_i = f) = \frac{1}{2}$. La formule (38.73) de la probabilité conditionnelle ainsi que l'indépendance donnent :

$$P(X_1 = f | X_2 = f) = \frac{P(X_1 = f, X_2 = f)}{P(X_2 = f)}. \tag{38.163}$$

Le numérateur vaut $\frac{1}{4}$ et le dénominateur vaut $\frac{1}{2}$; le résultat vaut $\frac{1}{2}$. Fin de l'histoire.

Simulation Voici un petit programme qui simule la situation. Il retourne clairement 1/2.

```
#! /usr/bin/python3
# -*- coding: utf8 -*-

"""
Vous frappez à la porte d'une famille qui a deux enfants. Une ↩
   fille ouvre la porte et vous dit "Je suis l'aînée".
Quelle est la probabilité que l'autre soit une fille ?
"""


import random

def famille():
    """
    return a pair of 'f' and 'g'.
    """
    a=[random.choice( ['g','f'] )]
    a.append(random.choice( ['g','f'] ))
    return a

def toctoc():
    """
    - Create a family with two children.
    - Pick the second one, the elder.
```

```
    - if it is a 'g', return None.
    - if it is a 'f', return the other one.
    """
    F=famille()
    if F[1] != 'f':
        return None
    else :
        if F[0]=='f':
            return 1
        else :
            return 0

N_girl_opens=0
N_girl_other=0
for k in range(1,10000):
    res=toctoc()
    if res is not None:
        N_girl_opens = N_girl_opens+1
        N_girl_other = N_girl_other + res

proba=N_girl_other/N_girl_opens
print(proba)        # ~0.5, intuitively correct.
```

tex/sage/simul_famille_aine.py

38.2.11.2 « Bonjour »

Nous frappons à la porte, une fille ouvre en disant « bonjour », sans préciser si elle est la première ou la seconde. Quelle est la probabilité que l'autre soit une fille ? Naïvement on croirait que la probabilité est également $\frac{1}{2}$. Un raisonnement moins naïf montre le contraire. Et nous allons voir qu'un raisonnement encore moins naïf montre que la probabilité est bien $\frac{1}{2}$.

Premier raisonnement Voici le raisonnement qui est, à mon avis, faux. Vu que l'enfant qui ouvre la porte est une fille, la famille a une des compositions suivantes : fg, ff ou gf. Le cas où une fille ouvre la porte *et* que l'autre est également une fille est seulement le cas ff dont la probabilité est $\frac{1}{3}$. Pour justifier cela nous considérons le couple de variables aléatoires (X_1, X_2) et le conditionnement $A = \{X_1 = f\} \cup \{X_2 = f\}$: évidemment $P(A) = \frac{3}{4}$. Nous calculons facilement la loi du couple (X_1, X_2) conditionné à A :

$$P(X_1 = f, X_2 = f|A) = \frac{P(\{X_1 = f, X_2 = f\} \cap A)}{P(A)} = \frac{1/4}{3/4} = \frac{1}{3}; \tag{38.164}$$

de même,

$$P(X_1 = g, X_2 = f|A) = \frac{1}{3} \tag{38.165a}$$

$$P(X_1 = f, X_2 = g|A) = \frac{1}{3} \tag{38.165b}$$

$$P(X_1 = g, X_2 = g|A) = 0 \tag{38.165c}$$

Donc sachant A, la probabilité que la famille soit constituée de deux filles est $\frac{1}{3}$.

Simulation Pour en avoir le cœur net, nous avons écrit ce programme :

```
#! /usr/bin/python3
# -*- coding: utf8 -*-

"""
Vous frappez à la porte d'une famille qui a deux enfants. Une fille ouvre la porte.
Quelle est la probabilité que l'autre soit une fille ?
"""


import random

def famille():
    """
    return a pair of 'f' and 'g'
    """
    a=[random.choice( ['g','f'] )]
    a.append(random.choice( ['g','f'] ))
    return a

def toctoc():
    """
    - Create a family with two children.
    - Choose one (the one who opens the door)
    - if it is a 'g', return None.
    - if it is a 'f', return the other one.
    """
    F=famille()
    s=random.choice([0,1])
    if F[s] != 'f':
        return None
    else :
        t=(s+1)%2
        if F[t]=='f':
            return 1
        else :
            return 0

N_girl_opens=0
N_girl_other=0
for k in range(1,10000):
    res=toctoc()
    if res is not None:
        N_girl_opens = N_girl_opens+1
        N_girl_other = N_girl_other + res

proba=N_girl_other/N_girl_opens
print(proba)        # ~0.5, intuitively correct.
```

tex/sage/simul_famille_simple.py

Le faisant tourner, la réponse est sans appel : la fréquence observée est beaucoup plus proche de 0.5 que de 0.33 ou 0.66.

La vraie réponse Nous considérons les variables aléatoires $X_0, X_1 : \Omega_E \to \{f, g\}$ avec probabilité $\frac{1}{2}$. De plus nous considérons une nouvelle variable aléatoire qui donne le numéro de l'enfant qui ouvre la porte :

$$\sigma : \Omega_C \to \{0, 1\}. \tag{38.166}$$

Nous devons calculer

$$P\big(X_{1-\sigma} = f | X_\sigma = f\big) = \frac{P\big(X_{1-\sigma} = f, X_\sigma = f\big)}{P(X_\sigma = f)}. \tag{38.167}$$

L'univers de toute cette expérience est, donnée en triples pour les valeurs (X_1, X_2, σ) :

(1) $g, g, 0$ (3) $g, f, 0$ (5) $f, g, 0$ (7) $f, f, 0$
(2) $g, g, 1$ (4) $g, f, 1$ (6) $f, g, 1$ (8) $f, f, 1$.

Cela fait 8 cas en tout. Nous avons

$$\{X_\sigma = f\} = \{(g, f, 1), (f, g, 0), (f, f, 0), (f, f, 1)\}. \tag{38.168}$$

et

$$\{X_{1-\sigma} = f\} \cap \{X_\sigma = f\} = \{(f, f, 0), (f, f, 1)\}. \tag{38.169}$$

Donc

$$P\big(X_{1-\sigma} = f, X_\sigma = f\big) = \frac{2}{8} = \frac{1}{4} \tag{38.170}$$

et

$$P(X_\sigma = f) = \frac{4}{8} = \frac{1}{2}. \tag{38.171}$$

Au final,

$$P\big(X_{1-\sigma} = f | X_\sigma = f\big) = \frac{1/4}{1/2} = \frac{2}{4} = \frac{1}{2}. \tag{38.172}$$

38.2.11.3 Le parent qui répond aux questions

Nous avons une famille de deux enfants dont nous savons qu'au moins un des deux est une fille. Quelle est la probabilité que la famille contienne deux filles ? Cela est a priori la même question que celle où une fille ouvre la porte sans dire si elle est l'aînée ou non.

Simulation Commençons par la simulation :

```
#! /usr/bin/python3
# -*- coding: utf8 -*-

"""
Vous savez qu'une famille a deux enfants.
Vous demandez à un parent si il y a une fille.
Réponse : Oui.
Question : quelle est la probabilité que ce soient deux filles ?
"""


import random

def famille():
    """
    return a pair of 'f' and 'g'.
```

```
    """
    a=[random.choice( ['g','f'] )]
    a.append(random.choice( ['g','f'] ))
    return a

def toctoc():
    """
    - Create a family with two children.
    - If 'gg', there are no girls -> return None
    - If 'gf', there is a girl but the other is a boy -> 0
    - If 'fg', there is a girl but the other is a boy -> 0
    - If 'ff', there is a girl and the other is a girl -> 1
    """
    F=famille()
    if F==['g','g'] :
        return None
    if F==['g','f']:
        return 0
    if F==['f','g']:
        return 0
    if F==['f','f']:
        return 1

N_at_least_one_girl=0
N_two_girls=0
for k in range(1,10000):
    res=toctoc()
    if res is not None:
        N_at_least_one_girl=N_at_least_one_girl+1
        N_two_girls=N_two_girls+res

proba=N_two_girls/N_at_least_one_girl
print(proba)        # ~0.333. Beware !
```

tex/sage/simul_famille_une_fille.py

Et là, bing, la réponse est clairement plutôt 0.33 que 0.5.

Résolution Nous avons les variables aléatoires X_1 et X_2 qui valent 0 ou 1 suivant que l'enfant soit une fille ou un garçon ; ce sont des variables aléatoires indépendantes et identiquement distribuées. Nous définissons la variable aléatoire somme

$$S = X_1 + X_2 \tag{38.173}$$

qui compte le nombre de filles. La question est de calculer

$$P(S = 2|S \geqslant 1) = \frac{P(S = 2 \cap S \geqslant 1)}{P(S \geqslant 1)} = \frac{P(S = 2)}{P(S \geqslant 1)}. \tag{38.174}$$

L'événement $S = 2$ est réduit au singleton $\{ff\}$ et sa probabilité est $\frac{1}{4}$. Au contraire l'événement $S \geqslant 1$ est l'ensemble $\{fg, gf, ff\}$ et sa probabilité est $\frac{3}{4}$. Nous avons donc

$$P(S = 2|S \geqslant 1) = \frac{1/4}{3/4} = \frac{1}{3}. \tag{38.175}$$

Et là, la réponse est 1/3 et non 1/2 comme d'aucuns auraient pu le croire.

Précision Notons que l'événement $S \geqslant 1$ n'est pas le même que l'événement $X_\sigma = f$. En effet

$$S \geqslant 1 = \{(ff,1),(ff,2),(fg,1),(fg,2),(gf,1),(gf,2)\} \tag{38.176}$$

tandis que

$$\{X_\sigma = f\} = \{(ff,1),(ff,2),(fg,1),(gf,2)\}. \tag{38.177}$$

38.2.11.4 À propos des simulations

Si vous lisez ces lignes avec l'intention de passer l'agrégation en utilisant Sage à l'épreuve de modélisation, vous devez être capable de refaire les trois simulations. Notez de plus que les simulations ici sont données avec python3 alors que Sage utilise Python2. Je ne vous dit pas si ça change quelque chose.

Allez oui, je vous dit. Si vous changez dans `simul_famille_une_fille.py` la première ligne pour utiliser python2 au lieu de python3, le résultat affiché sera `0` et non `0.333`. La raison est que dans Python2, l'opérateur `/` entre deux entiers est une **division entière**. Autrement dit : le résultat 0.33 est arrondi à zéro.

Solution : forcer python à interpréter le `/` comme une vraie division. Pour Sage, ça donne ceci comme début de programme :

```
1 #! /usr/bin/sage
2 # -*- coding: utf8 -*-
3
4 from __future__ import division
```

tex/frido/codeSnip_3.py

Importez toujours `division` de `__future__` .

Ah oui, et dernière remarque : pour autant que je le sache, le jour de l'oral, vous n'aurez que Sage en mode notebook. Je ne sais pas si l'import fonctionne aussi bien.

Sinon vous pouvez forcer la division dans les `float` de la façon suivante : `a/float(b)`.

38.2.12 Inégalité de Jensen

Proposition 38.52 (Inégalité de Jensen)**.**
Soit g une fonction convexe[4] *sur $\mathbb{R}$ et une variable aléatoire $Y \in L^1(\Omega,\mathcal{A},P)$ telle que $g\circ Y$ soit également L^1. Alors*

$$g\big(E(Y|\mathcal{F})\big) \leqslant E\big((g\circ Y)|\mathcal{F}\big). \tag{38.178}$$

Démonstration. La convexité de g et la proposition 19.98 nous donnent deux suites (a_n) et (b_n) dans $\mathbb{R}$ telles que pour tout $x \in \mathbb{R}$,

$$g(x) = \sup_{n\in\mathbb{N}}(a_n x + b_n). \tag{38.179}$$

Nous avons alors

$$a_n E(Y|\mathcal{F}) \overset{p.s.}{=} E\big(a_n Y + b_n|\mathcal{F}\big) \leqslant E(g\circ Y|\mathcal{F}). \tag{38.180}$$

L'inégalité est due au fait que $g\circ Y$ est le supremum sur les n de a_nY+b_n. Pour chaque n, l'inégalité (38.180) est fausse sur un ensemble de mesure nulle $R_n \subset \Omega$. L'union

$$R = \bigcup_{n\in\mathbb{N}} R_n \tag{38.181}$$

est encore de mesure nulle. Sur $\Omega\backslash R$, nous avons

$$a_n E(Y|\mathcal{F}) + b_n \leqslant E(g\circ Y|\mathcal{F}). \tag{38.182}$$

4. Définition 19.85.

Vu que cela est vrai presque partout et pour tout n nous passons a supremum et nous avons encore presque partout l'inégalité

$$\sup_{n\in\mathbb{N}}\big(a_nE(Y|\mathcal{F})+b_n\big)\leqslant E(g\circ Y|\mathcal{F}). \tag{38.183}$$

□

Si nous ne nous intéressons pas à $E(Y|\mathcal{F})$ mais seulement à $E(Y)$, alors une démonstration plus simple est donnée sur Wikipédia[417].

38.2.13 Fonction de répartition

Définition 38.53.
*Si X est une variable aléatoire réelle, nous définissons sa **fonction de répartition** par*

$$\begin{aligned}F_X\colon\mathbb{R}&\to[0,1]\\ F_X(x)&=P(X\leqslant x).\end{aligned} \tag{38.184}$$

Remarque 38.54.
La fonction de répartition est discontinue en a si $P(X=a)>0$. En particulier nous ne pouvons pas dire

$$P(X\geqslant a)=1-F_X(a). \tag{38.185}$$

38.2.14 Fonction caractéristique

Définition 38.55.
*La **fonction caractéristique** de la variable aléatoire $X\colon\Omega\to\mathbb{R}$ est la fonction réelle définie par*

$$\Phi_X(t)=E(e^{itX}). \tag{38.186}$$

Une autre façon d'écrire la définition est

$$\Phi_X(t)=\int_{\mathbb{R}}e^{itX}dP_X(x), \tag{38.187}$$

ou encore, si X a une densité f_X,

$$\Phi_X(t)=\int_{\mathbb{R}}e^{itx}f_X(x)dx \tag{38.188}$$

Nous reconnaissons la transformée de Fourier :

$$\Phi_X(t)=\hat{f}_X(-t/2\pi). \tag{38.189}$$

La proposition suivante se déduit en utilisant le théorème de dérivation sous l'intégrale 19.30.

Proposition 38.56.
Soit X une variable aléatoire qui accepte un moment d'ordre $r\geqslant 1$. Alors la fonction caractéristique Φ_X est r fois continument dérivable et

$$\Phi_X^{(r)}(t)=E\big((iX)^re^{itX}\big). \tag{38.190}$$

Démonstration. Nous étudions la fonction

$$\Phi(t)=\int_{\Omega}e^{itX(\omega)}dP(\omega). \tag{38.191}$$

Nous considérons la fonction

$$\begin{aligned}f\colon\mathbb{R}\times\Omega&\to\mathbb{R}\\ (t,\omega)&\mapsto e^{itX(\omega)}.\end{aligned} \tag{38.192}$$

et nous regardons si ce contexte vérifie les hypothèses du théorème 19.30.

(1) Étant donné que X est mesurable, f sera mesurable.

(2) La fonction $t \mapsto e^{itX(\omega)}$ est absolument continue pour chaque ω.

(3) Note : par rapport aux notations du théorème 19.30, nous avons ici $A = \mathbb{R}$. Prenons donc un intervalle (compact) $[a, b] \subset \mathbb{R}$ et calculons

$$\frac{\partial f}{\partial t}(t, \omega) = iX(\omega)e^{itX}, \tag{38.193}$$

et

$$\int_a^b \int_\Omega \left| iX(\omega)e^{itX(\omega)} \right| d\omega\, dt = \int_a^b \int_\Omega |X(\omega)| d\omega\, dt. \tag{38.194}$$

Par hypothèse X accepte un moment d'ordre 1, de sorte que l'intégrale par rapport à ω converge vers un nombre qui ne dépend pas de t. L'intégrale sur t ne pose alors aucun problèmes.

Par conséquent nous pouvons effectuer la première dérivation :

$$\Phi'(t) = \frac{d\Phi}{dt}(t) = \int_\Omega iX(\omega)e^{itX(\omega)} d\omega = E(iXe^{itX}) \tag{38.195}$$

et la fonction Φ' est absolument continue. Ce dernier point est important parce que c'est lui qui permet de faire la récurrence et passer à l'ordre deux.

Le résultat ressort alors en dérivant successivement l'expression (38.193). □

Exemple 38.57
Sachant la fonction caractéristique de X, nous pouvons calculer les moments. Par exemple

$$E(X^2) = \Phi''_X(0). \tag{38.196}$$

△

Théorème 38.58.
Si $\Phi_X = \Phi_Y$, alors $P_X = P_Y$.

Notons que cela n'implique pas que $X = Y$. En effet X et Y peuvent même être définis sur des espaces probabilisés différents.

Dans le cas d'une variable aléatoire vectorielle, nous définissons $\Phi_X : \mathbb{R}^d \to \mathbb{R}$ par

$$\Phi_X(v) = E\left(e^{i\langle v, X\rangle}\right) \tag{38.197}$$

38.2.15 Fonction génératrice des moments, transformée de Laplace

Soit X une variable aléatoire. Sa **transformée de Laplace** ou **fonction génératrice des moments** est la fonction

$$M_X(t) = E(e^{tX}) \tag{38.198}$$

pour chaque t tel que cette espérance existe.

Théorème 38.59 ([418]).
Soit X une variable aléatoire réelle et

$$I_X = \{t \in \mathbb{R} \text{ tel que } E(e^{tX}) \text{ existe}\}. \tag{38.199}$$

La fonction

$$\begin{aligned} M_X : I_X &\to \mathbb{R} \\ t &\mapsto E(e^{tX}) \end{aligned} \tag{38.200}$$

*est la **transformée de Laplace** de X.*

(1) I_X *est un intervalle contenant* 0.

(2) Si I_X *n'est pas réduit à* $\{0\}$ *alors* M_X *se développe en série entière*

$$M_X(t) = \sum_{n=0}^{\infty} \frac{E(X^n)}{n!} t^n. \tag{38.201}$$

(3) Si X *et* Y *sont des variables aléatoires indépendantes, alors* $I_{X+Y} = I_X \cap I_Y$ *et*

$$M_{X+Y} = M_X M_Y \tag{38.202}$$

sur I_{X+Y}.

Démonstration. Le fait que 0 soit dans I_X est évident : $E(1) = 1$. Pour montrer que I_X est un intervalle nous prenons $z \in I_X$ et $0 < s < z$ ou $z < s < 0$, puis nous montrons que $s \in I_X$. Il faut remarquer que dans tous les cas,

$$e^{sX} \leqslant 1 + e^{zX}. \tag{38.203}$$

En effet soit sX et zX sont tous deux à gauche de zéro et alors ils sont tous deux plus petit que 1 ; soit ils sont tous deux à droite de 0 et alors $e^{zX} > e^{sX}$ par croissance de l'exponentielle. Nous avons donc dans tous les cas que

$$E(e^{sX}) = \int_{\mathbb{R}} f_X(x) e^{sX} dx \leqslant \int_{\mathbb{R}} f_X(x)(1 + e^{zx}) = 1 + E(e^{zX}). \tag{38.204}$$

Soit maintenant $a > 0$ tel que $[-a, a] \in I_X$. Étant donné que $e^{a|X|} < e^{aX} e^{-aX}$, l'espérance $E(e^{a|X|})$ existe toujours pour $|t|$. Nous avons

$$\left| M_X(t) - \sum_{n=0}^{N} \frac{E(X^n)}{n!} t^n \right| = \left| E\Big(e^{tX} - \sum_{n=0}^{N} \frac{X^n}{n!} t^n \Big) \right| \tag{38.205a}$$

$$= \left| E\Big(\sum_{n=N+1}^{\infty} \frac{X^n}{n!} t^n \Big) \right| \tag{38.205b}$$

$$\leqslant E\left(\sum_{n=N+1}^{\infty} \frac{|tX|^n}{n!} \right). \tag{38.205c}$$

Maintenant le but est de prendre la limite $N \to \infty$ en inversant la limite et l'espérance par le théorème de la convergence dominée (16.140). L'intégrale à traiter est

$$\lim_{N\to\infty} \int_\Omega \sum_{n=N+1}^{\infty} \frac{|tX(\omega)|^n}{n!} dP(\omega). \tag{38.206}$$

L'intégrante est uniformément borné (en N) par $e^{tX(\omega)}$, qui est intégrable par hypothèse (choix de t). Du coup

$$\lim_{N\to\infty} \int_\Omega \sum_{n=N+1}^{\infty} \frac{|tX(\omega)|^n}{n!} dP(\omega) = E\left(\lim_{N\to\infty} \sum_{n=N+1}^{\infty} \frac{|tX|^n}{n!} \right) = 0. \tag{38.207}$$

□

38.2.16 Loi d'une variable aléatoire

La **loi** de la variable aléatoire X, notée P_X est la mesure image de P par X, c'est à dire

$$P_X(B) = P(X \in B) \tag{38.208}$$

pour tout borélien $B \subset \mathbb{R}^d$. Note :

$$P(X \in B) = P\big(\{\omega \in \Omega \text{ tel que } X(\omega) \in B\}\big) = P\big(X^{-1}(B)\big). \tag{38.209}$$

En particulier P_X est une mesure de probabilité sur $\mathbb{R}^d$ parce que

$$P_X(\mathbb{R}^d) = P(\Omega) = 1. \tag{38.210}$$

Si Q est une mesure de probabilité sur $\mathbb{R}^d$, nous notons $X \sim Q$ si $P_X = Q$. Nous disons alors que « X suit la loi Q ».

La proposition suivante permet de calculer en pratique les intégrales qui définissent par exemple l'espérance mathématique d'une variable aléatoire.

Proposition 38.60 (Théorème de transfert[419])**.**
Si X est une variable aléatoire, alors

$$E(f \circ X) = \int_\Omega f\big(X(\omega)\big) dP(\omega) = \int_{\mathbb{R}^d} f(x) dP_X(x) \tag{38.211}$$

dès que $f\colon \mathbb{R}^d \to \bar{\mathbb{R}}$ est telle qu'une des deux intégrales existe. En particulier, ça marche si f est borélienne.

En utilisant cette proposition nous trouvons une formule pratique pour l'espérance d'une variable aléatoire réelle :

$$E(X) = \int_\Omega X(\omega) dP(\omega) = \int_{\mathbb{R}} x dP_X(x), \tag{38.212}$$

en vertu de la proposition 38.60 appliquée à la fonction $f(x) = x$.

Proposition 38.61.
Une variable aléatoire réelle X est intégrable si et seulement si $P(x = \pm\infty) = 0$ et

$$\int_{\mathbb{R}} |x| dP_X(x) < \infty. \tag{38.213}$$

Le lien entre la densité f_X de la variable aléatoire X et sa loi est

$$P_X(A) = \int_A f_X(x) dx \tag{38.214}$$

pour tout ensemble mesurable $A \subset \mathbb{R}$. Le lien entre la mesure de Lebesgue et celle de la loi de X est alors donné par

$$dP_X(x) = f_X(x) dx. \tag{38.215}$$

En particulier l'espérance de X peut être calculée à partir de sa densité via la formule

$$E(X) = \int_{\mathbb{R}} x dP_X(x) = \int_{\mathbb{R}} x f_X(x) dx. \tag{38.216}$$

38.2.17 Changement de variables

Théorème 38.62.
Soit $\mathcal{O}$, un ouvert de $\mathbb{R}^n$ et $\mathcal{O}'$ un ouvert de $\mathbb{R}^m$ ainsi qu'un difféomorphisme C^1 $\varphi\colon \mathcal{O} \to \mathcal{O}'$. Soit $X\colon \Omega \to \mathbb{R}^n$ une variable aléatoire prenant presque sûrement ses valeurs dans $\mathcal{O}$. Si nous supposons que X a la densité f_X, alors la variable aléatoire $Y = \varphi(X)$ accepte la densité $f_Y\colon \mathcal{O}' \to \mathbb{R}$ donnée par

$$f_Y(v) = f_X\big(\varphi^{-1}(v)\big) |J_{\varphi^{-1}}(v)|. \tag{38.217}$$

Démonstration. Nous devons vérifier la relation

$$P(Y \in B) = \int_B f_Y(v) dv \tag{38.218}$$

pour tout borélien $B \subset \mathcal{O}'$. Nous avons

$$\begin{aligned} P(Y \in B) &= \int_{\mathbb{R}^m} \mathbb{1}_B(v) dP_Y(v) && (38.219a)\\ &= E(\mathbb{1}_B \circ Y) && (38.219b)\\ &= E\big((\mathbb{1}_B \circ \varphi) \circ X\big) && (38.219c)\\ &= \int_{\mathbb{R}^n} (\mathbb{1}_B \circ \varphi)(u) dP_X(u) && (38.219d)\\ &= \int_{\mathbb{R}^n} (\mathbb{1}_B \circ \varphi)(u) f_X(u) du. && (38.219e) \end{aligned}$$

À ce niveau, nous utilisons la formule de changement de variables du théorème 16.232. Nous trouvons alors

$$P(Y \in B) = \int_{\mathbb{R}^m} \mathbb{1}_B\big(\varphi^{-1}(v)\big) f_X\big(\varphi^{-1}(v)\big) |J_{\varphi^{-1}}(v)| dv. \tag{38.220}$$

□

38.3 Convergence

Soient X_i des variables aléatoires réelles définies sur le même espace de probabilité $(\Omega, \mathcal{A}, P)$. Nous disons que X_i converge **presque sûrement** vers la variable aléatoire X et nous notons

$$X_n \xrightarrow{p.s.} X \tag{38.221}$$

si

$$P\big(\{\omega \in \Omega \text{ tel que } X_n(\omega) \to X(\omega)\}\big) = 1 \tag{38.222}$$

où la convergence $X_n(\omega) \to X(\omega)$ est la convergence usuelle dans $\mathbb{R}$.

Lemme 38.63.
Nous avons $X_n \xrightarrow{p.s.} X$ si et seulement s'il existe un événement $A \in \mathcal{A}$ tel que $P(A) = 1$ et tel que $X_n(\omega) \to X(\omega)$ pour tout $\omega \in A$.

Lemme 38.64.
Si X est une variable aléatoire à valeurs dans $\mathbb{R} \cup \{\infty\}$, alors

$$X \wedge n \xrightarrow{p.s.} X \tag{38.223}$$

Démonstration. Si $\omega \in \{X = \infty\}$ alors $(X \wedge n)(\omega) = n$ et d'accord. Si par contre $\omega \in \{X < \infty\}$ alors il existe N tel que si $n \geqslant N$ alors $n \geqslant T(\omega)$ et pour ces grandes valeurs de n nous avons $(T \wedge n)(\omega) = T(\omega)$. □

Définition 38.65 ([420]).
*Nous disons que les variables aléatoires réelles X_n convergent **en probabilité** vers la variable aléatoire X si pour tout $\eta > 0$, on a*

$$P\big(|X_n - X| \geqslant \eta\big) \to 0, \tag{38.224}$$

et on note

$$X_n \xrightarrow{P} X. \tag{38.225}$$

Définition 38.66 (Convergence en loi).
*Nous disons que X_n converge vers X **en loi** vers la variable aléatoire X et nous notons*

$$X_n \xrightarrow{\mathscr{L}} X \tag{38.226}$$

si pour toute fonction continue et bornée g nous avons

$$E\big(g(X_n)\big) \to E\big(g(X)\big) = \int g dP_X. \tag{38.227}$$

Proposition 38.67.
Deux autres caractérisations de la convergence en loi.

(1) *Nous avons* $X_n \xrightarrow{\mathscr{L}} X$ *si et seulement si*

$$\Phi_{X_n}(v) \to \Phi_X(v) \tag{38.228}$$

pour tout $v \in \mathbb{R}^d$. *Ici* Φ_X *est la fonction caractéristique de* X.

(2) *Dans la définition de la convergence en loi nous pouvons indifféremment utiliser les fonctions continues et bornées, les fonctions continues à support compact ou les fonctions bornées uniformément continues.*

Proposition 38.68.
Les types de convergence sont reliées par les implications suivantes :

$$\textit{presque sûre} \;\Rightarrow\; \textit{en probabilité} \;\Rightarrow\; \textit{en loi.} \tag{38.229}$$

La convergence en loi n'implique pas la convergence en probabilité, et par conséquent pas non plus la convergence presque certaine.

Dans le cas particulier $d = 1$ nous avons quelques critères supplémentaires.

Proposition 38.69.
Supposons que les variables aléatoires X_n *soient réelles, et notons* F_n *la fonction de répartition de* X_n. *Si* $F_n(x) \to F(x)$ *pour tout* x *dans l'ensemble des points de continuité de* F, *alors* $X_n \xrightarrow{\mathscr{L}} X$.

Proposition 38.70.
Si les X_n *sont des variables aléatoires réelles positives, et si* X *est une variable aléatoire positive, alors* $X_n \xrightarrow{\mathscr{L}} X$ *si les transformées de Laplace des fonctions de répartition convergent ponctuellement, c'est à dire si*

$$E\big(e^{-\alpha X_n}\big) \to E\big(e^{-\alpha X}\big) \tag{38.230}$$

pour tout $\alpha \geqslant 0$.

Proposition 38.71.
Si les X_n *et* X *sont des variables aléatoires réelles discrètes à valeurs dans* $\{x_0, x_1, \ldots\}$ *alors* $X_n \xrightarrow{\mathscr{L}} X$ *si et seulement si*

$$P(X_n = x_k) \to P(X = x_k) \tag{38.231}$$

pour tout $k \in \mathbb{N}$.

Proposition 38.72 ([413])**.**
Soient X_n *et* X *des variables aléatoires réelles. Nous avons*

$$X_n \xrightarrow{\mathscr{L}} X \tag{38.232}$$

si et seulement si pour tout t *où* F_X *est continue,*

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t). \tag{38.233}$$

Proposition 38.73 ([413])**.**
Soit X_n *une suite de variables aléatoires* $\Omega \to \mathbb{R}^d$ *et* $a \in \mathbb{R}^d$. *Si* $X_n \xrightarrow{\mathscr{L}} a$, *alors*

$$X_n \xrightarrow{P} a. \tag{38.234}$$

Démonstration. Quitte à passer aux composantes, nous pouvons supposer que $d = 1$. Soit $\eta > 0$; nous savons que l'inégalité $|x| > a$ a pour solution $x > a$ ou $x < -a$. Dans notre cas,

$$\begin{aligned}
P\big(|X_n - a| > \eta\big) &= P(X_n - a > \eta) + P(X_n - a < -\eta) && \text{(38.235a)}\\
&= P(X_n > \eta + a) + P(X < a - \eta) && \text{(38.235b)}\\
&= 1 - P(X_n \leqslant \eta + a) + P(X_n \leqslant a - \eta) - P(X_n = a - \eta) && \text{(38.235c)}\\
&\leqslant 1 - F_{X_n}(\eta + a) + F_{X_n}(a - \eta) && \text{(38.235d)}
\end{aligned}$$

où la majoration est l'oubli du terme $P(X_n = a-\eta)$, lequel est positif ou nul et F_{X_n} est la fonction de répartition de X_n, définition 38.53. Nous allons utiliser la proposition 38.72. La fonction de répartition de la variable aléatoire constante $X = a$ est donnée par

$$F_a(t) = P(a \leqslant t) = \mathbb{1}_{[0,\infty[}(t-a). \tag{38.236}$$

Par conséquent, la convergence en loi $X_n \xrightarrow{\mathscr{L}} a$ nous montre que

$$F_{X_n}(t) \to \mathbb{1}_{[0,\infty[}(t-a) \tag{38.237}$$

pour tout $t \neq a$ parce que $t = 0$ est un point de discontinuité de $\mathbb{1}_{[0,\infty[}$. Nous avons par conséquent

$$P\big(|X_n - a| > \eta\big) = 1 - \mathbb{1}_{[0,\infty[}(\eta) + \mathbb{1}_{[0,\infty[}(-\eta) = 1 - 1 + 0 = 0 \tag{38.238}$$

parce que $\eta > 0$. □

Le lemme de Slutsky sera utilisé en combinaison avec la proposition 38.75 pour calculer des intervalles de confiance, voir par exemple ce qui se passe autour de l'équation (39.126).

Lemme 38.74 (Slutsky[421])**.**
Soient X_n et Y_n des suites de variables aléatoires réelles telles que

$$\begin{aligned} X_n &\xrightarrow{\mathscr{L}} X \\ Y_n &\xrightarrow{P} a \in \mathbb{R}. \end{aligned} \tag{38.239}$$

Alors $(X_n, Y_n) \xrightarrow{\mathscr{L}} (X, a)$.

Démonstration. Étant donné que $Y_n \xrightarrow{\mathscr{L}} a$, nous avons $Y_n \xrightarrow{P} a$ par la proposition 38.73. Soit une fonction $f\colon \mathbb{R}^2 \to \mathbb{R}^2$; nous devons prouver que

$$E\big(f(X_n, Y_n)\big) \to E\big(f(X, a)\big). \tag{38.240}$$

Soit $\epsilon > 0$. Nous avons

$$E\big(\|f(X_n, Y_n) - f(X, a)\|\big) \leqslant E\big(\|f(X_n, Y_n) - f(X_n, a)\|\big) + E\big(\|f(X_n, a) - f(X, a)\|\big). \tag{38.241}$$

La fonction $g(t) = f(t, a)$ étant continue et bornée, la convergence en loi $X_n \xrightarrow{\mathscr{L}} X$ donne

$$E\big(\|f(X_n, a) - f(X, a)\|\big) \to 0. \tag{38.242}$$

Étudions à présent le premier terme du membre de droite de (38.241). Pour tout $\eta > 0$ et toute variables aléatoires Z et Z' nous avons

$$E(Z) = E(Z\mathbb{1}_{|Z'|<\eta}) + E(Z\mathbb{1}_{|Z'|\geqslant\eta}). \tag{38.243}$$

Nous décomposons donc le premier terme de (38.241) en

$$\begin{aligned} E\big(\|f(X_n, Y_n) - f(X_n, a)\|\big) = {} & E\big(\|f(X_n, Y_n) - f(X_n, a)\|\mathbb{1}_{|Y_n-a|<\eta}\big) \\ & + E\big(\|f(X_n, Y_n) - f(X_n, a)\|\mathbb{1}_{|Y_n-a|\geqslant\eta}\big). \end{aligned} \tag{38.244}$$

Choisissons maintenant une valeur de η telle que

$$|(x, y) - (x', y')| < \eta \Rightarrow |f(x, y) - f(x', y')| \leqslant \epsilon. \tag{38.245}$$

Un tel η existe par l'uniforme continuité de f. Dans le premier terme, $|Y_n - a| < \eta$, par conséquent

$$\|(X_n, Y_n) - (X_n, a)\| = |Y_n - a| < \eta \tag{38.246}$$

et donc

$$\|f(X_n,Y_n)-f(X_n,a)\|\leqslant\epsilon. \tag{38.247}$$

Le premier terme devient donc

$$E\big(\|f(X_n,Y_n)-f(X_n,a)\|\mathbb{1}_{|Y_n-a|<\eta}\big)\leqslant\epsilon E(\mathbb{1}_{|Y_n-a|<\eta})\leqslant\epsilon \tag{38.248}$$

parce que $E(\mathbb{1}_A)=P(A)\leqslant 1$. Pour le second terme de (38.244) nous effectuons la majoration

$$\|f(X_n,Y_n)-f(X_n,a)\|\leqslant 2\|f\|_\infty \tag{38.249}$$

tandis que la convergence $Y_n\xrightarrow{P}a$ entraine

$$P\big(|Y_n-a|\geqslant\eta\big). \tag{38.250}$$

□

Proposition 38.75 ([422]).
Soient X_i des variables aléatoires telles que

$$X_i\xrightarrow{\mathscr{L}}X \tag{38.251}$$

et h, une fonction mesurable sur l'espace d'arrivée de X_i. Soit C l'ensemble des points de continuité de h au sens

$$C=\{\omega\in\Omega\text{ tel que }h\text{ est continue en }X_i(\omega)\}. \tag{38.252}$$

Alors si $P(X\in C)=1$, nous avons

$$h(X_i)\xrightarrow{\mathscr{L}}h(X). \tag{38.253}$$

Une conséquence de cette proposition couplée au lemme de Slutsky est le résultats suivant, qui est donné sous le nom de théorème de Slutsky sur wikipédia.

Corollaire 38.76.
En reprenant les notations du lemme de Slutsky, si

$$X_n\xrightarrow{\mathscr{L}}X \tag{38.254a}$$
$$Y_n\xrightarrow{P}a, \tag{38.254b}$$

alors

$$X_n+Y_n\xrightarrow{\mathscr{L}}X+a \tag{38.255a}$$
$$X_nY_n\xrightarrow{\mathscr{L}}aX \tag{38.255b}$$
$$Y_n^{-1}X_n\xrightarrow{\mathscr{L}}a^{-1}X \tag{38.255c}$$
$$\tag{38.255d}$$

pourvu que a soit inversible.

Lemme 38.77 (Borel-Cantelli).
Soit (A_n) une suite d'événements (avec $A_n\in\mathcal{A}$ pour tout n).

(1) Si $\sum_{n=0}^{\infty}P(A_n)$ converge, alors

$$P(A_n\ i.s.)=0. \tag{38.256}$$

(2) Si la somme $\sum_n P(A_n)$ diverge, et si de plus les A_i sont indépendants, alors

$$P(A_n\ i.s.)=1. \tag{38.257}$$

La notation $P(A_n \text{ i.s.})$ signifie « infiniment souvent », c'est à dire

$$P(A_n \text{ i.s.}) = P\Big(\bigcap_{N\in\mathbb{N}} \bigcup_{k\geqslant N} A_k\Big) = P(\limsup A_n) \tag{38.258}$$

Une façon de paraphraser le lemme de Borel-Cantelli est que nous avons l'alternative

$$P(\limsup A_n) = \begin{cases} 0 & \text{si } \sum_{n\geqslant 0} P(A_n) < \infty \\ 1 & \text{sinon.} \end{cases} \tag{38.259}$$

Proposition 38.78.
Soit X_n, une suite de variables aléatoires et X une variable aléatoires. Si

$$\sum_n P\big(\|X_n - X\| > \eta\big) < \infty \tag{38.260}$$

pour tout ϵ, alors $X_n \xrightarrow{p.s.} X$.

Démonstration. Fixons ϵ et considérons les événements $A_n = \|X_n - X\| > \epsilon$. L'hypothèse dit que

$$\sum_n P(A_{n,\epsilon}) < \infty \tag{38.261}$$

et le lemme de Borel-Cantelli implique que

$$P(\limsup \|X_n - X\| > \epsilon) = 0. \tag{38.262}$$

Un élément ω est dans $\limsup A_n$ s'il est contenu dans tous les A_n, par conséquent, pour chaque ϵ nous avons l'inclusion

$$\{\omega \in \Omega \text{ tel que } X_n(\omega) \to X(\omega)\} \subset \complement \limsup A_n. \tag{38.263}$$

Nous pouvons aller plus loin et écrire

$$\{\omega \in \Omega \text{ tel que } X_n(\omega) \to X(\omega)\} = \complement\{\omega \in \Omega \text{ tel que } \|X_n - X\| > \epsilon, \forall \epsilon > 0\}. \tag{38.264}$$

Or la probabilité de l'ensemble

$$\{\omega \in \Omega \text{ tel que } \|X_n - X\| > \epsilon\} \tag{38.265}$$

est 0 pour chaque ϵ, et par conséquent la probabilité du membre de droite de (38.264) est 1. □

Exemple 38.79
Considérons une suite de 0 et de 1 dans laquelle le 1 arrive avec une probabilité p et le 0 avec une probabilité $1-p$. Une telle suite est modélisée par une suite de variables aléatoires de Bernoulli $(X_n)_{n\in\mathbb{N}}$ indépendantes de paramètre p.

Question : une telle suite contient elle une infinité de 1 ? Considérons les événements indépendants $A_n = \{X_n = 1\}$. Nous avons

$$\sum_n P(A_n) = \sum_n P(X_n = 1) = \sum_n p = \infty. \tag{38.266}$$

Par Borel-Cantelli et son expression (38.259), nous avons alors

$$P(\limsup A_n) = 1. \tag{38.267}$$

Donc une infinité d'événements A_n se produisent, et nous avons bien une infinité de 1 dans la suite.

Remarque : dans ce raisonnement nous pouvons considérer une probabilité non constante p_n tant que la série $\sum_n p_n$ diverge. △

Exemple 38.80
À propos de maximum. La fonction $h\colon \mathbb{R}^d \to \mathbb{R}$ donnée par $h(x_1,\dots,x_d) = \max_i\{x_i\}$ est une fonction continue. Nous voudrions prouver que si on a une famille (finie en $i = 1,\dots,l$) de suites (en n) variables aléatoires $X_n^{(i)} \xrightarrow{p.s.} a$ convergeant toutes vers la même limite a, alors

$$M_n = \max_i\{X_n^{(i)}\} \xrightarrow{p.s.} a. \tag{38.268}$$

D'abord si nous avons l suite numériques $(x_n^{(i)})$, alors la suite

$$M_n = \max_i x_n^{(i)} \tag{38.269}$$

converge vers la même limite. En effet si $\epsilon > 0$ est donné, il suffit de prendre N_i l'entier tel que $|x_n^{(i)} - a| \leqslant \epsilon$ pour tout $n > N_i$. Et ensuite on prend $N > \max\{N_i\}$.

Si maintenant au lieu de suites numériques, nous avons des variables aléatoires, le résultat reste valable. Nous cherchons à prouver que

$$P\Big(\{\omega \in \Omega \text{ tel que } \max\{X_n^{(i)}(\omega)\} \to a\}\Big) = 1. \tag{38.270}$$

Par ce que nous venons de dire sur les suites numériques, un élément ω n'est pas dans cet ensemble seulement s'il y a un ∞ pour lequel $X_n^{(i)}$ ne converge pas vers a. Or cela, pour chaque i est un événement de probabilité zéro.

Les ω qui ne fonctionneront pas dans l'équation (38.270) sont ceux de la réunion d'un ensemble fini d'ensembles de probabilité nulle. C'est donc de probabilité nulle. △

38.4 Loi des grands nombres, théorème central limite

38.4.1 Loi des grands nombres

Lemme 38.81 (Inégalité de Markov[423])**.**
Soit une variable aléatoire $X \in L^p$ et $\epsilon > 0$. Nous avons

$$P(|X| \geqslant \epsilon) \leqslant \frac{1}{\epsilon^r} E(|X|^r). \tag{38.271}$$

Démonstration. Nous avons

$$E(|X|^r) \geqslant \int_{|X|\geqslant\epsilon} X(\omega)dP(\omega) \geqslant \epsilon^r \int_{|X|\geqslant\epsilon} dP = \epsilon^r P(|X| \geqslant \epsilon). \tag{38.272}$$

□

Corollaire 38.82 ([423])**.**
Soit ϕ, une fonction croissante et positive ou nulle sur l'intervalle I. Soit aussi une variable aléatoire $Y\colon \Omega \to \mathbb{R}$ telle que $P(Y \in I) = 1$. Alors pour tout $b \in I$ tel que $\phi(b) > 0$ nous avons

$$P(Y \geqslant b) \leqslant \frac{E\big[\phi(Y)\big]}{\phi(b)}. \tag{38.273}$$

Théorème 38.83 (Loi forte des grands nombres)**.**
Soit (X_n) une suite de variables aléatoires réelles

(1) indépendantes et identiquement distribuées,

(2) intégrables (c'est à dire dans L^1),

alors

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s.} E(X_1). \tag{38.274}$$

Note : étant donné que les variables aléatoires sont identiquement distribuées, nous avons évidemment $E(X_1) = E(X_2) = \dots$

Problèmes et choses à faire

Est-ce que les variables aléatoires doivent vraiment être réelles ?

Corollaire 38.84.
Si les variables aléatoires réelles X_n sont

(1) indépendantes et identiquement distribuées,

(2) dans L^2

alors

$$\bar{X}_n \xrightarrow{P} E(X_1). \tag{38.275}$$

Démonstration. Nous voulons prouver que pour tout $\eta > 0$,

$$P\big(|\bar{X}_n - E(X_1)| > \eta\big) \to 0. \tag{38.276}$$

Remarquons d'abord que les variables aléatoires X_n étant identiquement distribuées, $E(\bar{X}_n) = E(X_1)$ parce que $E(X_i) = E(X_1)$ pour tout i. L'inégalité de Markov avec $r = 2$ nous donne

$$P\big(|\bar{X}_n - E(\bar{X}_n)| > \eta\big) \leqslant \frac{1}{\eta^2} E\big(|\bar{X}_n - E(\bar{X}_n)|^2\big) \tag{38.277a}$$

$$= \frac{1}{\eta^2} \operatorname{Var}(\bar{X}_n) \tag{38.277b}$$

$$= \frac{1}{n\eta^2} \operatorname{Var}(X_1) \tag{38.277c}$$

où nous avons utilisé la la proposition 38.27 : $\operatorname{Var}(\bar{X}_n) = \operatorname{Var}(X_1)/n$. Au final nous avons prouvé que

$$P\big(|\bar{X}_n - E(\bar{X}_n)| > \frac{1}{n\eta^2} \operatorname{Var}(X_1), \tag{38.278}$$

qui tend vers zéro lorsque $n \to \infty$. □

Proposition 38.85.
Soient X_n des variables aléatoires indépendantes et identiquement distribuées avec $X_n \geqslant 0$. Nous acceptons $E(X_1) = \infty$, c'est à dire que nous relaxons la condition $X_n \in L^1$ par rapport à la loi des grands nombres.

Alors

$$\bar{X}_n \xrightarrow{p.s.} E(X_1) \in [0, \infty]. \tag{38.279}$$

Démonstration. Si $E(X_1) < \infty$, nous sommes dans le cas de la loi des grands nombres. Pour chaque $N \in \mathbb{N}$ nous considérons la suite de variables aléatoires

$$X_n^{(N)} = \min(X_n, N). \tag{38.280}$$

Nous avons évidement $\bar{X}_n^{(N)} \leqslant \bar{X}_n$. Les variables aléatoires $X_n^{(N)}$ étant bornées par N, elles vérifient la loi des grands nombres pour chaque N séparément. Par conséquent nous avons pour chaque N la limite

$$\bar{X}_n^{(N)} \to E(X_1^{(N)}) \tag{38.281}$$

Nous supposons que $E(X_1) = \infty$, par conséquent $\lim_{N\to\infty} E(X_1^{(N)}) = \infty$. Soit $\eta > 0$ et choisissons N de telle manière à avoir

$$E(X_1^{(N)}) > \eta + 1. \tag{38.282}$$

La limite (38.281) nous permet de trouver n_0 tel que pour tout $n > n_0$ nous ayons $\bar{X}_n^{(N)} > \eta$. Au final,

$$\eta < \bar{X}_n^{(N)} \leqslant \bar{X}_n, \tag{38.283}$$

ce qui montre que $\bar{X}_n \to \infty$. □

Exemple 38.86
La loi des grands nombres justifie la pratique courante d'approximer une grandeur physique par la moyenne empirique d'un grand nombre de mesures. △

Exemple 38.87
Citons ici le dernier paragraphe de *Le mystère de Marie Roget* par Edgar Allan Poe, traduit par Charles Baudelaire [5].

> Rien, par exemple, n'est plus difficile que de convaincre le lecteur non spécialiste que, si un joueur de dés a amené les six deux fois coup sur coup, ce fait est une raison suffisante de parier gros que le troisième coup ne ramènera pas les six. Une opinion de ce genre est généralement rejetée tout d'abord par l'intelligence. On ne comprend pas comment les deux coups déjà joués, et qui sont maintenant complètement enfouis dans le Passé, peuvent avoir de l'influence sur le coup qui n'existe que dans le Futur. La chance pour amener les six semble être précisément ce qu'elle était à n'importe quel moment, c'est-à-dire soumise seulement à l'influence de tous les coups divers que peuvent amener les dés. Et c'est là une réflexion qui semble si parfaitement évidente, que tout effort pour la controverser est plus souvent accueilli par un sourire moqueur que par une condescendance attentive.

Dans le cours de la nouvelle, Edgar Poe cite et utilise la théorie des probabilités avec une justesse inaccoutumée dans la littérature. Mais dans ce paragraphe final, Poe montre de façon la plus formelle qu'il n'a *rien* compris à la loi des grands nombres. △

38.4.2 Théorème central limite

Lemme 38.88.
Soit $z_n \to z$ une suite convergente dans $\mathbb{C}$. Alors

$$\left(1 + \frac{z_n}{n}\right)^n \to e^z. \tag{38.284}$$

Théorème 38.89.
Si les variables aléatoires X_n sont

(1) indépendantes et identiquement distribuées de loi parente X,

(2) $X_1 \in L^2(\Omega, \mathcal{A})$,

alors nous notons $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $m = E(X_1)$ et $\sigma^2 = \mathrm{Var}(X_1)$ et nous avons

$$\frac{\bar{X}_n - E(X)}{\sqrt{\mathrm{Var}(\bar{X}_n)}} = \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1). \tag{38.285}$$

Démonstration. Nous allons écrire la démonstration dans le cas de variables aléatoires réelles. La proposition 38.67 dit que la suite X_n converge en loi vers X si et seulement si les fonctions caractéristiques convergent ponctuellement. Nous devons donc prouver, pour chaque [6] $t \in \mathbb{R}$, que

$$\Phi_{\frac{S_n - nm}{\sigma\sqrt{n}}}(t) \to \Phi_{\mathcal{N}(0,1)}(t). \tag{38.286}$$

5. Disponible sur https://fr.wikisource.org/wiki/Le_Mystère_de_Marie_Roget
6. Chuck Norris peut *vraiment* le faire pour *chaque* $t \in \mathbb{R}$

Supposons dans un premier temps que $E(X_i) = 0$ et $\sigma(X_i) = 1$. Dans ce cas nous considérons la fonction

$$\Phi_{\frac{S_n}{\sqrt{n}}} = E\big(e^{i\frac{t}{\sqrt{n}}\sum_{k=1}^{n} X_k}\big) \tag{38.287a}$$
$$= \prod_{k=1}^{n} E\big(e^{i\frac{t}{\sqrt{n}}X_1}\big) \tag{38.287b}$$
$$= \prod_{k=1}^{n} \Phi_{X_1}\left(\frac{t}{\sqrt{n}}\right) \tag{38.287c}$$
$$= \Phi_{X_1}\left(\frac{t}{\sqrt{n}}\right)^n. \tag{38.287d}$$

Cette quantité est a priori complexe ; nous ne pouvons donc pas immédiatement passer au logarithme. Nous pouvons par contre utiliser un développement en puissances de t en nous servant de la proposition 38.56 et de l'hypothèse comme quoi $X_1 \in L^2$:

$$\Phi_{X_1}(t) = \Phi_{X_1}(0) + \Phi'_{X_1}(0)t + \Phi''_{X_1}(0)\frac{t^2}{2} + \alpha(t)t^2 \tag{38.288}$$

où α est une fonction qui a la propriété $\lim_{x\to 0}\alpha(x) = 0$.

En utilisant les hypothèses et la formule de dérivation de la fonction caractéristique,

$$\Phi_{X_1}(0) = 1 \tag{38.289a}$$
$$\Phi'_{X_1}(0) = E\big((iX)\big) = 0 \tag{38.289b}$$
$$\Phi''_{X_1}(0) = E(-X^2) = -\operatorname{Var}(X_1) = -1. \tag{38.289c}$$

Nous avons donc

$$\Phi_{X_1}\left(\frac{t}{\sqrt{n}}\right) = \underbrace{1 - \frac{1}{2}\frac{t^2}{n}}_{\in\mathbb{R}} + \underbrace{\frac{t^2}{n}\alpha\left(\frac{t}{\sqrt{n}}\right)}_{\in\mathbb{C}}, \tag{38.290}$$

de telle sorte que, en considérant une valeur fixée de t,

$$\Phi_{\frac{S_n}{\sqrt{n}}}(t) = \left(1 - \frac{\frac{t^2}{2} + \beta_n}{n}\right)^n \tag{38.291}$$

où $\beta_n = t^2\alpha(t/\sqrt{n})$. Nous avons bien entendu $\lim_{n\to\infty}\beta_n = 0$.

Nous pouvons appliquer le lemme 38.88 pour obtenir la limite

$$\lim_{n\to\infty}\Phi_{\frac{S_n}{\sqrt{n}}}(t) = e^{-t^2/2}. \tag{38.292}$$

La convergence (38.286) est par conséquent prouvée dans le cas où $E(X_i) = 0$ et $\operatorname{Var}(X_i) = 1$.

Considérons maintenant des variables aléatoires avec $E(X_i) = m$ et $\operatorname{Var}(X_i) = \sigma^2$. Elles peuvent être écrites sous la forme

$$X_i = \sigma X_i + m \tag{38.293}$$

où X'_i est d'espérance nulle et de variance un. Nous avons alors

$$S_n = \sigma\sum_{i=1}^{n} X'_i + nm, \tag{38.294}$$

et

$$\frac{S_n - nm}{\sigma\sqrt{n}} = \frac{S'_n}{\sqrt{n}} \tag{38.295}$$

où $S'_n = \sum_i X'_i$. L'étude de la variable aléatoire

$$\frac{S_n - nm}{\sigma\sqrt{n}} \tag{38.296}$$

revient donc à celle de $S'_n/\sqrt{n}$ qui vient d'être effectuée. □

38.90.
À propos du théorème central limite 38.89. Si pour une certaine variable aléatoire X on a $E(X) = m$, alors nous n'avons pas forcément $P(X = m + a) = P(X = m - a)$. Est-ce que le théorème central limite permet cependant d'affirmer que dans un certaine mesure nous avons

$$P(\bar{X}_n = m + a) = P(\bar{X}_n = m - a) \tag{38.297}$$

lorsque n est grand ?

Tel quelle, l'équation (38.297) est en général fausse pour chaque n parce qu'il existe des distributions non symétriques. Mais bien entendu les deux membres tendent vers zéro pour $n \to \infty$. Mais cela n'est pas lié à la symétrie de la distribution gaussienne. C'est seulement le fait que la gaussienne n'a pas de masses ponctuelles.

Par contre, il y a effectivement une assurance de symétrie pour $\bar{X}_n$ lorsque $n \to \infty$. Le fait est que $X_n \xrightarrow{\mathscr{L}} X$ où X est une gaussienne. La fonction de répartition de X est continue partout et la proposition 38.72 nous dit que pour tout $x \in \mathbb{R}$,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(t). \tag{38.298}$$

Vu que le nombre $P\big(\bar{X}_n \in B(m + a, \delta)\big)$ peut être exprimé avec des sommes et différences F_{X_n}, nous avons

$$\lim_{n\to\infty} P\big(\bar{X}_n \in B(m + a, \delta)\big) = P\big(X \in B(m + a, \delta)\big). \tag{38.299}$$

Par symétrie de la gaussienne le membre de droite est égal à $P\big(X \in B(m - a, \delta)\big)$ et nous avons bien

$$\lim_{n\to\infty} P\big(\bar{X}_n \in B(m + a, \delta)\big) = \lim_{n\to\infty} P\big(\bar{X}_n \in B(m - a, \delta)\big). \tag{38.300}$$

Remarque 38.91.
Le théorème central limite s'applique quelle que soit la distribution des variables aléatoires X_i (dans les limites de hypothèses) ; en particulier il ne dit rien sur la moyenne des X_i. Il dit seulement que l'écart de la moyenne « mesurée » à la moyenne « théorique » est une variable aléatoire gaussienne si on a mesuré assez de fois.

Autrement dit, si la durée d'attente à la poste est de 5 minutes, et si j'y vais 2000 fois, alors la probabilité que ma moyenne d'attente soit de 4 minutes est la même que la probabilité qu'elle soit de 6 minutes [7].

Problèmes et choses à faire

La remarque 38.91 est une interprétation personnelle. J'aimerais avoir l'avis de quelqu'un de plus compétent.

Remarque 38.92.
Nous pouvons obtenir la limite (38.292) d'une façon alternative. Nous considérons la détermination du logarithme complexe sur $\mathbb{C}\backslash\mathbb{R}^-$; cela est une fonction analytique (théorème 29.61) vérifiant l'équation

$$e^{\ln(z)} = z \tag{38.301}$$

pour tout $z \in \mathbb{C}\backslash\mathbb{R}^-$ et le développement

$$\ln(1 + z) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{z^n}{n}. \tag{38.302}$$

En particulier, $\ln(1 + z) = z + z\alpha(z)$ où $\lim_{z\to 0} \alpha(z) = 0$. Nous reprenons à l'équation (38.291) en

7. Et en l'occurrence, cette probabilité est nulle parce qu'on est en train de parler de variable aléatoire continue, mais vous voyez l'idée.

fixant t. Nous avons

$$\Phi_{\frac{S_n}{\sqrt{n}}}(t) = \exp\left[\ln\left(\Phi_{\frac{S_n}{\sqrt{n}}}(t)\right)\right] \tag{38.303a}$$

$$= \exp\left[n\ln\left(1+\frac{-\frac{t^2}{2}-\beta_n}{n}\right)\right] \tag{38.303b}$$

$$= \exp\left[-\frac{t^2}{2}-\beta_n+\big(-\frac{t^2}{2}-\beta_n\big)\alpha\left(\frac{-\frac{t^2}{2}-\beta_n}{n}\right)\right]. \tag{38.303c}$$

À la limite $n\to 0$ nous tombons sur $e^{-t^2/2}$.

Remarque 38.93.
Étant donné que la variable aléatoire

$$\frac{S_n-nm}{\sigma\sqrt{n}} \tag{38.304}$$

converge en loi vers $\mathcal{N}(0,1)$, nous avons la convergence des fonctions de répartition partout où la fonction de répartition de la normale est continue [8] (donc sur tout $\mathbb{R}$). En particulier,

$$\left|P\left(\frac{S_n-nm}{\sigma\sqrt{n}}\leqslant x\right)-\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy\right|\to 0. \tag{38.305}$$

Nous avons la borne de **Berry-Esséen** qui donne une estimation de la vitesse de convergence : si $X\in L^3$, alors il existe une constante C, indépendante de x, des X_i et de n telle que

$$\left|P\left(\frac{S_n-nm}{\sigma\sqrt{n}}\leqslant x\right)-\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy\right|\leqslant\frac{X\mu_3}{\sigma^3\sqrt{n}} \tag{38.306}$$

où $\mu_3=E\big(|X_1-m|^3\big)$ est le moment d'ordre 3 de X. La chose à retenir est que la convergence est à la vitesse de $1/\sqrt{n}$.

En dimension $d>1$, nous avons encore un théorème central limite.

Théorème 38.94.
Si $d>1$, et si nous avons des variables aléatoires X_n à valeurs dans $\mathbb{R}^d$ avec

(1) les X_n sont indépendantes et identiquement distribuées

(2) les X_n sont dans L^2.

Alors nous notons $X_1=(X_1^{(1)},\dots,X_1^{(d)})$. Nous avons

$$\frac{S_n-nm}{\sqrt{n}}\xrightarrow{\mathscr{L}}\mathcal{N}(0,\Sigma) \tag{38.307}$$

où Σ est ma matrice de covariance du vecteur aléatoire X_1 :

$$\Sigma=\big(\operatorname{Cov}(X_1^{(1),\dots,X_1^{(d)}})\big)_{i,j=1,\dots,d}. \tag{38.308}$$

38.4.3 Marche aléatoire

Nous considérons un mobile qui se déplace sur l'axe $\mathbb{Z}$. À chaque pas de temps, nous supposons qu'il va faire un pas à gauche avec une probabilité p et un pas à droite avec une probabilité $(1-p)$. Nous nous demandons quel est le mouvement du mobile sur le long terme.

La position S_n du mobile à l'instant n est donnée par

$$S_n=\sum_{i=1}^{n}X_i \tag{38.309}$$

8. Proposition 38.72.

où X_i est le pas effectué à l'instant i. Ce sont des variables de Bernoulli indépendantes avec

$$X_i \overset{\mathscr{L}}{=} p\delta_{-1} + (1-p)\delta_1 \tag{38.310}$$

c'est à dire

$$P(X_i = -1) = p \tag{38.311a}$$
$$P(X_i = 1) = 1 - p. \tag{38.311b}$$

Ces variables vérifient les hypothèses de la loi des grands nombres :

(1) elles sont indépendantes et identiquement distribuées,

(2) elles sont intégrables.

Pour le second point, le calcul est

$$\int_\Omega |X_i| dP = \int_{\mathbb{R}} |x| dP_X = \int_{\mathbb{R}} |x|(p\delta_{-1} + (1-p)\delta_1) = |1-p| + |p| = 1. \tag{38.312}$$

Nous avons par conséquent

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s.} E(X_1) = (1-2p) \tag{38.313}$$

et

$$\frac{S_n}{n} \to (1-2p). \tag{38.314}$$

Si $p \neq 1/2$ nous pouvons conclure que

(1) si $p > 1/2$, alors $S_n \xrightarrow{p.s.} -\infty$

(2) si $p < 1/2$, alors $S_n \xrightarrow{p.s.} \infty$.

De plus nous connaissons la vitesse de divergence : elle est linéaire. Le mobile suit essentiellement l'équation $p(n) = (1-2p)n$.

Remarque 38.95.
Cela ne traite pas le cas $p = 1/2$. Dans ce cas, nous pouvons simplement dire que $S_n = o(n)$.

38.5 Les lois usuelles

38.5.1 Loi de Bernoulli

Une expérience de Bernoulli consiste à tirer au hasard un 0 ou un 1 avec une probabilité p de tomber sur 1 et $1-p$ de tomber sur zéro. Il s'agit donc d'une expérience qui réussit ou qui rate.

Le cas typique est une urne avec des boules indiscernables blanches ou noires. La probabilité p est la proportion de blanches dans l'urne (avec remise entre les tirages). Dans ce cas, nous avons l'espace de probabilité $(\Omega, \mathcal{A}, P)$ où Ω représente l'ensemble des boules, $\mathcal{A}$ est l'ensemble des parties de Ω et P est l'équiprobabilité sur Ω. Une variable aléatoire est une application

$$\begin{aligned} X\colon \Omega &\to \{0,1\} \\ \omega &\mapsto \text{couleur de la boule } \omega. \end{aligned} \tag{38.315}$$

Nous notons $\mathscr{B}(1,p)$ la loi de Bernoulli. Elle a une expression très simple :

$$\mathscr{B}(0,1)\big(\{1\}\big) = p \tag{38.316a}$$
$$\mathscr{B}(0,1)\big(\{0\}\big) = 1-p \tag{38.316b}$$

Une variable aléatoire réelle est de **Bernoulli** de paramètre p $(0 < p < 1)$ si

$$X\colon \Omega \to \mathbb{R} \tag{38.317}$$

avec $P(x=1)=p$ et $P(X=0)=1-p$. En tant que mesure sur $\mathbb{R}$, nous avons

$$P_X = p\delta_1 + (1-p)\delta_0. \tag{38.318}$$

Une fonction h qui réalise le supremum de la formule (16.289) est par exemple une fonction en escalier qui vaut en x le plus petit entier plus grand ou égal à x. L'espérance d'une loi de Bernoulli est alors

$$E(x) = p. \tag{38.319}$$

Étant donné que la variable aléatoire X prend seulement les valeurs 0 et 1, nous avons pour tout ensemble mesurable B

$$P_{X^2}(B) = P(X^2 \in B) = P(X \in B), \tag{38.320}$$

et par conséquent $P_{X^2} = P_X$ et $E(X^2) = E(X)$. Nous trouvons donc la variance

$$\operatorname{Var}(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p). \tag{38.321}$$

38.5.2 Loi binomiale

Une expérience binomiale consiste à répéter n expériences de Bernoulli de paramètre p et de compter le nombre de réussites. Une telle expérience peut être réalisée selon la procédure suivante.

Soit une urne contenant N boules dont une proportion p de 1 et $1-p$ de 0. Une expérience binomiale de paramètres n et p consistera à prendre n boules *avec remise* et à compter le nombre de 1 obtenus.

En termes d'espaces probabilisé, nous avons Ω qui est l'ensemble des tuples de taille n à valeurs dans $\{0,1\}$, la tribu $\mathcal{A}$ est l'ensemble des parties de Ω, et la probabilité P est l'équiprobabilité :

$$P(\omega) = \frac{1}{N^n} \tag{38.322}$$

si il y a N boules dans l'urne. Nous construisons alors la variable aléatoire

$$X(\omega) = \sum_{i=1}^{n} \omega_i \tag{38.323}$$

où ω est une suite de taille n de 0 et de 1.

Calculons $P(X=k)$. Il s'agit de considérer tous les sous-ensembles de taille n de Ω contenant exactement k fois 1. Il y a $\binom{n}{k}$ manière de décider lesquelles des n boules seront blanches. Ensuite, chaque boule blanche peut être choisie parmi les m boules disponibles, et chaque boule noire peut être choisie parmi les $(N-m)$ disponibles. Nous avons donc

$$P(X=k) = \binom{n}{k} \frac{m^k (N-m)^{n-k}}{N^k}. \tag{38.324}$$

En effet la mesure de probabilité sur Ω est la mesure de comptage renormalisée par le cardinal de Ω qui vaut N^m. Étant donné que $p = m/N$, nous transformons facilement (38.324) en

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{38.325}$$

Une variable aléatoire de loi binomiale étant une somme de variables aléatoires de Bernoulli indépendantes, l'espérance [9] et la variance [10] s'obtiennent en sommant les espérances et variances termes à terme :

$$E(X) = np \tag{38.326}$$

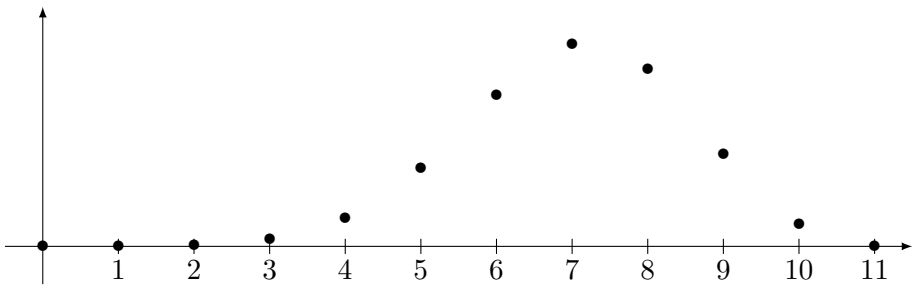
et

$$\operatorname{Var}(X) = \sum_{i=1}^{n} \operatorname{Var}(X_i) = np(1-p) \tag{38.327}$$

en vertu de (38.321).

La loi binomiale lorsque $p = 0.7$ et $n = 10$.

9. Valable même sans indépendance, proposition 38.25
10. Lemme 38.23.



38.5.3 Loi multinomiale

La loi multinomiale $\mathscr{M}(n;k;p_1,\dots,p_k)$ consiste à effectuer n épreuves d'une démarche aléatoire qui peut avoir k issues différentes avec probabilités $p_1,\dots,p_k$. Les variables aléatoires multinomiales sont N_i avec les contraintes

$$\sum_{i=1}^{k} N_i = n \tag{38.328a}$$

$$\sum_{i=1}^{k} p_i = 1. \tag{38.328b}$$

La fonction de probabilité multinomiale est

$$P(N_1 = n_1,\dots,N_k = n_k) = \frac{n!}{n_1!\dots n_k!}p_1^{n_1}\dots p_k^k. \tag{38.329}$$

Chacune des N_i est une binomiale de probabilité p_i.

38.5.4 Loi géométrique

Soit (X_n) une suite indépendante et identiquement distribuée de lois de Bernoulli de paramètre p. Alors la variable aléatoire

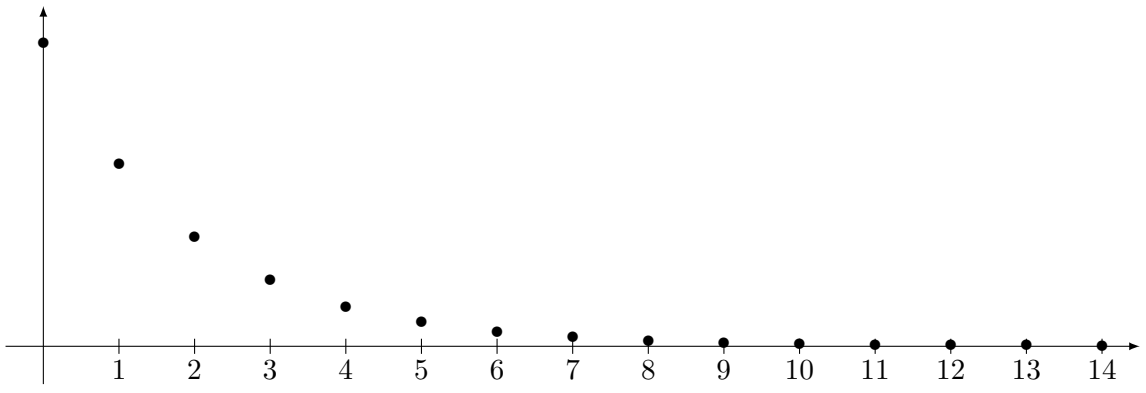
$$Z = \inf\{n \geqslant 1 \text{ tel que } X_n = 1\} \tag{38.330}$$

est une loi géométrique de paramètre p.

La loi géométrique compte donc le nombre d'expériences de Bernoulli à effectuer avant que le premier succès soit au rendez-vous. Nous avons

$$P(Z = k) = P(X_k = 1)P(X_1,\dots,X_{k-1} = 0) = p(1-p)^{k-1} \tag{38.331}$$

La loi géométrique de paramètre $p = 0.2$.



Note : si p est trop grand, on ne voit vite plus rien parce que la probabilité d'attendre longtemps est vite très faible.

38.5.5 Loi de Poisson

Une variable aléatoire Z suit une **loi de Poisson** de paramètre λ, notée $\mathscr{P}(\lambda)$ si

$$P(Z = k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{38.332}$$

pour tout $k \in \mathbb{N}$.

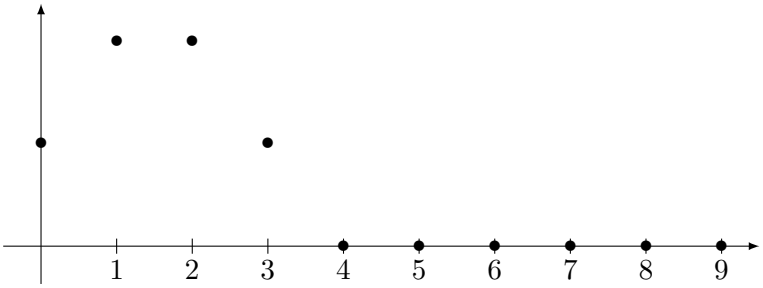
La loi de Poisson est une loi de probabilité discrète qui décrit le comportement du nombre d'évènements se produisant dans un laps de temps fixé, si ces évènements se produisent avec une fréquence moyenne connue et indépendamment du temps écoulé depuis l'évènement précédent.

Si un événement se produit en moyenne p fois par seconde, la probabilité d'observer l'événement k fois durant n secondes est donnée par $P(Z = k)$ où Z est une loi de Poisson de paramètre $\lambda = pn$.

Théorème 38.96 (wikipedia)**.**
L'espérance et la variance d'une variable aléatoire de Poisson sont λ :

$$E(X) = \lambda \tag{38.333a}$$
$$\mathrm{Var}(X) = \lambda. \tag{38.333b}$$

La loi de Poisson de paramètre $\lambda = 2$.



38.5.6 Loi exponentielle

La loi de exponentielle représente le temps qu'il faut attendre pour qu'une particule se désintègre si elle a en permanence une probabilité [11] λdt de se désintégrer entre t et $t + dt$. L'espérance est donc $1/\lambda$.

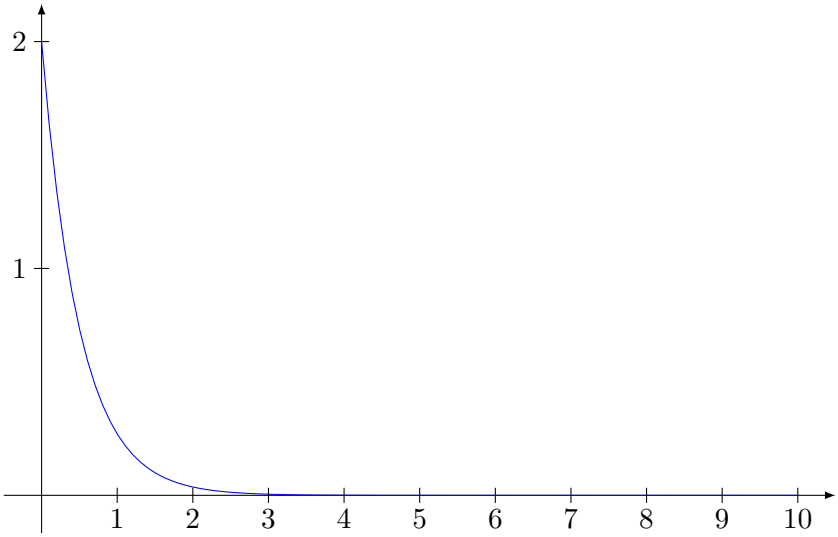
Il se passe donc en moyenne λ événements par seconde. La proposition 38.103 nous montrera que le nombre d'événements se produisant en une seconde suit une loi de Poisson de paramètre λ.

Plus formellement, la loi exponentielle de paramètre λ, notée $\mathscr{E}(\lambda)$ est la loi de densité

$$f_X : x \mapsto \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geqslant 0 \\ 0 & \text{sinon.} \end{cases} \tag{38.334}$$

Densité de la loi exponentielle de paramètre $\lambda = 2$.

11. Étant donné que λ n'est pas limité à 1, en réalité ce n'est pas une probabilité. Je suis preneur d'une bonne interprétation physique de ce paramètre.



Proposition 38.97.
Si $X \sim \mathscr{E}(\lambda)$, alors la fonction de répartition de X est donnée par

$$F(x) = P(X \leqslant x) = \begin{cases} 1 - e^{-\lambda x} & si\ x \geqslant 0 \\ 0 & sinon. \end{cases} \tag{38.335}$$

La fonction caractéristique est donnée par

$$E(e^{itX}) = \frac{\lambda}{\lambda - it}. \tag{38.336}$$

L'espérance et la variance sont données par

$$\begin{aligned} E(X) &= \frac{1}{\lambda} \\ \mathrm{Var}(X) &= \frac{1}{\lambda^2}. \end{aligned} \tag{38.337}$$

Démonstration. Pour la fonction caractéristique,

$$E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} \lambda e^{-\lambda x} \mathbb{1}_{[0,\infty[}(x) dx \tag{38.338a}$$

$$= \lambda \int_0^{\infty} e^{x(-\lambda+it)} dx \tag{38.338b}$$

$$= \lim_{A\to\infty} \left[\frac{e^{x(-\lambda+it)}}{it-\lambda} \right]_{x=0}^{x=A} \tag{38.338c}$$

$$= \lim_{A\to\infty} \frac{e^{A(it-\lambda)}}{it-\lambda} - \frac{1}{it-\lambda}. \tag{38.338d}$$

Le premier terme est nul parce que si on prend la norme,

$$\left| \frac{e^{A(-\lambda+it)}}{-\lambda+it} \right| = \frac{e^{-\lambda A}}{|it-\lambda|} \to 0. \tag{38.339}$$

En ce qui concerne l'espérance nous faisons le calcul suivant :

$$E(X) = \int_{\mathbb{R}} x f_X(x) dx = \lambda \int_{\mathbb{R}^+} x e^{-\lambda x} dx = \frac{1}{\lambda}. \tag{38.340}$$

Pour la variance, nous utilisons la formule (38.58). Nous avons

$$E(X^2) = \int_{\mathbb{R}} x^2 f_X(x) = \int_0^{\infty} x^2 \lambda e^{-\lambda x} = \frac{2}{\lambda^2}. \tag{38.341}$$

Donc $\mathrm{Var}(X) = E(X^2) - E(X)^2 = \frac{1}{\lambda^2}$. □

La loi exponentielle est une loi **sans mémoire** en ce sens que

$$P(X > x + y | X > y) = P(X > x). \tag{38.342}$$

En effet nous utilisons la règle de la probabilité conditionnelle

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{38.343}$$

Ici,

$$P(X > x + y | X > y) = \frac{P(X > x + y)}{P(X > y)} = e^{-\lambda x}. \tag{38.344}$$

La proposition suivante montre que la loi exponentielle est à peu près la seule à être sans mémoire. D'où son importance dans l'étude des machines dont les pièces ne subissent pas d'usure.

Proposition 38.98.
Soit X, une variable aléatoire admettant une densité continue f par rapport à la mesure de Lebesgue. Si elle est sans mémoire, alors elle est exponentielle.

Démonstration. Nous posons $\varphi(x) = P(X \geqslant x)$. Cela est la fonction de répartition de X (à part que cette dernière est $1 - \varphi(x)$ mais c'est pas grave), et est donnée par

$$\varphi(x) = 1 - \int_0^x f(t)dt. \tag{38.345}$$

Cette dernière intégrale vérifie les hypothèses du théorème 38.98, de telle sorte que φ soit une fonction dérivable et $\varphi'(x) = f(x)$.

D'autre part en utilisant la définition de la probabilité conditionnelle la propriété de ne pas avoir de mémoire donne

$$\varphi(x + y) = \varphi(x)\varphi(y) \tag{38.346}$$

et de plus $\varphi(0) = 0$. Calculons la dérivée de φ :

$$\varphi'(x) = \lim_{\epsilon \to 0} \frac{\varphi(x + \epsilon) - \varphi(x)}{\epsilon} = \varphi(x) \lim_{\epsilon \to 0} \frac{\varphi(\epsilon) - 1}{\epsilon} = \varphi(x)\varphi'(0). \tag{38.347}$$

Donc φ vérifie l'équation différentielle de l'exponentielle.

□

Exemple 38.99
Une machine a une durée de vie représentée par une variable aléatoire suivant une loi de Poisson de paramètre λ. Soit T_y la variable aléatoire qui représente la temps de vie restant sachant que la machine a déjà vécu un temps y. Nous voulons trouver la fonction de répartition de T_y. Nous avons

$$P(T_y > x) = P(X > x + y | X > y) = P(X > x) = e^{-\lambda x}. \tag{38.348}$$

Dans ce cas, la loi de T_y ne dépend pas de y. Cela signifie que la machine ne vieilli pas et surtout que le modèle n'est pas réaliste. △

Proposition 38.100.
Si $X \sim \mathscr{E}(\lambda)$ et $Y \sim \mathscr{E}(\mu)$ sont indépendantes, alors

(1) $P(X < Y) = \frac{\lambda}{\lambda+\mu}$
(2) $P(X > Y) = \frac{\mu}{\lambda+\mu}$
(3) $P(X = Y) = 0$
(4) $\min(X, Y) \sim \mathscr{E}(\lambda + \mu)$.

De plus les variables aléatoires exponentielles ont une propriété d'absence de mémoire :

$$P(X > t + s|X > s) = P(X > t) = e^{-\lambda t}. \tag{38.349}$$

Démonstration. Étant donné que X et Y sont indépendantes, la densité conjointe est le produit des densités (38.20). Nous avons donc

$$P(X > Y) = \int_D \lambda e^{-\lambda x} \mu e^{-\mu y} dxdy \tag{38.350}$$

où D est le domaine $D = \{(x, y) \in \mathbb{R}^2 \text{ tel que } x, y > 0, x > y\}$. Nous avons donc

$$P(X > Y) = \lambda\mu \int_0^\infty dx \int_0^x dy e^{-\lambda x} e^{-\mu y} = \frac{\mu}{\lambda + \mu}. \tag{38.351}$$

```
sage: var('a,b')
(a, b)
sage: f(x,y)=exp(-a*x)*exp(-b*y)
sage: assume(a>0)
sage: assume(b>0)
sage: a*b*f.integrate(y,0,x).integrate(x,0,oo)
(x, y) |--> a*b/(a^2 + a*b)
```

Pour trouver la loi de $\min(X, Y)$, nous écrivons

$$\begin{aligned}
P\big(\min(X, Y) > t\big) &= P(X > t, Y > t) && && (38.352a)\\
&= P(X > t)P(Y > t) && \text{par indépendance} && (38.352b)\\
&= \big(1 - F_X(t)\big)\big(1 - F_Y(t)\big) && && (38.352c)\\
&= e^{-(\lambda+\mu)t} \mathbb{1}_{[0,\infty[}(t) && && (38.352d)\\
&= 1 - F_Z(t) && && (38.352e)
\end{aligned}$$

où $Z \sim \mathscr{E}(\lambda + \mu)$. □

38.5.7 Approximation de la binomiale par une Poisson

Proposition 38.101.
Soit (X_n) une suite de variables aléatoires avec $X_n \sim \mathscr{B}(n, p_n)$ telle que np_n converge vers une constante $\lambda > 0$. Alors $X_n \xrightarrow{\mathscr{L}} \mathscr{P}(\lambda)$.

Démonstration. Commençons par écrire la loi binomiale sous une forme plus adaptée au passage à la limite :

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n_k} = \frac{n(n - 1) \dots (n - k + 1)}{k!} p^k (1 - p)^{n-k}. \tag{38.353}$$

Le produit au numérateur contient k termes dans lesquels nous mettons n en évidence. Nous trouvons

$$P(X = k) = \frac{(np)^k \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{k-1}{n}\right)}{k!} p^k (n - p)^{n-k}. \tag{38.354}$$

Lorsque nous passons à la limite, tous les facteurs du type $1-l/n$ tendent vers 1 ainsi que $(1-p_n)^{-k}$. Les facteurs dont la limite n'est pas 1 sont donc

$$P(X_n = k) \simeq \frac{(np_n)^k}{k!} (1 - p_n)^k. \tag{38.355}$$

Nous avons

$$\lim_{n\to\infty} (1 - p_n)^n = \lim_{n\to\infty} \left(1 - \frac{np_n}{n}\right)^n = e^{-\lambda}. \tag{38.356}$$

La thèse est alors obtenue en remettant les morceaux ensemble. □

Exemple 38.102
Considérons un serveur informatique qui reçoit des requêtes. Toutes les 10^{-3} s il reçoit une requête avec une probabilité $p = 0.05$. La variable aléatoire qui consiste à donner le nombre de requêtes effectivement effectuées en une seconde suit une loi binomiale $\mathscr{P}(1000, p)$.

Déterminons la probabilité que le serveur reçoive 20 requêtes en une seconde. Nous approximons $\mathscr{B}(1000, 0.05)$ par $\mathscr{P}(50)$, et la réponse est

$$e^{-50}\frac{50^{20}}{20!} \simeq 7 \cdot 10^{-7}. \tag{38.357}$$

△

38.5.8 Loi de Poisson et loi exponentielle

Soient $X_1, \dots, X_n$ des variables aléatoires réelles indépendantes de loi exponentielle de paramètre λ. En utilisant le produit de convolution, nous pouvons trouver la fonction de densité de la somme (voir point 38.2.3). Commençons avec deux variables aléatoires X et Y. Les densités sont

$$f_X(x) = \mathbb{1}_{[x\geqslant 0]}\lambda e^{-\lambda x} \tag{38.358a}$$
$$f_Y(y) = \mathbb{1}_{[y\geqslant 0]}\lambda e^{-\lambda y}, \tag{38.358b}$$

et la densité conjointe est alors

$$f_{X+Y}(x) = \int_{\mathbb{R}} \mathbb{1}_{[x-t\geqslant 0]}\lambda e^{-\lambda(x-t)}\mathbb{1}_{[t\geqslant 0]}\lambda e^{-\lambda t}dt \tag{38.359a}$$
$$= \lambda^2 e^{-\lambda x}\int_0^x 1\,dt \tag{38.359b}$$
$$= x\lambda^2 e^{-\lambda x}. \tag{38.359c}$$

Par récurrence si $S = X_1 + \cdots + X_n$ nous trouvons

$$f_S(x) = x^{n-1}\lambda^n e^{-\lambda x}. \tag{38.360}$$

Proposition 38.103 ([424]).
Soit $(T_k)_{k\in\mathbb{N}}$ une suite de variables aléatoires indépendantes de loi $\mathscr{E}(\lambda)$. Nous considérons la variable aléatoire $S_n = \sum_{i=1}^n X_i$ et pour chaque $t \in \mathbb{R}^+$ nous considérons

$$N_t = \max\{n \geqslant 1 \text{ tel que } \sum_{i=1}^n T_i \leqslant t\}. \tag{38.361}$$

Alors $N_t \sim \mathscr{P}(\lambda t)$.

Démonstration. Ce que nous devons calculer est

$$P(N_t = k) = P(S_n \leqslant t \leqslant S_{n+1}). \tag{38.362}$$

Nous introduisons la variable aléatoire $V_{n=1} = (X_1, \dots, X_{n+1})$ ainsi que l'ensemble

$$A_{n+1} = \{x \in \mathbb{R}^{n+1} \text{ tel que } x_1 + \cdots + x_n \leqslant t \leqslant x_1 + \cdots + x_n + x_{n+1}\}. \tag{38.363}$$

Le problème est donc de calculer

$$P(S_n \leqslant t \leqslant S_{n+1}) = P(V_{n+1} \in A_{n+1}^+) = \int_{A_{n+1}^+} f_{n+1}(x)dx \tag{38.364}$$

où A_{n+1}^+ est la partie de A_{n+1} dans laquelle $x_i \geqslant 0$ pour tout i et f_{n+1} est la fonction de densité conjointe des variables aléatoires X_i. Nous effectuons le changement de variables

$$s_k = \sum_{i \leqslant k} x_i \tag{38.365a}$$

$$x_k = s_k - s_{k-1} \tag{38.365b}$$

dont le déterminant vaut 1. D'autre part par indépendance des variables aléatoires X_i, la fonction de partition jointe f_{n+1} s'exprime sous la forme

$$f_{n+1}(x_1, \ldots, x_{n+1}) = f_{X_1}(x_1) \ldots f_{X_{n+1}}(x_{n+1}) \tag{38.366a}$$

$$= \lambda^{n+1} e^{-\lambda(x_1 + \cdots + x_{n+1})} \tag{38.366b}$$

$$= \lambda^{n+1} e^{-\lambda s_{n+1}}. \tag{38.366c}$$

En ce qui concerne les bornes de l'intégrale dans les variables s_i, nous voulons que tous les x_i soient positifs, par conséquent $s_1 \geqslant 0$ et ensuite l'équation $x_k = s_k - s_{k-1}$ demande $s_k \geqslant s_{k-1}$. Les bornes sont donc données par l'ensemble

$$0 \leqslant s_1 \leqslant s_2 \leqslant \ldots \leqslant s_n \leqslant t \leqslant s_{n+1}, \tag{38.367}$$

c'est à dire $B_n \times]t, \infty[$ où

$$B_n = \{(s_1, \ldots, s_n) \text{ tel que } 0 \leqslant s_1 \leqslant s_2 \leqslant \ldots \leqslant s_n \leqslant t\}. \tag{38.368}$$

Le théorème de Fubini nous permet de décomposer l'intégrale :

$$P(S_n \leqslant t \leqslant S_{n+1}) = \int_{B_n \times]t,\infty[} \lambda^{n+1} e^{-\lambda s_{n+1}} ds_1 \ldots ds_{n+1} \tag{38.369a}$$

$$= \lambda^{n+1} \left(\int_{B_n} ds_1 \ldots ds_n \right) \underbrace{\left(\int_t^\infty e^{-\lambda s_{n+1}} ds_{n+1} \right)}_{= \lambda^{-1} e^{-\lambda t}} \tag{38.369b}$$

$$= \lambda^n e^{-\lambda t} \operatorname{Vol}(B_n) \tag{38.369c}$$

où $\operatorname{Vol}(B_n)$ est le volume de B_n qui reste à calculer. L'ensemble $C^n = [0, t]^n$ se décompose en cellules disjointes (à ensemble de mesure nulle près) de la forme

$$C_\sigma = \{0 \leqslant s_{\sigma(1)} \leqslant s_{\sigma(2)} \leqslant \ldots \leqslant s_{\sigma(n)} \leqslant t\} \tag{38.370}$$

pour chaque permutation $\sigma \in S_n$. Il y a exactement $n!$ telles cellules dans C^n. Par conséquent

$$t^n = \operatorname{Vol}(C^n) = n! \operatorname{Vol}(C_\sigma) = n! \operatorname{Vol}(B_n) \tag{38.371}$$

et $\operatorname{Vol}(B_n) = \frac{t^n}{n!}$. Finalement nous avons

$$P(n_t = n) = P(S_n \leqslant t \leqslant S_{n+1}) = \frac{(\lambda t)^n}{n!} e^{-(\lambda t)}. \tag{38.372}$$

□

38.5.9 Loi normale

La loi normale de paramètres m et $\sigma > 0$, notée $\mathcal{N}(m, \sigma^2)$ est la loi donnée par la densité

$$\gamma_{m,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{x - m}{\sigma} \right)^2 \right]. \tag{38.373}$$

Proposition 38.104.
Si la variable aléatoire réelle X suit une loi normale $\mathcal{N}(m,\sigma^2)$, alors nous avons $E(X)=m$ et $\text{Var}(X)=\sigma^2$.

Démonstration. L'espérance d'une variable aléatoire se calcule à partir de la formule (38.216) :

$$E(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}}x\exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right]dx \tag{38.374a}$$

$$=\frac{1}{\sigma\sqrt{2\pi}}\sigma\int_{\mathbb{R}}(\sigma u+m)e^{-u^2/2} \tag{38.374b}$$

où nous avons effectué le changement de variable $u=(x-m)/\sigma$. Nous utilisons ensuite l'intégrale remarquable

$$\int_{\mathbb{R}}e^{-u^2/2}du=\sqrt{2\pi}. \tag{38.375}$$

En ce qui concerne la variance, nous avons le même genre de calculs. □

La **loi normale réduite** est la densité

$$\gamma(x)=\gamma_{0,1}(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}. \tag{38.376}$$

La variable aléatoire X suit la loi $\mathcal{N}(m,\sigma^2)$ si et seulement si la variable aléatoire $Z=\frac{X-m}{\sigma}$ suit la loi normale réduite $\mathcal{N}(0,1)$.

Proposition 38.105.
La fonction caractéristique de la distribution normale $\mathcal{N}(m,\sigma^2)$ est

$$\Phi_{\mathcal{N}(m,\sigma^2)}(t)=\exp\left(itm-\frac{\sigma^2t^2}{2}\right). \tag{38.377}$$

Démonstration. En suivant la formule (38.188), l'intégrale à calculer est

$$\Phi_{\mathcal{N}(m,\sigma^2)}(t)=\frac{1}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}}e^{itx}e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}dx. \tag{38.378}$$

Nous reconnaissons une transformée de Fourier. Afin de la calculer sans encombres, nous passons par les fonctions intermédiaires suivantes :

$$\begin{aligned}g(x)&=e^{-\frac{1}{2}x^2}\\h(x)&=g\left(\frac{x}{\sigma}\right)\\k(x)&=h(x-m).\end{aligned} \tag{38.379}$$

La fonction caractéristique que nous cherchons est $\frac{1}{\sigma\sqrt{2\pi}}\hat{k}(t)$. Les formules liées à la transformée de Fourier nous donnent

$$\hat{k}(t)=\hat{h}(t)e^{itm} \tag{38.380a}$$

$$\hat{h}(t)=\sigma\hat{g}(\sigma t) \tag{38.380b}$$

$$\hat{g}(t)=\int_{\mathbb{R}}e^{-itx}e^{-\frac{1}{2}x^2}dx=\sqrt{2\pi}e^{-t^2/2}. \tag{38.380c}$$

Attention : l'intégrale à calculer est une transformée de Fourier *inverse*, d'où la formule (38.380a) qui a un signe de différence avec la formule usuelle. En recombinant toutes ces expressions nous trouvons

$$\Phi_{\mathcal{N}(m,\sigma^2)}=e^{-\sigma^2t^2/2}e^{itm}, \tag{38.381}$$

ce qu'il nous fallait. □

Exemple 38.106
Une espérance qui sert de temps en temps est celle de $X = e^{\beta Z}$ lorsque $Z \sim \mathcal{N}(0,1)$. Elle se calcule en remarquant que $x^2 - 2\beta x = (x-\beta)^2 - \beta^2$, donc

$$E(e^{\beta Z}) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{\beta x} e^{-x^2/2} dx \tag{38.382a}$$
$$= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{1}{2}(x-\beta)^2} e^{\beta^2/2} \tag{38.382b}$$
$$= \frac{e^{\beta^2/2}}{\sqrt{2\pi}}\int_{e^{-y^2/2}} dy \tag{38.382c}$$
$$= e^{\beta^2/2}. \tag{38.382d}$$

△

38.5.10 Vecteurs gaussiens

Source : [195, 425].

Définition 38.107.
*Un vecteur aléatoire $X\colon \Omega \to \mathbb{R}^d$ est un **vecteur gaussien** si toutes les combinaisons linéaires de ses composantes sont des variables aléatoires normales. En d'autres termes, X est un vecteur gaussien si pour tout vecteur u, la variable aléatoire $u \cdot X$ est gaussienne.*

Le vecteur moyenne d'un vecteur gaussien est $E(X) = \big(E(X_1), \dots, E(X_n)\big)$ et sa matrice de **variance-covariance** est la matrice

$$K_X = \mathrm{Var}(X) = E\Big[\big(X - E(X)\big) \otimes \big(X - E(X)\big)\Big] \tag{38.383}$$

où l'opération $\otimes$ est celle introduite autour de l'équation (12.213). Cela n'est rien d'autre que la matrice de covariance de la variable aléatoire $X\colon \Omega \to \mathbb{R}^d$.

Lemme 38.108.
Si les variables aléatoires réelles $X_1, \dots, X_n$ sont des variables aléatoires réelles gaussiennes indépendantes, alors le vecteur $X = (X_1, \dots, X_n)$ est un vecteur gaussien.

Démonstration. Nous devons montrer que si X et Y sont des variables aléatoires gaussiennes indépendantes, alors $X+Y$ est encore gaussienne. L'indépendance nous assure les égalités suivantes pour la fonction caractéristique :

$$\Phi_{X+Y}(t) = E\big(e^{it(X-Y)}\big) = E\big(e^{itX}\big)E\big(e^{itY}\big). \tag{38.384}$$

Dans le cas où X et Y sont gaussiens nous trouvons

$$\Phi_{X+Y}(t) = \exp\left(im_1 t - \frac{\sigma_1^2 t^2}{2}\right)\exp\left(im_2 t - \frac{\sigma_2^2 t^2}{2}\right) = \exp\left(i(m_1+m_2)t - \frac{(\sigma_1^2+\sigma_2^2)t^2}{2}\right). \tag{38.385}$$

Étant donné que la loi d'une variable aléatoire est entièrement déterminée par sa fonction caractéristique (théorème 38.58), nous déduisons que $X+Y$ est une normale de moyenne m_1+m_2 et de variance $\sigma = \sigma_1^2 + \sigma_2^2$. □

Proposition 38.109.
La fonction caractéristique d'un vecteur gaussien est donnée par

$$\Phi_X(u) = \exp\left(iu \cdot E(X) - \frac{1}{2}u \cdot K_X u\right) \tag{38.386}$$

où K_X est la matrice de covariance de X.

Démonstration. Nous considérons la variable aléatoire réelle gaussienne $u \cdot X$. Son espérance $m = E(u \cdot X) = u \cdot E(X)$. Nous commençons par établir la formule suivante :

$$u^t K_X u = u \cdot K_X u = E\Big[\big([X - E(X)] \cdot u\big)^2\Big]. \tag{38.387}$$

Utilisant la linéarité de l'espérance,

$$\sum_{kl} E(A_{kl})u_k u_l = \sum_{kl} E(A_{kl}u_k u_l), \tag{38.388}$$

nous trouvons

$$\begin{aligned}
K_X(u,u) &= E\Big([X - E(X)] \otimes [X - E(X)]\Big)(u,u) && \text{(38.389a)}\\
&= E\Big(\big([X - E(X)] \otimes [X - E(X)]\big)(u,u)\Big) && \text{(38.389b)}\\
&= E\Big(\big([X - E(X)] \cdot u\big)\big([X - E(X)] \cdot u\big)\Big) && \text{(38.389c)}\\
&= E\left(\big([X - E(X)] \cdot u\big)^2\right). && \text{(38.389d)}
\end{aligned}$$

Par la linéarité du produit scalaire et de l'espérance,

$$[X - E(X)] \cdot u = u \cdot X - E(u \cdot X), \tag{38.390}$$

ce qui nous ramène à la variable aléatoire $u \cdot X$. Nous avons alors

$$K_X(u,u) = E\left[\big([X - E(X)] \cdot u\big)^2\right] = \mathrm{Var}(u \cdot X). \tag{38.391a}$$

Nous avons donc obtenu une forme pour la variance de la variable aléatoire $u \cdot X$. Étant donné que $u \cdot X$ est gaussienne de moyenne $m = u \cdot E(X)$ et de variance $\sigma^2 = K_X(u,u)$, nous avons

$$\Phi_{u \cdot X}(t) = \exp\big(itm - \frac{1}{2}\sigma^2 t^2\big). \tag{38.392}$$

Par ailleurs nous avons $\Phi_X(u) = \Phi_{u \cdot X}(1)$ parce que

$$\Phi_X(u) = E\big(e^{iu \cdot X}\big) = \Phi_{u \cdot X}(1). \tag{38.393}$$

En utilisant la forme (38.392) pour $\Phi_{u \cdot X}$ nous trouvons

$$\Phi_X(u) = \Phi_{u \cdot X}(1) = \exp\big(im - \frac{1}{2}\sigma^2\big) = \exp\Big(iE(u \cdot X) - \frac{1}{2}u \cdot KXu\Big). \tag{38.394}$$

□

Théorème 38.110.
Soit $X = (X_1, \dots, X_d)$, un vecteur gaussien. Les composantes sont indépendantes si et seulement si elles sont non corrélées.

Démonstration. Nous savons que les variables aléatoires indépendantes sont non corrélées. Nous devons donc surtout prouver le contraire. Le fait que les variables aléatoires X_i soient non corrélées signifie que la matrice de covariance est

$$K_X = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2 \end{pmatrix} \tag{38.395}$$

où $\sigma_k^2 = \text{Var}(X_k)$. Notons $m_k = E(X_k)$. Si $u \in \mathbb{R}^d$, nous avons en vertu de la proposition 38.109 que

$$\Phi_X(u) = \exp\Big(i(u_1m_1 + \cdots + u_dm_d) - \frac{1}{2}(\sigma_1^2u_1^2 + \cdots + \sigma_d^2u_d^2)\Big) \tag{38.396a}$$
$$= \exp\big(iu_1m_1 - \frac{1}{2}\sigma_1^2u_1^2\big)\ldots\exp\big(iu_dm_d - \frac{1}{2}\sigma_d^2u_d^2\big) \tag{38.396b}$$
$$= \Phi_{X_1}(u_1)\ldots\Phi_{X_d}(u_d). \tag{38.396c}$$

Les variables aléatoires X_i sont donc indépendantes parce que la fonction caractéristique se factorise. □

Exemple 38.111
Nous donnons à présent un exemple de deux variables aléatoires gaussiennes qui ne forment pas un vecteur gaussien. Pour ce faire nous devons chercher des variables aléatoires non indépendantes. Soit $Y \sim \mathcal{N}(0,1)$ et ϵ une variable aléatoire (indépendante de Y) donnée par $P(\epsilon = -1) = \frac{1}{2}$, $P(\epsilon = 1) = \frac{1}{2}$. Nous considérons le vecteur $(Y, \epsilon Y)$.

D'abord montrons que ϵY est une variable aléatoire gaussienne. Soit A un borélien de $\mathbb{R}$. Nous avons

$$P(\epsilon Y \in A) = P(\epsilon = -1, Y \in A) + P(\epsilon = -1, -Y \in A). \tag{38.397}$$

Par indépendance et par symétrie de Y [12] nous trouvons

$$P(\epsilon Y \in A) = P(\epsilon = 1)P(Y \in A) + P(\epsilon = -1)P(-Y \in A) \tag{38.398a}$$
$$= \frac{1}{2}P(Y \in A) + \frac{1}{2}P(-Y \in A) \tag{38.398b}$$
$$= P(Y \in A). \tag{38.398c}$$

Nous avons donc $Y \sim \epsilon Y$, et donc ϵY est gaussienne.

En ce qui concerne la covariance, nous savons que $E(Y) = E(\epsilon) = 0$, donc

$$\text{Cov}(Y, \epsilon Y) = E(Y \cdot \epsilon Y) = E(\epsilon Y^2) = E(\epsilon)E(Y^2) = 0. \tag{38.399}$$

Note : $E(Y^2) = 1$.

Les variables aléatoires Y et ϵY ne sont pas indépendantes. En effet si elles l'étaient, Y serait aussi indépendante de $(\epsilon Y)^2 = Y^2$, alors que Y et Y^2 ne sont pas indépendantes. Donc $X = (Y, \epsilon Y)$ n'est pas un vecteur gaussien. △

Théorème 38.112 ([195]).
Soit $m \in \mathbb{R}^d$ et $K \in \mathbb{M}(d, \mathbb{R})$ une matrice symétrique et positive. Alors il existe un vecteur gaussien de moyenne m et de matrice de covariance K.

Démonstration. Nous effectuons la preuve avec $m = 0$. Nous choisissons l'espace probabilisé $(\Omega, \mathcal{A}) = (\mathbb{R}^r, \mathcal{B}\text{or}(\mathbb{R}^r))$ où r est le rang de K muni de la probabilité de densité

$$\gamma(u) = \frac{1}{(2\pi)^{r/2}}\exp\big(-\frac{1}{2}\|u\|^2\big). \tag{38.400}$$

Nous considérons la variable aléatoire

$$\begin{aligned} Y_0 \colon \Omega &\to \mathbb{R}^r \\ u &\mapsto u. \end{aligned} \tag{38.401}$$

C'est une variable aléatoire gaussienne de loi $P_{Y_0} = \gamma\lambda_d$ où λ_d est la mesure de Lebesgue sur $\mathbb{R}^d$. Sa densité (38.400) s'écrit comme le produit de r gaussiennes indépendantes ; sa matrice de covariance est donc $\mathbb{1}_{r\times r}$.

12. C'est à dire que $P(Y \in A) = P(Y \in -A) = P(-Y \in A)$.

Étant donné que K est symétrique et positive, il existe une matrice $d \times r$ telle que $K = AA^t$. Pour voir cela, remarquons qu'il existe une matrice $d \times d$ qui fait le travail. En effet K se diagonalise par une orthogonale (théorème 11.174) :

$$K = ADA^t = A\sqrt{D}\sqrt{D}A^t \tag{38.402}$$

où D est une matrice diagonale contenant $d-r$ zéros et $\sqrt{D}$ est la matrice que l'on imagine. Donc la matrice $L = A\sqrt{D}$ est une matrice telle que $LL^t = K$. Maintenant, étant donné que les $d-r$ dernières lignes de D sont vides, les $d-r$ dernières lignes de L n'ont pas d'importance et peuvent être choisies nulles, voire même ne pas exister. La matrice $A \in \mathbb{M}_{d\times r,\mathbb{R}}$ qui réalise $AA^t = K$ est la « troncature » de L à ses r premières lignes.

Nous considérons la variable aléatoire $Y = AY_0 \colon \Omega \to \mathbb{R}^d$. Étant donné que AY_0 est une transformation linéaire d'un vecteur gaussien, c'est un vecteur gaussien. Nous avons encore $m(Y) = 0$ et

$$\operatorname{Var}(Y) = E(Y \otimes Y) = E\big((AY_0) \otimes (AY_0)\big) = AE(Y_0 \otimes Y_0)A^t = AA^t = K \tag{38.403}$$

parce que $E(Y_0 \otimes Y_0) = \mathbb{1}$. Nous avons utilisé les formules du produit tensoriel introduit en (12.213), et en particulier la formule (12.216). □

Lorsqu'une matrice symétrique et positive K est donné, nous avons créé un vecteur gaussien de covariance K en créant $X = AY$ où Y est le vecteur gaussien « le plus simple » et où A est donné par $AA^t = K$. La proposition suivante montre l'inverse : un vecteur gaussien X peut se réduire au vecteur gaussien « le plus simple » en utilisant la transformation $Y = A^{-1}X$ où A est encore donnée par $AA^t = K$. Ce résultat nous permettra de voir les vecteurs gaussiens généraux comme des « changements de coordonnées » par rapport au vecteur gaussien simple $Y = (Y_1, \dots, Y_d)$ avec $Y_i \sim \mathcal{N}(0,1)$.

Proposition 38.113.
Soit $Y = (Y_1, \dots, Y_n)$ le vecteur gaussien formé des variables aléatoires indépendantes $Y_i \sim \mathcal{N}(0,1)$. Soit K une matrice symétrique et positive et la matrice A telle que $AA^t = K$. Alors le vecteur $X = AY$ est gaussien de covariance K.

Démonstration. Nous montrons que la covariance de $X = AY$ est donnée par K. Nous avons

$$\begin{aligned}
K_X &= E\big([X - E(X)] \otimes [X - E(X)]\big) && (38.404a)\\
&= E\big(A(Y - E(Y)) \otimes A(Y - E(Y))\big) && (38.404b)\\
&= E\Big(A\big([Y - E(Y)] \otimes [Y - E(Y)]\big)A^t\Big) && (38.404c)\\
&= E(AA^t) && (38.404d)\\
&= K_X. && (38.404e)
\end{aligned}$$

Nous avons utilisé le lemme 12.104 ainsi que le fait que

$$[Y - E(Y)] \otimes [Y - E(Y)] = K_Y = \mathbb{1}. \tag{38.405}$$

Notons que $E(Y) = 0$, mais cela ne joue pas ici. □

Théorème 38.114.
Un vecteur gaussien $X \colon \Omega \to \mathbb{R}^d$ possède une densité par rapport à la mesure de Lebesgue si et seulement si sa matrice de covariance est inversible. Dans ce cas nous avons la densité

$$f_X(x) = \frac{1}{(2\pi)^{d/2}} \frac{1}{\sqrt{|\det(K_X)|}} \exp\left(-\frac{1}{2}[X - E(X)] \cdot K_X^{-1}[X - E(X)]\right). \tag{38.406}$$

Démonstration. Nous supposons que $E(X) = 0$. En utilisant la proposition 38.113, nous posons $X = AY$ où Y est un vecteur gaussien $Y \sim \mathcal{N}(0, \mathbb{1})$ et $AA^t = K_X$. Si K n'est pas inversible, alors

A n'est pas inversible non plus. Notons $r < d$ le rang de A. Étant donné que $X = AY$, la variable aléatoire X prend presque sûrement ses valeurs dans l'image de A, c'est à dire dans un sous-espace de dimension $r < d$ de $\mathbb{R}^d$. Ce sous-espace est de mesure nulle pour la mesure de Lebesgue, mais de mesure 1 pour la mesure P_X. La mesure P_X ne peut donc pas avoir de densité par rapport à celle de Lebesgue.

Supposons maintenant que K_X soit inversible. La matrice A l'est aussi. Nous anticipons l'utilisation du théorème de changement de variable 16.232. Ici le changement de variable sera la transformation linéaire A dont le jacobien vaut

$$|\det(A^{-1})| = \frac{1}{\det A} = \frac{1}{\sqrt{|\det K_X|}}. \tag{38.407}$$

Soit γ la densité de Y. Nous posons

$$f_X(x) = \frac{1}{\sqrt{|\det K_X|}}\gamma(A^{-1}X) \tag{38.408}$$

où γ est le produit des densités de d gaussiennes usuelles $\mathcal{N}(0,1)$. Nous allons d'abord montrer que cette formule est bien la fonction (38.406) et ensuite que X admet f_X comme densité. Nous avons

$$\gamma(A^{-1}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|A^{-1}x\|^2\right), \tag{38.409}$$

et

$$\begin{aligned}
\|A^{-1}x\|^2 &= \langle A^{-1}x, A^{-1}x\rangle && \text{(38.410a)}\\
&= \langle (A^{-1})^t A^{-1}x, x\rangle && \text{(38.410b)}\\
&= \langle (AA^t)^{-1}x, x\rangle && \text{(38.410c)}\\
&= x \cdot K_X^{-1}x, && \text{(38.410d)}
\end{aligned}$$

ce qui nous donne bien la formule (38.406). Nous vérifions maintenant que f_X est bien une densité pour X. Soit B un borélien de $\mathbb{R}^d$. Nous avons d'abord

$$P(X \in B) = P(AY \in B) = P(Y \in A^{-1}B). \tag{38.411}$$

Ici nous avons utilisé le fait que A était bijectif. Nous avons ensuite

$$P(X \in B) = \int_{A^{-1}B} \gamma(t)dy = \int_B \gamma\big(A^{-1}x\big)|J_{A^{-1}}(x)|dx = \int_B f_X(x)dx. \tag{38.412}$$

C'est ici que nous avons utilisé le théorème de changement de variable 16.232. □

38.5.11 Variable aléatoire de Rademacher

Une variable aléatoire **de Rademacher** est une variable aléatoire de loi

$$\epsilon \sim \delta_0 - \delta_1, \tag{38.413}$$

sur l'ensemble $\Omega = \{0,1\}$. C'est la variable aléatoire qui prend valeur 1 ou -1 avec probabilité $\frac{1}{2}$.

Parmi les propriétés évidentes de cette variable aléatoire nous avons $E(\epsilon) = 0$ et $E(\epsilon^2) = 1$.

Proposition 38.115 (Inégalité de Khintchine[4])**.**
Soient $r_1, \dots, r_n$ des variables aléatoires indépendantes et identiquement distribuées de Rademacher et une combinaison linéaire

$$X = \sum_{i=1}^n a_i r_i \tag{38.414}$$

avec $a_i \in \mathbb{R}$. Alors

$$\|X\|_2 \leqslant \sqrt{e}E\big(|X|\big). \tag{38.415}$$

Démonstration. Pour rappel la définition est que

$$\|X\|_2 = \sqrt{E\big(|X|^2\big)}. \tag{38.416}$$

Vu que c'est de la quantité $E\big(|X|\big)$ que nous voulons parler, nous notons l'inégalité

$$E\big(|X|\big) = \int_\Omega \big|X(\omega)\big|dP(\omega) \geqslant \Big|\int_\Omega X(\omega)dP(\omega)\Big| = \big|E(X)\big|. \tag{38.417}$$

Nous supposons que $\sum_{j=1}^n a_j^2 = 1$; sinon nous multiplions les a_j parce qu'il faut pour l'avoir et ce facteur sortira des deux côtés de (38.415).

Nous allons passer par la variable aléatoire intermédiaire

$$Y = \prod_{j=1}^n (1 + ia_j r_j) \tag{38.418}$$

et pour presque tout $\omega \in \Omega$ nous avons

$$\begin{aligned}
|Y(\omega)| &= \prod_{j=1}^n |1 + ia_j r_j(\omega)| && (38.419a)\\
&= \prod_{j=1}^n \sqrt{1 + a_j^2 \underbrace{r_j(\omega)^2}_{=1}} && (38.419b)\\
&= \prod_{j=1}^n \sqrt{1 + a_j^2} && (38.419c)\\
&\leqslant \prod_{j=1}^n \sqrt{e^{a_j^2}} && (38.419d)\\
&= \sqrt{\prod_{j=1}^n e^{a_j^2}} && (38.419e)\\
&= \sqrt{e^{\sum a_j^2}} && (38.419f)\\
&= \sqrt{e}. && (38.419g)
\end{aligned}$$

Donc $\|Y\|_\infty \leqslant \sqrt{e}$ où $\|.\|_\infty$ est la norme supremum sur Ω. Cela nous permet de donner une première inégalité à propos de $E(|X|)$. D'abord

$$\big|E(XY)\big| = \Big|\int_\Omega X(\omega)Y(\omega)dP(\omega)\Big| \leqslant \int_\Omega \big|X(\omega)\big|\|Y\|_\infty dP(\omega) = \|Y\|_\infty E\big(|X|\big), \tag{38.420}$$

ensuite en remplaçant $\|Y\|_\infty$ par la majoration que nous venons de donner de $|Y(\omega)|$,

$$\sqrt{e}E\big(|X|\big) \geqslant \big|E(XY)\big|. \tag{38.421}$$

Il nous reste à prouver que $\big|E(XY)\big| \geqslant \|X\|_2$.

Pour ce faire nous commençons par noter que pour chaque j nous avons $E(r_j) = 0$ et $E(ia_j r_j^2) = ia_j$; en utilisant l'indépendance des r_j et le lemme 38.22 nous avons alors

$$\begin{aligned}
E(r_j Y) &= E\Big(r_j(1 + ia_j r_j)\prod_{k\neq j}(1 + ia_k r_k)\Big) && (38.422a)\\
&= E\big(r_j(1 + ia_j r_j)\big)\prod_{k\neq j} E(1 + ia_k r_k) && (38.422b)\\
&= ia_j \prod_{k\neq j}\big(1 + ia_k E(r_k)\big) && (38.422c)\\
&= ia_j. && (38.422d)
\end{aligned}$$

Par conséquent, en utilisant la proposition 38.25 dans le cas non indépendant,

$$E(XY) = \sum_{j=1}^{n} a_j E(r_j Y) = \sum_j a_j i a_j = i \sum_j a_j^2 = i. \tag{38.423}$$

Nous pouvons compléter l'équation (38.421) en

$$\sqrt{e} E\big(|X|\big) \geqslant \big|E(XY)\big| = 1, \tag{38.424}$$

et nous nous empressons de montrer que $\|X\|_2 = 1$. En effet $\|X\|_2 = \sqrt{E(|X|^2)}$ alors que

$$\begin{aligned}
\|X\|_2^2 &= E\big((\sum_i a_i r_i)^2\big) && \text{(38.425a)}\\
&= E\Big(\sum_k a_k^2 r_k^2 + 2\sum_{i\neq j} a_i a_j r_i r_j\Big) && \text{(38.425b)}\\
&= \sum_k a_k^2 + 2\sum_{i\neq j} a_i a_j E(r_i r_j) && \text{(38.425c)}\\
&= 1 && \text{(38.425d)}
\end{aligned}$$

□

38.5.12 Loi de Student

Définition 38.116.
La loi χ^2 à d degrés de liberté est la loi de la variable aléatoire $Y_1^2 + \cdots + y_n^2$ si les (Y_i) sont des variables aléatoires normales indépendantes centrées et réduites.

*La loi de **Student** à d degrés de liberté est la loi de la variable aléatoire*

$$\frac{X}{\sqrt{K/d}} \tag{38.426}$$

où $X \sim \mathcal{N}(0,1)$ et $K \sim \chi^2(d)$ sont des variables aléatoires indépendantes. Cette loi est notée $\mathscr{T}(d)$

Nous avons une illustration de la densité de la loi $\chi^2(10)$ à la figure 38.1.

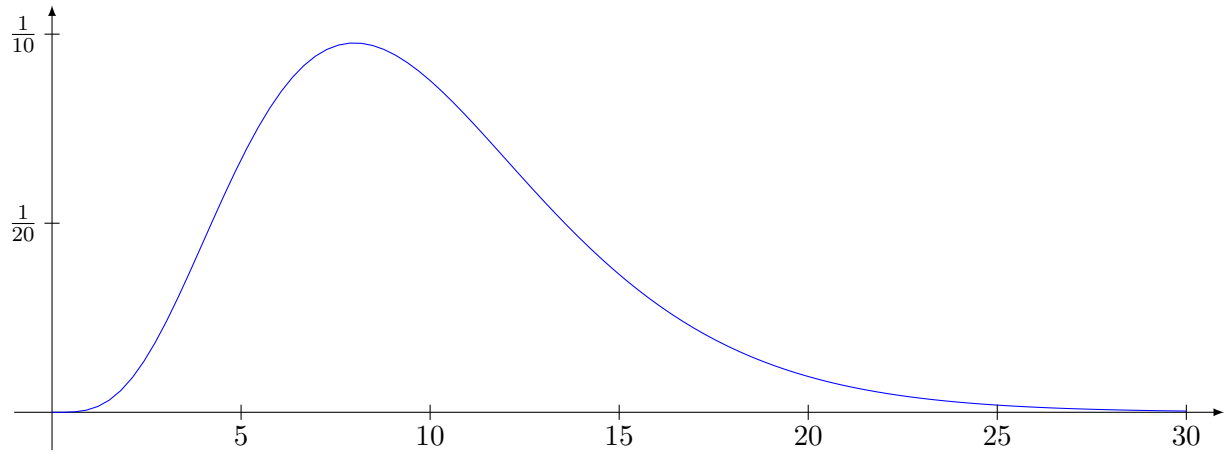


FIGURE 38.1 – La densité de $\chi^2(10)$.

L'importance de cette loi sera dans le théorème de Cochrane 39.15.

38.5.13 Indépendance, covariance et variance de somme

Si X et Y sont des variables aléatoires réelles, nous avons défini la covariance par (38.67) :

$$\mathrm{Cov}(X,Y) = E\Big[\big(X-E(X)\big)\big(Y-E(Y)\big)\Big] \tag{38.427}$$

Une clef est le lemme 38.22 qui dit que lorsque X et Y sont indépendantes, alors $E(XY) = E(X)E(Y)$. Nous avons les liens suivants.

(1) Si X et Y sont indépendantes, alors $\mathrm{Cov}(X,Y) = 0$.

(2) La réciproque n'est pas vraie par l'exemple 38.24.

(3) Si Z est un vecteur gaussien, les composantes Z_i sont indépendantes si et seulement si la matrice de covariance est diagonale, c'est à dire que les Z_i sont deux à deux non corrélées ; c'est le théorème 38.110.

(4) En ce qui concerne la variance d'une somme,

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y). \tag{38.428}$$

Donc lorsque X et Y ne sont pas corrélées, la variance est sympa avec la somme. En particulier lorsqu'elles sont indépendantes, mais pas seulement.

38.6 Estimation des grands écarts

Si S_n est une somme de variables aléatoires de Bernoulli X_i indépendantes de probabilité p, l'espérance de S_n/n est p, et nous voudrions savoir quelle est la probabilité d'avoir un taux de succès un peu plus important :

$$P\big(\frac{S_n}{n} \geqslant p+\epsilon\big). \tag{38.429}$$

La loi des grands nombres nous permet de dire que ça ne va pas être très grand. En effet si nous posons

$$Y_i = X_i - p - \epsilon \tag{38.430}$$

et

$$Z_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{S_n}{n} - p - \epsilon, \tag{38.431}$$

la loi des grands nombres (théorème 38.83) nous indique que

$$Z_n \xrightarrow{p.s.} E(Y_1) = -\epsilon. \tag{38.432}$$

La proposition 38.68 sur le lien entre les types de convergence nous donne immédiatement la convergence en loi. C'est à dire que pour tout $\eta > 0$,

$$P\Big(|Z_n + \epsilon| \geqslant \eta\Big) \to 0. \tag{38.433}$$

En prenant $\eta = \frac{\epsilon}{2}$,

$$P(Z_n = 0) \leqslant P\big(|Z_n+\epsilon| \geqslant \frac{\epsilon}{2}\big) \to 0. \tag{38.434}$$

Tout cela montre que

$$\lim_{n\to\infty} P\big(\frac{S_n}{n} \geqslant p+\epsilon\big) = 0. \tag{38.435}$$

Autrement dit, la probabilité de tomber à une distance fixée de la moyenne tend vers zéro lorsque le nombre d'essais augmente. Rien d'étonnant.

Le théorème suivant nous indique la vitesse de convergence. Elle est exponentielle et le coefficient est donné en fonction de p et de ϵ.

Théorème 38.117 ([4]).
Soient des variables aléatoires $X_i \sim \mathscr{B}(p)$ et

$$S_n = \sum_{i=1}^{n} X_i \sim \mathscr{B}(n,p). \tag{38.436}$$

Pour $\epsilon \in]0, 1-p[$, nous définissons

$$h_+(\epsilon) = (p+\epsilon)\ln\left(\frac{p+\epsilon}{p}\right) + (1-p-\epsilon)\ln\left(\frac{1-p-\epsilon}{1-p}\right). \tag{38.437}$$

Alors

(1) $h_+(\epsilon) > 0$.

(2) Pour tout $n \geqslant 1$, nous avons

$$P\left(\frac{S_n}{n} \geqslant p+\epsilon\right) \leqslant e^{-nh_+(\epsilon)}. \tag{38.438}$$

(3) L'estimation (38.438) est optimale au sens que

$$\lim_{n\to\infty} \frac{1}{n}\ln\left(P\big(\frac{S_n}{n} \geqslant p+\epsilon\big)\right) = -h_+(\epsilon). \tag{38.439}$$

Démonstration. Pour tout $t \geqslant 0$ nous avons :

$$\begin{aligned}
P\left(\frac{S_n}{n} \geqslant p+\epsilon\right) &= P\big(S_n \geqslant np+n\epsilon\big) && \text{(38.440a)}\\
&\leqslant E\Big(e^{t(S_n-np-n\epsilon)}\Big) && \text{(38.440b)}\\
&= e^{-nt(p+\epsilon)}E\left(e^{tS_n}\right) && \text{(38.440c)}\\
&= e^{-nt(p+\epsilon)}\sum_{k=0}^{\infty} e^{tk}P(S_n=k) && \text{(38.440d)}\\
&= e^{-nt(p+\epsilon)}\sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k(1-p)^{n-k} && \text{(38.440e)}\\
&= e^{-nt(p+\epsilon)}\big((1-p)+pe^t\big)^n && \text{(38.440f)}\\
&= \exp\Big(-n\big(t(p+\epsilon)-\ln(1-p+pe^t)\big)\Big). && \text{(38.440g)}
\end{aligned}$$

Justifications :
— L'inégalité (38.440b) est l'inégalité de Markov (corollaire 38.82) avec $\phi(x) = e^{tx}$.
— La ligne (38.440d) est une utilisation du théorème de transfert 38.60.
Nous posons maintenant

$$h = \sup_{t>0}\big(t(p+\epsilon)-\ln(1-p+pe^t)\big). \tag{38.441}$$

Pour cette valeur de h nous avons

$$P\left(\frac{S_n}{n} \geqslant p+\epsilon\right) \leqslant e^{-nh}. \tag{38.442}$$

Nous considérons la fonction

$$\begin{aligned}
g\colon \mathbb{R}^+ &\to \mathbb{R}\\
t &\mapsto t(p+\epsilon)-\ln(1-p+pe^t).
\end{aligned} \tag{38.443}$$

Elle vérifie $g(0) = 0$ et

$$g'(t) = (p+\epsilon) - \frac{pe^t}{1-p+pe^t} \tag{38.444a}$$

$$g'(0) = (p+\epsilon) - \frac{p}{1-p+p} = \epsilon > 0. \tag{38.444b}$$

Par conséquent sur un voisinage de $t = 0$ la fonction g est strictement croissante et nous concluons que g prend (au moins) quelques valeurs strictement positives. Du coup nous avons

$$h = \|g\|_\infty > 0. \tag{38.445}$$

Nous cherchons maintenant pour quelle valeur de t est réalisé le maximum de g. D'abord résoudre $g'(t) = 0$ donne

$$t_0 = \ln\left(\frac{(p+\epsilon)(1-p)}{p(1-p-\epsilon)}\right) \tag{38.446}$$

Vu que $g'(t) \to p + \epsilon - 1 < 0$, nous avons $\lim_{t\to\infty} g(t) = -\infty$ et donc le t_0 trouvé est bien un maximum et non un minimum.

Il est maintenant loisible de calculer une valeur pour h : il suffit de calculer $g(t_0)$. La calcul n'est pas très compliqué et donne

$$h = g(t_0) = (p+\epsilon)\ln\left(\frac{p+\epsilon}{p}\right) + (p+\epsilon-1)\ln\left(\frac{1-p}{1-p-\epsilon}\right), \tag{38.447}$$

ce qui est bien $h = h_+(\epsilon)$. Cela démontre les points (1) et (2).

Nous montrons à présent l'aspect optimal de l'estimation. Nous savons déjà que

$$\frac{1}{n}\ln\left(P\big(\frac{S_n}{n} \geqslant p+\epsilon\big)\right) \leqslant h_+(\epsilon). \tag{38.448}$$

Nous posons $k_n = [n(p+\epsilon)]$. Vu que $p + \epsilon < 1$ et que $S_n \leqslant n$, nous avons

$$P\big(\frac{S_n}{n} \geqslant p+\epsilon\big) = P\big(S_n \geqslant n(p+\epsilon)\big) \geqslant P(S_n = k_n) = \binom{n}{k_n} p^{k_n}(1-p)^{n-k_n}. \tag{38.449}$$

C'est maintenant que nous utilisons la formule de Stirling (lemme 9.16) pour chacune des factorielles intervenant dans le coefficient binomial. Nous trouvons :

$$P\left(\frac{S_n}{n} \geqslant p+\epsilon\right) \geqslant P(S_n = k_n) = \clubsuit = \frac{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}\alpha(n) p^{k_n}(1-p)^{n-k_n}}{\left(\frac{k_n}{e}\right)^{k_n}\sqrt{2\pi k_n}\alpha(k_n)\left(\frac{n-k_n}{e}\right)^{n-k_n}\sqrt{2\pi(n-k_n)}}. \tag{38.450}$$

Nous savons que $k_n = n(p+\epsilon) + \sigma(n)$ avec σ borné par 1. Par conséquent $n - k_n \to \infty$ et nous pouvons regrouper les coefficients en α en

$$\beta(n) = \frac{\alpha(n)}{\alpha(k_n)\alpha(n-k_n)} \to 1. \tag{38.451}$$

Nous remarquons aussi que les e se simplifient. Nous récrivons $\clubsuit$ sous la forme

$$\clubsuit = \underbrace{\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k_n(n-k_n)}}\beta(n)}_{=A(n)} \frac{n^n p^{k_n}(1-p)^{n-k_n}}{k_n^{k_n}(n-k_n)^{n-k_n}} \tag{38.452a}$$

$$= A(n) n^n \left(\frac{p}{k_n}\right)^{k_n}\left(\frac{1-p}{n-k_n}\right)^{n-k_n} \tag{38.452b}$$

$$= A(n)\left(\frac{np}{k_n}\right)^{k_n}\left(\frac{n(1-p)}{n-k_n}\right)^{n-k_n}. \tag{38.452c}$$

Nous passons au logarithme et nous étudions $\frac{1}{n}\ln\big(P(S_n=k_n)\big)$. Nous avons les termes suivants à étudier :

$$\frac{1}{n}\ln\big(P(S_n=k_n)\big)=-\frac{1}{2n}\ln(2\pi)+\frac{1}{2n}\ln\left(\frac{n}{k_n(n-k_n)}\right)+k_n\ln\left(\frac{np}{k_n}\right)+(n-k_n)\ln\left(\frac{n(1-p)}{n-k_n}\right)+\frac{1}{n}\ln\big(\alpha(n)\big).\tag{38.453}$$

Nous étudions terme à terme la limite de cela lorsque $n\to\infty$.

(1) Le terme $\frac{1}{2n}\ln(2\pi)$ ne pose pas de problèmes. Il tend vers zéro.

(2) Si nous remplaçons k_n par $n(p+\epsilon)+\sigma(n)$ nous voyons que ce qui est dans le logarithme est majoré par $\frac{1}{P(n)}$ pour un certain polynôme P. Ce terme est dans le cas $\frac{\ln(P(n))}{n}$ qui tend vers zéro lorsque $n\to\infty$.

(3) Pour ce terme nous remplaçons k_n par $n(p+\epsilon)+k_n-n(p+\epsilon)$. Nous devons alors étudier la limite de

$$(p+\epsilon)\ln\left(\frac{np}{k_n}\right)+\frac{k_n-n(p+\epsilon)}{n}\ln\left(\frac{np}{k_n}\right).\tag{38.454}$$

Ce qui est dans les logarithmes est encadré de la façon suivante :

$$\frac{n(p+\epsilon)}{np}\leqslant\frac{k_n}{np}\leqslant\frac{n(p+\epsilon)+1}{np}.\tag{38.455}$$

Donc la limite de k_n/np est $(p+\epsilon)/p$. Les logarithmes restent bornés. Pour le second terme de (38.454), le numérateur du coefficient est borné par 1. Donc le second terme tend vers zéro et le tout tend vers

$$(p+\epsilon)\ln\left(\frac{p}{p+\epsilon}\right).\tag{38.456}$$

(4) Nous devons enfin étudier le dernier terme. La combinaison $\frac{n-k_n}{n}$ s'étudie de la façon suivante :

$$\frac{n-k_n}{n}=\frac{n-n(p+\epsilon)+n(p+\epsilon)-k_n}{n}=\frac{n(1-p-\epsilon)+n(p+\epsilon)-k_n}{n}\to1-p-\epsilon\tag{38.457}$$

parce que $n(p+\epsilon)-k_n$ est borné par 1. Sachant cela, notre terme a pour limite

$$\frac{n-k_n}{n}\ln\left(\frac{n(1-p)}{n-k_n}\right)\to(1-p-\epsilon)\ln\left(\frac{1-p}{1-p-\epsilon}\right).\tag{38.458}$$

En remettant tous les morceaux bouts à bout,

$$\frac{1}{n}P(S_n=k_n)\to(1+\epsilon)\ln\left(\frac{p}{p+/e}\right)+(1-p-\epsilon)\ln\left(\frac{1-p}{1-p-\epsilon}\right)=-h_+(\epsilon).\tag{38.459}$$

Étant donné que nous avions déjà prouvé que $P\big(\frac{S_n}{n}\geqslant p+\epsilon\big)\geqslant P(S_n=k_n)$, nous avons

$$\liminf_{n\to\infty}\frac{1}{n}\ln\Big(P\big(\frac{S_n}{n}\geqslant p+\epsilon\big)\Big)\geqslant\lim_{n\to\infty}\frac{1}{n}P(S_n=k_n)=-h_+(\epsilon).\tag{38.460}$$

En combinant avec (38.448) nous trouvons que

$$\lim_{n\to\infty}\frac{1}{n}\ln\Big(P\big(\frac{S_n}{n}\geqslant p+\epsilon\big)\Big)=h_+(\epsilon).\tag{38.461}$$

Pour cette dernière déduction nous utilisons le fait que si (a_n) est une suite telle que $a_n\leqslant l$ et $\liminf_{n\to\infty}a_n=l$, alors (a_n) admet une limite qui vaut l. □

38.7 Simulations de réalisations de variables aléatoires

Le générateur de base que possède un système informatique est un générateur de nombres pseudo-aléatoires de nombres entiers entre 0 et $m-1$ généré par une suite du type

$$x_{n+1} = (ax_n + b) \mod m. \tag{38.462}$$

38.7.1 Générateur uniforme

38.7.1.1 Première méthode

Une première façon de générer une variable aléatoire de loi uniforme sur $]0,1[$ est de diviser par $m-1$ la suite de x_n. En effet nous avons la proposition suivante.

Proposition 38.118.
Si (Y_n) est une suite de variables aléatoires indépendantes et identiquement distribuées de loi uniforme sur $\{0,\dots,n-1\}$. Alors

$$\frac{Y_n}{n} \xrightarrow{\mathscr{L}} \mathscr{U}[0,1]. \tag{38.463}$$

Démonstration. Nous prouvons la convergence en loi en passant par la fonction de répartition et la proposition 38.69. La fonction de répartition de la densité $\mathscr{U}[0,1]$ est

$$F_X(x) = x\mathbb{1}|_{[0,1]}. \tag{38.464}$$

La fonction de répartition de la variable aléatoire (discrète) $X_n = \frac{Y_n}{n}$ est

$$P(\frac{Y_n}{n} \leqslant x) = P(Y_n \leqslant nx) = \frac{\lfloor nx \rfloor}{n} \tag{38.465}$$

où $\lfloor a \rfloor$ est le plus grand entier inférieur à a. Nous avons évidemment

$$\lim_{n\to\infty} \frac{\lfloor nx \rfloor}{n} = x, \tag{38.466}$$

ce qui montre la convergence des fonctions de répartitions et donc la convergence en loi qui nous intéresse. □

38.7.1.2 Seconde méthode

Soit (Y_n) une suite de variables aléatoires indépendantes et identiquement distribuées selon la loi $Y_n \sim \mathscr{U}\{0,\dots,m-1\}$. Alors la série de variables aléatoires

$$Z = \sum_{k=0}^{\infty} \frac{Y_k}{m^{k+1}} \tag{38.467}$$

est une série qui converge presque sûrement parce que Y_k est borné par m. Avec probabilité zéro nous avons $Z = \sum_k 1/m^k$ qui converge. Nous avons

$$Z \sim \mathscr{U}[0,1]. \tag{38.468}$$

L'argument pour montrer cette loi est qu'en base m, la variable aléatoire Z a un développement décimal $Z = 0.Y_1Y_2Y_3Y_4\cdots$.

38.7.2 Simulation par inversion

Nous cherchons maintenant à simuler une loi X de fonction de répartition F.

Définition 38.119.
Soit $f : \mathbb{R} \to [0,1]$ une fonction croissante, continue à droite et telle que

$$\lim_{x\to-\infty} f(x) = 0 \tag{38.469a}$$
$$\lim_{x\to\infty} f(x) = 1. \tag{38.469b}$$

*L'**inverse généralisé** de f, notée f^{-1} est la fonction définie par*

$$f^{-1}(t) = \inf\{x \text{ tel que } f(x) \geqslant t\}. \tag{38.470}$$

Remarque 38.120.
L'inverse généralisé d'une fonction bijective est la vraie fonction réciproque usuelle.

Proposition 38.121.
Soit f une fonction admettant un inverse généralisé f^{-1}. Alors nous avons $f^{-1}(t) \leqslant a$ si et seulement si $t \leqslant f(a)$.

La continuité à droite joue pour démontrer cette proposition.

Proposition 38.122.
Si F est la fonction de répartition de la variable aléatoire X et si V est une variable aléatoire de loi uniforme $\mathscr{U}[0,1]$, alors $F^{-1}(U)$ a la même loi que X.

Démonstration. Nous montrons que les fonctions de répartition de X et de $F^{-1}(U)$ sont identiques. En utilisant la proposition 38.121, nous avons

$$\begin{aligned} P\big(F^{-1}(U) \leqslant y\big) &= P\big(U \leqslant F(y)\big) && \text{(38.471a)}\\ &= F(y) && \text{(38.471b)}\\ &= P(X \leqslant y). && \text{(38.471c)} \end{aligned}$$

Donc $F^{-1}(U)$ est la fonction de répartition de X. □

La difficulté de la méthode par inversion est qu'il faut être capable de calculer l'inverse de la fonction de répartition de la loi à simuler.

38.7.2.1 Loi exponentielle

La loi exponentielle est une loi qui peut être simulée par inversion. La fonction de répartition vaut

$$F(x) = 1 - e^{\lambda x}, \tag{38.472}$$

et l'inverse vaut

$$F^{-1}(x) = -\frac{1}{\lambda}\ln(1-y). \tag{38.473}$$

Par conséquent, une bonne formule pour simuler une loi exponentielle est

$$-\frac{1}{\lambda}\ln(1-U). \tag{38.474}$$

Notez que U étant uniforme, nous pouvons tout autant prendre $-\ln(U)/\lambda$.

38.7.3 Algorithme de Box-Muller

Il s'agit de simuler une loi gaussienne. La proposition est la suivante.

Proposition 38.123.
Si U et V sont des variables aléatoires indépendantes de même loi uniforme sur $[0,1]$, alors le couple

$$(X,Y) = \big(\sqrt{-2\ln(U)}\cos(2\pi V), \sqrt{-2\ln(U)}\sin(2\pi V)\big) \tag{38.475}$$

vérifie

(1) X est indépendante de Y

(2) X et Y sont de loi $\mathcal{N}(0,1)$.

Démonstration. Nous allons montrer la proposition en utilisant les fonctions tests. Soit donc $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ une fonction bornée et mesurable. Soient Z et W, deux variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$. Nous allons montrer que

$$F\big(\varphi(X,Y)\big) = E\Big[\varphi\big(\sqrt{-2\ln(U)}\cos(2\pi V), \sqrt{-2\ln(U)}\sin(2\pi V)\big)\Big]. \tag{38.476}$$

Par indépendance de U et V, la densité du couple est le produit des densités, donc en passant aux coordonnées polaires,

$$\begin{aligned}
\diamondsuit &= E\big[\varphi(Z,W)\big] && (38.477a)\\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{\infty}^{\infty}\varphi(x,y)e^{-x^2/2}e^{-y^2/2}dxdy && (38.477b)\\
&= \frac{1}{2\pi}\int_0^{2\pi}d\theta\int_0^{\infty}\varphi(r\cos\theta, r\sin\theta)e^{-r^2/2}rdr. && (38.477c)
\end{aligned}$$

Nous posons $u = e^{-r^2/2}$ et $v = \frac{\theta}{2\pi}$. En particulier $r = \sqrt{-2\ln(u)}$ et

$$\begin{aligned}
\diamondsuit &= \int_0^1\int_0^1 \varphi\big(\sqrt{-2\ln(u)}\cos(2\pi v), \sqrt{-2\ln(u)}\sin(2\pi v)\big)dudv && (38.478a)\\
&= E\Big(\varphi\big(\sqrt{-2\ln(U)}\cos(2\pi V), \sqrt{-2\ln(U)}\sin(2\pi V)\big)\Big) && (38.478b)
\end{aligned}$$

parce que mesure $dudv$ est la densité de la loi uniforme. □

En pratique, la formule

$$(x,y) \mapsto \big(\sqrt{-2\ln x}\cos(2\pi y), \sqrt{-2\ln x}\sin(2\pi y)\big) \tag{38.479}$$

est une façon d'obtenir deux gaussiennes à partir de deux variables uniformes.

38.7.4 Méthode du rejet

La méthode du rejet permet de simuler des lois à densité. Soit f la densité de la loi à simuler. Nous faisons les hypothèses suivantes.

(1) Il existe une densité g d'une variable aléatoire facile à simuler.

(2) Il existe un $k \geqslant 0$ tel que $f(x) \leqslant kg(x)$.

Remarque 38.124.
Le k de la seconde hypothèse est nécessairement plus grand que 1. En effet,

$$1 = \int f \leqslant k\int g = k \tag{38.480}$$

parce que f et g sont des densités et ont donc une intégrale égale à 1.

Proposition 38.125.
Soient (X_n) et (U_n) des suites de variables aléatoires indépendantes au sens où non seulement les X_i et U_k sont indépendants entre eux, mais de plus X_i est indépendant de U_j pour tout i et j. Nous supposons que les X_i sont indépendantes et identiquement distribuées, de densité g et que les U_i sont indépendantes et identiquement distribuées de loi uniforme.

Nous introduisons la variable aléatoire à valeurs dans $\mathbb{N}$

$$p(\omega) = \inf\{n \geqslant 0 \text{ tel que } \alpha\big(X_n(\omega)\big) \geqslant U_n(\omega)\} \tag{38.481}$$

où α est la fonction définie par

$$\alpha(x) = \begin{cases} \frac{f(x)}{kg(x)} & si\ g(x) \neq 0 \\ 0 & si\ g(x) = 0. \end{cases} \tag{38.482}$$

Alors la variable aléatoire Y définie par

$$Y(\omega) = X_{p(\omega)}(\omega) \tag{38.483}$$

admet f pour densité.

Démonstration. D'abord étant donné que $f(x) \leqslant kg(x)$ nous avons $\alpha(x) \in [0,1]$. Nous pouvons a priori avoir $p(\omega) = \infty$, ce qui rendrait caduque la définition de $Y(\omega)$. Montrons donc pour commencer que $P(p = \infty) = 0$. En utilisant l'indépendance nous avons

$$\begin{aligned} P\big(\alpha(X_n) < U_n, \forall n\big) &= \lim_{N\to\infty} \prod_{i=1}^{N} P\big(\alpha(X_i) < U_i\big) && \text{(38.484a)} \\ &= \lim_{N\to\infty} P\big(\alpha(X_1) < U_1\big)^N. && \text{(38.484b)} \end{aligned}$$

Pour conclure nous devons prouver que $P\big(\alpha(X_1) < U_1\big) < 1$. Pour cela nous calculons

$$\begin{aligned} P\big(\alpha(X_1) < U_1\big) &= \int_{\mathbb{R}} dx \int_0^1 du \mathbb{1}_{\alpha(x)<u} g(x) && \text{(38.485a)} \\ &= \int_{\mathbb{R}} g(x)\big(1 - \alpha(x)\big)dx && \text{(38.485b)} \\ &= \int_{\mathbb{R}} \left(g(x) - \frac{f(x)}{k}\right) && \text{(38.485c)} \\ &= 1 - \frac{1}{k} && \text{(38.485d)} \\ &< 1. && \text{(38.485e)} \end{aligned}$$

L'équation (38.484) nous permet donc de conclure que $P\big(\alpha(X_n) < U_n, \forall n\big) = 0$. Par conséquent la variable aléatoire $Y(\omega) = X_{p(\omega)}(\omega)$ a un sens.

Nous devons maintenant prouver que Y a bien f pour densité. Pour cela nous considérons un ensemble mesurable $A \in \mathcal{B}\text{or}(\mathbb{R})$ et nous montrons que $P(Y \in A) = \int_A f(x)dx$. Nous avons

$$P(Y \in A) = P(X_p \in A) = \sum_{i=1}^{\infty} P(X_j \in A, p = j). \tag{38.486}$$

Par ailleurs nous avons

$$\begin{aligned} P(X_j \in A, p = j) &= P\big(X_j \in A, \alpha(X_j) \geqslant U_j, \alpha(X_m) < U_m\, \forall m \leqslant j-1\big) \\ &= P\big(X_j \in A, \alpha(X_j) \geqslant U_j\big) P\big(\alpha(X_1) < U_1\big)^{j-1} \\ &= P\big(X_j \in A, \alpha(X_j) \geqslant U_j\big) \left(1 - \frac{1}{k}\right)^{j-1}. \end{aligned} \tag{38.487}$$

Étant donné que $g(x)dx$ est la densité de X_j et que du est la densité de U, nous avons

$$\begin{aligned} P\big(X_j \in A, \alpha(X_j) \geqslant U_j\big) &= \int_{\mathbb{R}} \int_0^1 \mathbb{1}_{x\in A} \mathbb{1}_{\alpha(x)\geqslant u} g(x) du dx && \text{(38.488a)} \\ &= \int_{\mathbb{R}} g(x) \mathbb{1}_{x\in A} \underbrace{\int_0^1 \mathbb{1}_{\alpha(x)\geqslant u} du}_{\alpha(x)}\, dx. && \text{(38.488b)} \\ &= \int_{\mathbb{R}} \mathbb{1}_{x\in A} \frac{f(x)}{k} dx && \text{(38.488c)} \\ &= \frac{1}{k} P(X \in A). && \text{(38.488d)} \end{aligned}$$

En remplaçant dans l'équation (38.487) nous trouvons

$$P(X_j \in A, p = j) = \frac{1}{k} P(X \in A) \left(1 - \frac{1}{k}\right)^{j-1}. \tag{38.489}$$

Et enfin, l'équation (38.486) donne

$$P(Y \in A) = \frac{1}{k} P(X \in A) \sum_{i=1}^{\infty} \left(1 - \frac{1}{k}\right)^{j-1} = P(X \in A). \tag{38.490}$$

□

38.7.5 Simuler une loi géométrique à l'ordinateur

Si (X_n) est une suite de variables aléatoires indépendantes et identiquement distribuées avec $X_n \sim \mathscr{B}(p)$, alors

$$Z = \min\{k \geqslant 1 \text{ tel que } X_k = 1\} \sim \mathscr{G}(p). \tag{38.491}$$

Nous avons alors $P(Z = k) = (1 - p)^k p$.

Si nous avons un générateur de lois de Bernoulli de paramètre p, alors nous on simulons jusqu'à obtenir 1 et nous comptons combien de simulations ont été nécessaires.

38.7.6 Simuler une loi exponentielle à l'ordinateur

Nous pouvons utiliser la méthode de l'inversion. Étant donné que la fonction de répartition de la loi exponentielle est $F(x) = 1 - e^{-\lambda x}$, nous avons $F^{-1}(y) = \frac{1}{\lambda} \ln(1 - y)$. Par conséquent à partir d'un générateur uniforme U, nous pouvons calculer

$$F^{-1}(U) = \frac{1}{\lambda} \ln(U) \tag{38.492}$$

qui suivra une loi exponentielle d'espérance $1/\lambda$.

38.7.7 Simuler une loi de Poisson à l'ordinateur

Nous savons du point 38.5.8 que si les T_i sont des variables aléatoires indépendantes et identiquement distribuées de loi $\mathscr{E}(\lambda)$, alors nous avons

$$\max\{n \geqslant 1 \text{ tel que } \sum_i T_i \leqslant 1\} \sim \mathscr{P}(\lambda). \tag{38.493}$$

La façon usuelle pour créer une loi exponentielle est d'avoir un générateur de loi uniforme U_i et d'écrire que

$$-\frac{1}{\lambda} \ln(U_i) \sim \mathscr{E}(\lambda). \tag{38.494}$$

Nous devons donc faire la somme de telles variables aléatoires et voir à partir de quel moment la somme dépasse 1. Le calcul est le suivant :

$$-\sum_{i=1}^{n} \frac{1}{\lambda} \ln(U_i) \leqslant 1 \tag{38.495}$$

implique

$$\prod_{i=1}^{n} U_i \leqslant e^{-\lambda}. \tag{38.496}$$

En pratique, la variable aléatoire qui se comporte comme une loi de Poisson de paramètre λ est

$$N = \max\{n \geqslant 1 \text{ tel que } \prod_{i=1}^{n} U_i \geqslant e^{-\lambda}\}. \tag{38.497}$$

Nous générons donc des nombres aléatoires entre 1 et 1 et nous effectuons le produit jusqu'à ce qu'il passe en dessous de $e^{-\lambda}$. À ce moment, nous retournons le nombre de nombres qu'il a fallu générer.

38.8 Sage

Nous allons montrer maintenant quelques trucs importants dans l'utilisation de Sage pour réaliser des petits graphiques.

Remarque 38.126.
Dans ce qui suit, nous allons parler de « Sage », mais en réalité nous allons surtout parler du module `scipy` qui fait partie des modules hyper-usuels de Python. Les remerciements vont donc au moins autant du côté de l'équipe de `scipy` que vers celle de Sage.

38.8.1 Loi exponentielle

Il faut savoir que la définition d'une loi continue retourne automatiquement la loi centrée réduite. Pour avoir une loi exponentielle de moyenne donnée, il faut donc préciser de façon plus maligne que ce que l'on croit.

```
from scipy import stats

X=stats.expon(scale=5)
print(X.mean())      # retourne 5

P=plot( X.pdf,x,0,10 )
show(P)              # Affiche le graphique
```

tex/frido/code_sage1.py

38.8.2 Inverser des lois

Pour trouver des intervalles de confiance, il faut souvent calculer des inverses de loi. Bien entendu Sage le fait. Ce que sage connaît, c'est l'inverse de la fonction de survie. Autrement dit si `X` est une variable aléatoire, `X.sf` est la fonction $x \mapsto 1-P(X<x)$ et `X.isf` en est l'inverse. Pour résoudre $P(X<\xi)=\alpha$, il faut résoudre $F(\xi)=\alpha$, c'est à dire

$$1-F(\xi)=1-\alpha, \tag{38.498}$$

ce qui se fait de la façon suivante : le programme suivant donne pour une loi normale centrée réduite la valeur de ξ pour laquelle $P(N<\xi)=0.05$:

```
from scipy import stats

N=stats.norm
print N.mean()       # 0
print N.var()        # 1

xi = N.isf(0.95)
print xi             # -1.64485

N.cdf(xi)            # Vérification : 0.05

# Graphiques de la fonction de densité et la cumulative.
P=plot(N.cdf,x,-10,10)
Q=plot(N.pdf,x,-10,10,color="red")
show(P+Q)
```

tex/frido/code_sage2.py

38.9 Monte-Carlo

Nous voudrions calculer une valeur approchée de l'intégrale

$$I = \int_a^b f(x)dx. \tag{38.499}$$

Les méthodes classiques consistent à discrétiser l'intervalle $[a, b]$ et en calculant une somme de la forme $\sum_i w_i f(x_i)$.

L'idée de Monté Carlo est de remplacer le découpage déterministe x_i par des variables aléatoires X_i en trois étapes.

(1) Pour cela nous commençons par écrire l'intégrale comme une espérance : $I = E(X)$ où X est une variable aléatoire définie sur un espace probabilisé $(\Omega, \mathcal{F}, P)$ à déterminer. Une contrainte est évidemment d'avoir $X \in L^1(\Omega, \mathcal{F}, P)$.

(2) Nous générons une suite indépendantes et identiquement distribuée de variables aléatoires (X_n) de même loi que X et la loi (forte) des grands nombres implique que

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{p.s.} E(X) = I. \tag{38.500}$$

(3) Le dernier point sera de donner un intervalle de confiance.

Exemple 38.127
Nous voudrions déterminer de façon approchée l'intégrale $I = \int_0^1 f(x)dx$. Si $U \sim \mathscr{U}[0, 1]$, alors

$$I = E\big(f(U)\big) \tag{38.501}$$

et il suffit de faire

$$I \simeq \frac{1}{n}\sum_{k=1}^{n} f(U_i). \tag{38.502}$$

où mes U_i sont indépendantes et identiquement distribuées de loi $\mathscr{U}[0, 1]$. △

Exemple 38.128
Supposons que la fonction à intégrer se présente sous la forme $f(x) = h(x)g(x)$ avec $g \geqslant 0$ et telle que l'intégrale $\int_{\mathbb{R}} g$ existe. Notons

$$c = \int_{\mathbb{R}} g \tag{38.503}$$

et

$$I = \int_{\mathbb{R}} h(x)c\frac{g(x)}{c}dx. \tag{38.504}$$

Nous avons alors $I = E\big(ch(Y)\big)$ où Y admet la densité $g(x)c$. △

Passons au cas de plusieurs variables et considérons l'intégrale

$$I = \int_{[0,1]^1} f(x_1, \ldots, x_d)dx_1 \ldots dx_d. \tag{38.505}$$

Nous écrivons

$$I = E\big(f(U_1, \ldots, U_d)\big) \tag{38.506}$$

où les U_i sont de loi uniformes sur $[0, 1]$. En pratique, nous générons une suite de variables aléatoires de (Z_k) de lois uniformes que nous regroupons par paquets :

$$V_k = (Z_{dk}, Z_{dk+1}, \ldots, Z_{d(k+1)-1}). \tag{38.507}$$

Ces variables aléatoires V_k sont indépendantes et identiquement distribuées de loi $\mathscr{U}[0, 1]^d$. Ensuite la loi des grands nombres nous indique que

$$I \sim \frac{1}{n}\sum_{k=1}^{n} f(V_k). \tag{38.508}$$

38.9.1 Intervalle de confiance

38.9.1.1 Principe

Nous supposons que nous travaillons sur une approximation de Monte-Carlo telle que la variable aléatoire choisie soit dans L^2. La loi des grands nombres nous dit que $\bar{X}_n \sim I$ tandis que le théorème central limite nous enseigne que

$$\frac{\bar{X}_n - E(X)}{\sigma/\sqrt{n}} \xrightarrow{\mathscr{L}} \mathcal{N}(0,1). \tag{38.509}$$

Par conséquent

$$P\left(\frac{\bar{X}_n - E(X)}{\sigma/\sqrt{n}} \in [-u,u]\right) \simeq P(-u \leqslant Z \leqslant u) \tag{38.510}$$

où $Z \sim \mathcal{N}(0,1)$. En remplaçant $E(X)$ par I et en effectuant les manipulations usuelles, nous trouvons que $P(I \in J_\alpha) = 1-\alpha$ si

$$J_\alpha = \left[\bar{X}_n - u_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X}_n + u_\alpha \frac{\sigma}{\sqrt{n}}\right] \tag{38.511}$$

où σ^2 est la variance de X. Si σ n'est pas connue, alors nous le remplaçons par un estimateur

$$S'_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2 \tag{38.512}$$

et nous considérons l'intervalle

$$J'_\alpha = \left[\bar{X}_n - u_\alpha \frac{S'_n}{\sqrt{n}} \bar{X}_n + u_\alpha \frac{S'_n}{\sqrt{n}}\right]. \tag{38.513}$$

Il y a deux façons de faire diminuer la longueur de l'intervalle de confiance : augmenter n ou diminuer σ. Pour le second point, le choix de X dans $I = E(X)$ est essentiel.

Exemple 38.129
Soit à calculer

$$I = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{\beta z} e^{-z^2/2} dz \tag{38.514}$$

avec $\beta > 0$. Nous introduisons la variable aléatoire $X = e^{\beta Z}$ avec $Z \sim \mathcal{N}(0,1)$. Nous avons alors

$$I = E(X). \tag{38.515}$$

Par ailleurs l'intégrale demandée vaut $e^{\beta^2/2}$. En appliquant les formules vues plus haut nous trouvons

$$J_\alpha = \left[\bar{X}_n - u_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X}_n + u_\alpha \frac{\sigma}{\sqrt{n}}\right] \tag{38.516}$$

où

$$\sigma^2 = \mathrm{Var}(X) = E(e^{2\beta Z}) - E(e^{\beta z})^2 = e^{2\beta^2} - e^{\beta^2}. \tag{38.517}$$

Nous avons utilisé la formule (38.382). Si nous choisissons $\beta = 2$, nous trouvons $\sigma^2 \simeq 2926$. Donc si nous voulons une longueur de J_α plus petite que 10^{-2} tout en demandant $\alpha = 0.05$ (ce qui implique $u_\alpha = 1.96$), nous devons avoir

$$1.96\frac{2973}{\sqrt{n}} < 10^{-2}, \tag{38.518}$$

c'est à dire environ $n = 10^{11}$, ce qui soit dit en passant est très largement au delà des capacités de la commande rand de scilab. △

Nous allons maintenant voir quelques méthodes pour réduire la variance.

38.9.1.2 Échantillonnage préférentiel

Nous devons calculer $I = \int_{\mathbb{R}} f(x)dx$. Pour cela nous introduisons artificiellement une densité g et nous écrivons

$$I = \int_{\mathbb{R}} \frac{f(x)}{g(x)} g(x)dx = E\left(\frac{f(Y)}{g(Y)}\right) \tag{38.519}$$

où Y est de densité g. Il faut essayer de trouver g de telle sorte à ce que

$$\operatorname{Var}\left(\frac{f(Y)}{g(Y)}\right) \tag{38.520}$$

soit la plus petite possible.

38.9.1.3 Méthode de la variable de contrôle

Soit $I = E(X)$. Nous introduisons une variable aléatoire Z et nous écrivons

$$I = E(X - Z) + E(Z). \tag{38.521}$$

Il faut alors choisir Z de telle sorte que $E(Z)$ soit calculable et que $X - Z$ ait une variance plus faible. En particulier, Z ne peut pas être indépendante de X.

38.9.1.4 Variables antithétiques

Soit $I = \int_0^1 f(x)dx$. La première idée (exemple 38.127) est d'écrire

$$I = E\big(f(U)\big) \tag{38.522}$$

où $U \sim \mathscr{U}[0,1]$, mais nous n'avons pas de garanties sur la variance de $f(U)$. Nous pouvons écrire

$$I = E\big(f(U)\big) = E\Big[\frac{1}{2}\big(f(U) - f(1-U)\big)\Big]. \tag{38.523}$$

Ici $1-U$ est encore une variable aléatoire uniforme sur $[0,1]$, mais il se fait que la variable aléatoire

$$Z = \frac{1}{2}\big(f(U) + f(1-U)\big) \tag{38.524}$$

a une variance inférieure à $\operatorname{Var}\big(f(U)\big)$. En effet, $f(U)$ et $f(1-U)$ ne sont pas indépendantes, par conséquent le résultat du lemme 38.23 n'est pas valide, par contre la proposition 38.28 reste vraie et nous avons

$$\operatorname{Var}(Z) = \frac{1}{4}\operatorname{Var}\big(f(U)\big) + \frac{1}{4}\operatorname{Var}\big(f(1-U)\big) + \frac{1}{2}\operatorname{Cov}\big(f(U), f(1-U)\big). \tag{38.525}$$

Nous avons $\operatorname{Var}\big(f(1-U)\big) = \operatorname{Var}\big(f(U)\big)$. En ce qui concerne le terme avec la covariance, nous lui appliquons l'équation (38.56) :

$$\begin{aligned}
\operatorname{Cov}\big(f(U), f(1-U)\big) &= E\Big(\big(f(U) - I\big)\big(f(1-U) - I\big)\Big) && \text{(38.526a)}\\
&\leqslant E\big((f(U) - I)^2\big)^{1/2} E\big((f(1-U) - I)^2\big)^{1/2} && \text{(38.526b)}\\
&= \operatorname{Var}\big(f(U)\big) && \text{(38.526c)}
\end{aligned}$$

où nous avons utilisé le fait que $E\big(f(U)\big) = E\big(f(1-U)\big) = I$. Au final nous avons bien obtenu

$$\operatorname{Var}(Z) \leqslant \operatorname{Var}\big(f(U)\big). \tag{38.527}$$

38.10 Résultats qui se démontrent avec des variables aléatoires

38.10.1 Nombres normaux

Tout nombre $x \in [0,1[$ admet un unique[13] développement en base $b \geqslant 2$:

$$x = \sum_{n=1}^{\infty} \frac{\epsilon_n(x)}{b^n} \tag{38.528}$$

avec $\epsilon_n(x) \in \mathcal{A} = \{0, \dots, b-1\}$.

Soit $k \geqslant 1$ et $r \in \mathcal{A}^k$; nous posons

$$N_x(r,n) = \text{Card}\big\{i \in \{1,\dots,n-k+1\} \text{ tel que } \epsilon_1(x) = r_1, \dots, \epsilon_{i+k-1} = r_k\big\}. \tag{38.529}$$

C'est le nombre d'occurrences du motif r (de longueur k) dans les n premières décimales de x.

Définition 38.130.
*Un nombre $x \in [0,1[$ est **normal** en base b si pour tout $r \in \{0,\dots,b-1\}^k$ nous avons*

$$\frac{N_x(b,n)}{n} \to \frac{1}{b^k}. \tag{38.530}$$

Un nombre est normal s'il est normal en toute base.

Proposition 38.131 ([426, 4]).
Au sens de la mesure de Lebesgue, presque tous les nombres de $[0,1[$ sont normaux.

Démonstration. Pour $x \in [0,1[$, nous notons $\epsilon_n(x)$ son développement en base b. Cela nous donne des variables aléatoires $\epsilon_i : [0,1[\to \mathcal{A}$ dont la loi de probabilité est donnée par

$$P(\epsilon_1 = d) = P\big([\frac{d}{b}, \frac{d+1}{b}[\big) = \frac{1}{b} \tag{38.531}$$

parce que l'intervalle $[\frac{d}{b}, \frac{d+1}{d}[$ est l'ensemble des nombres de $[0,1[$ dont la première décimale est d. Pour la loi des ϵ_i, il faut un peu plus découper, mais ça donne le même résultat : $P(\epsilon_i = d) = 1/b$. Ces variables aléatoires sont indépendantes et identiquement distribuées. Nous considérons aussi la variable aléatoire

$$\begin{aligned} N(r,n) : [0,1[&\to \mathbb{N} \\ x &\mapsto N_x(r,n) \end{aligned} \tag{38.532}$$

Pour un $r \in \mathcal{A}$ fixé, nous définissons encore la variable aléatoire

$$\begin{aligned} X_j : \mathcal{A} &\to \{0,1\} \\ x &\mapsto \begin{cases} 1 & \text{si } \epsilon_j(x) = b \\ 0 & \text{sinon.} \end{cases} \end{aligned} \tag{38.533}$$

Les variables aléatoires X_j sont des variables aléatoires de Bernoulli indépendantes et identiquement distribuées de paramètre $E(X_j) = P(X_j = 1) = P(\epsilon_1 = b) = \frac{1}{b}$. Nous pouvons utiliser dessus la loi forte des grands nombres (théorème 38.83). Pour dire que

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s.} E(X_1) = \frac{1}{b}. \tag{38.534}$$

Mais en réalité nous avons aussi $\sum_{j=1}^{n} X_j = N(r,n)$ parce que en appliquant à $x \in [0,1[$:

$$\sum_{j=1}^{n} \begin{cases} 1 & \text{si } \epsilon_j(x) = r \\ 0 & \text{sinon} \end{cases} = \text{Card}\big\{i \in \{1,\dots,n\} \text{ tel que } \epsilon_i(x) = r\big\} = N_x(r,n), \tag{38.535}$$

13. Nous excluons 1 parce que son développement en puissances négatives de b est zéro.

de sorte que l'équation (38.534) nous dit exactement que pour tout $r \in \mathcal{A}$,

$$\lim_{n\to\infty} \frac{N_x(r,n)}{n} = \frac{1}{b} \tag{38.536}$$

pour presque tout $x \in [0,1[$.

Il reste à prouver la même chose pour tout $r \in \mathcal{A}^k$. Voyons avec $k = 2$ et $r = (u,v) \in \mathcal{A}^2$. Nous posons

$$Y_j = \mathbb{1}_{\{\epsilon_j = u, \epsilon_{j+1} = v\}}, \tag{38.537}$$

et $N(r,n) = \sum_{j=1}^{n-1} Y_j$. Les Y_i sont encore des binomiales de paramètre $\frac{1}{b^2}$, mais elles ne sont pas indépendantes. En effet pour avoir $Y_1(x) = Y_2(x) = 1$, il faut que les trois premières décimales de x soit en même temps de la forme $uv.$ et $.uv$, donc

$$P(Y_1, Y_2 = 1) = b^3 \delta_{u,v} \tag{38.538}$$

alors que $P(Y_1 = 1)P(Y_2 = 2) = b^4$. Nous pouvons contourner ce problème en remarquant que les ϵ_i, eux, sont indépendants. Donc le lemme de regroupement 38.16 nous dit que la famille $\{Y_{2n}\}$ est une famille de variables aléatoires indépendantes (et idem pour la famille Y_{2n-1}). En effet, les variables aléatoires Y_{2n} correspondent à la partition 23, 45, 67, etc.

Nous appliquons la loi des grands nombres sur les deux familles indépendamment :

$$\frac{1}{n}\sum_{j=1}^{n} Y_{2j-1} \xrightarrow{p.s.} \frac{1}{b^2} \tag{38.539}$$

et

$$\frac{1}{n}\sum_{j=1}^{n} Y_{2j} \xrightarrow{p.s.} \frac{1}{b^2} \tag{38.540}$$

Pour rappel, le but pour l'instant est d'établir la limite $\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} Y_j = \frac{1}{b^2}$. Nous allons l'établir séparément pour les termes pairs et impairs de la suite. Pour les pairs :

$$\frac{1}{2n}\sum_{j=1}^{2n} Y_j = \frac{1}{2}\left(\frac{1}{n}\sum_{j=1}^{n} Y_{2j-1}\right) + \frac{1}{2}\left(\frac{1}{n}\sum_{j=1}^{n} Y_{2j}\right) \xrightarrow{p.s.} \frac{1}{b^2} \tag{38.541}$$

Pour les impairs [14],

$$\frac{1}{2n-1}\sum_{j=1}^{2n-1} Y_j = \frac{n}{2n-1}\left(\frac{1}{n}\sum_{j=1}^{n} Y_{2j-1}\right) + \frac{n}{2n-1}\left(\frac{1}{n}\sum_{j=1}^{n} Y_{2j}\right) \to \frac{1}{b^2} \tag{38.542}$$

parce que les deux parenthèses convergent vers $\frac{1}{b^2}$ alors que les coefficients devant convergent vers $\frac{1}{2}$.

Au final nous avons bien

$$\frac{N(r,n)}{n} = \frac{1}{n}\sum_{j=1}^{n-1} Y_j = \frac{n-1}{n}\left(\frac{1}{n-1}\sum_{j=1}^{n-1} Y_j\right) \to \frac{1}{b^2} \tag{38.543}$$

tant que $r \in \mathcal{A}^2$.

Pour prouver la même chose avec $r \in \mathcal{A}^k$, il suffit de faire le même raisonnement en divisant en plus de paquets : $\{Y_{kj+m}\}_{m=1,\dots,k-1}$ sont indépendants et nous utilisons k fois la loi des grands nombres.

Donc pour toute base b nous savons que les nombres normaux en base b forment un ensemble de mesure nulle dans $[0,1[$. Il reste à voir que leur union reste de mesure nulle. Cela est vrai parce que nous avons une union dénombrable et qu'une union dénombrable d'ensembles de mesure nulle est de mesure nulle par le lemme 16.34. □

14. Ici dans [4], la seconde somme va jusqu'à $n-1$ et je ne comprends pas pourquoi.

Remarque 38.132.
Un nombre x est normal en base b si et seulement si la suite $u_k = xb^k$ est équirépartie modulo 1 sur $[0,1]$ (c'est à dire quel la suite des parties fractionnelles des u_k est équirépartie). Pour le nombre $0.2357873\ldots$, nous parlons de la suite $0.2357873\ldots$; $0.357873\ldots$; $0.57897\ldots$ etc. C'est la suite des queues de suites de la suite de ses décimales [15].

38.10.2 Théorème de Bernstein

Théorème 38.133 (Théorème de Bernstein[4])**.**
Soit $f \in C^0\big([0,1],\mathbb{C}\big)$ et son module de continuité

$$\begin{aligned}\omega\colon [0,1] &\to \mathbb{R}\\ h &\mapsto \sup\{|f(u)-f(v)| \text{ tel que } |u-v|<h\}.\end{aligned} \tag{38.544}$$

*Pour $n \geqslant 0$ nous définissons le n^e **polynôme de Bernstein** de f par*

$$B_n(f)(x) = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} f\left(\frac{k}{n}\right). \tag{38.545}$$

Alors il existe C tel que pour tout $n \geqslant 1$:

(1)

$$\|f - B_n(f)\|_\infty \leqslant C\omega\left(\frac{1}{\sqrt{n}}\right). \tag{38.546}$$

(2)

$$B_n(f) \xrightarrow{unif} f \tag{38.547}$$

sur $[0,1]$.

(3) L'inégalité (38.546) est optimale : il existe une fonction $g \in C^0\big([0,1],\mathbb{C}\big)$ et $\delta > 0$ tels que pour tout $N \geqslant 1$, $\|g - B_n(g)\|_\infty \geqslant \frac{\delta}{\sqrt{n}}$. Cette fonction peut être choisie Lipschitzienne. Une telle fonction est donnée par exemple par $g(x) = |x - \frac{1}{2}|$.

Démonstration. Soit $x \in [0,1]$ et une suite de variables aléatoires de Bernoulli indépendantes [16] et identiquement distribuées $(X_i)_{i\geqslant 1}$ de paramètre x. Nous notons $S_n = \sum_{k=1}^n X_k$.

(1) Pour cette histoire de convergence, il faut majorer la quantité $\big|f(x) - B_n(f)(x)\big|$. Pour cela il y a trois astuces. La première est de se souvenir que $E\big(f(x)\big) = f(x)$, et la seconde est que le théorème de transfert 38.60 appliqué à $x \mapsto f(x/n)$ donne [17]

$$E\left(f(\frac{S_n}{n})\right) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) P(S_n = k) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} x^k (1-x)^{n-k}, \tag{38.548}$$

c'est à dire que

$$B_n(f)(x) = E\left(f(\frac{S_n}{n})\right). \tag{38.549}$$

Et enfin la troisième astuce est d'utiliser le lemme 19.123 pour avoir

$$\omega\left(|x - \frac{S_n}{n}|\right) = \omega\left(\frac{1}{\sqrt{n}}|\sqrt{n} - \frac{S_n}{\sqrt{n}}|\right) \leqslant \left(\sqrt{n}|x - \frac{S_n}{n}| + 1\right)\omega(\frac{1}{\sqrt{n}}). \tag{38.550}$$

15. C'est pas trop bien dit, mais on se comprend, non ?
16. Définition 38.9.
17. Nous avons aussi utilisé la formule de l'espérance pour les variables aléatoires discrètes.

À partir de là nous pouvons un peu calculer :

$$|f(x)-B_n(f)(x)| = \left|E\left(f(x)-f(\frac{S_n}{n})\right)\right| \tag{38.551a}$$
$$\leqslant E\left(|f(x)-f(\frac{S_n}{n})|\right) \tag{38.551b}$$
$$\leqslant E\left(\omega\left(|x-\frac{S_n}{n}|\right)\right) \tag{38.551c}$$
$$\leqslant \omega\left(\frac{1}{\sqrt{n}}\right)E\left(|\sqrt{n}x-\frac{S_n}{n}|+1\right). \tag{38.551d}$$

Le dernier facteur peut être récrit sous la forme

$$E\left(\sqrt{n}|x-\frac{S_n}{n}|+1\right) = \sqrt{n}E\left(|x-\frac{S_n}{n}|\right)+1, \tag{38.552}$$

et c'est là que nous pouvons utiliser l'inégalité de Hölder 28.23 :

$$E\big(|X|\big) = \|X\|_1 \leqslant \|X\|_2 \tag{38.553}$$

où $\|X\|_2$ désigne

$$\|X\|_2 = \sqrt{E\big(|X|^2\big)}. \tag{38.554}$$

Nous pouvons donc écrire

$$|f(x)-B_n(f)(x)| \leqslant \omega\left(\frac{1}{\sqrt{n}}\right)\left(\sqrt{n}\|x-\frac{S_n}{n}\|_2+1\right). \tag{38.555}$$

Nous étudions maintenant de plus près la quantité $\|x-\frac{S_n}{n}\|_2$. D'abord

$$E\left(|x-\frac{S_n}{n}|^2\right) = x^2-2\frac{x}{n}E(S_n)+\frac{1}{n^2}E(S_n^2). \tag{38.556}$$

Ensuite nous savons l'espérance de S_n (qui vaut $E(S_n)=nx$) par (38.326) et le lemme 38.22 nous permet de calculer $E(S_n^2)$ par indépendance des X_i qui composent S_n. Nous avons alors

$$E\left(|x-\frac{S_n}{n}|^2\right) = x^2-2x^2+\frac{1}{n^2}\sum_{1\leqslant i\neq j\leqslant n}E(X_i)E(X_j)+\frac{1}{n^2}\sum_{i=1}^{n}E(X_i^2) \tag{38.557a}$$
$$= -x^2+\frac{n^2-n}{n^2}x^2+\frac{nx}{n^2} \tag{38.557b}$$
$$= \frac{x(1-x)}{n}. \tag{38.557c}$$

Quelques justifications :

— $E(X_i)=E(X_i^2)=x$ parce que X_i est une variable aléatoire de Bernoulli de paramètre x.

— La première somme contient tous les couples (i,j) sauf les diagonaux ; il y en a donc n^2-n.

En recombinant le tout,

$$|f(x)-B_n(f)(x)| \leqslant \omega\left(\frac{1}{\sqrt{n}}\right)\left(\sqrt{n}\sqrt{\frac{x(1-x)}{n}}+1\right) \tag{38.558a}$$
$$= \omega\left(\frac{1}{\sqrt{n}}\right)\left(\sqrt{x(1-x)}+1\right) \tag{38.558b}$$
$$\leqslant \frac{3}{2}\omega\left(\frac{1}{\sqrt{n}}\right). \tag{38.558c}$$

La dernière majoration est une rapide étude de la fonction $x(1-x)$.

Étant donné que les majorations (38.558) sont valables pour tout x, en passant au supremum nous avons

$$\|f - B_n(f)\|_\infty \leqslant \frac{3}{2}\omega\left(\frac{1}{\sqrt{n}}\right) \to 0. \tag{38.559}$$

Ceci prouve les deux premiers points du théorème.

(2) Fait.

(3) Nous considérons la fonction

$$g(x) = \|x - \frac{1}{2}j\| \tag{38.560}$$

et nous vérifions qu'elle vérifie toutes les conditions. D'abord si $u, v \in [0,1]$ alors

$$\big|g(u) - g(v)\big| \leqslant |u - v| \tag{38.561}$$

et donc $\omega(h) \leqslant h$, ce qui signifie que g est 1-Lipschitz. Le principe de cette partie est de montrer que $\|g - B_n(g)\|_\infty$ est plus grand que d'autres trucs (et non plus petit que d'autres trucs comme d'habitude). Nous commençons par

$$\|g - B_n(g)\|_\infty \geqslant g(\frac{1}{2}) - B_n(g)(\frac{1}{2}). \tag{38.562}$$

Très vite nous nous rendons compte que $g(1/2) = 0$. Ensuite nous nous souvenons que

$$B_n(g)(\frac{1}{2}) = E\left(g(\frac{S_n}{n})\right) = E\left(|\frac{S_n}{n} - \frac{1}{2}|\right) = \frac{1}{2n}E\big(|2S_n - n|\big). \tag{38.563}$$

si nous posons $\epsilon_i = 2X_i - 1$, alors les ϵ_i sont des variables aléatoires de Rademacher indépendantes et identiquement distribuées qui satisfont à $2S_n - n = \sum_{i=1}^n \epsilon_i$. Nous utilisons la proposition 38.115 :

$$\|f - B_n(f)_\infty\| \geqslant \frac{1}{2n}E\big(|\sum_i \epsilon_i|\big) \geqslant \frac{1}{2n\sqrt{e}}\|\sum_{i=1}^n \epsilon_j\|_2. \tag{38.564}$$

Calculons ce qui est dans la norme :

$$\|\sum_{j=1}^n \epsilon_j\|_2^2 = E\left(\big(\sum_{j=1}^n \epsilon_j\big)^2\right) = \sum_{1\leqslant i\neq j\leqslant n} E(\epsilon_i)E(\epsilon_j) + \sum_{i=1}^n E(\epsilon_i^2) = 0 + n = n. \tag{38.565}$$

Nous finissons alors notre travail de majoration :

$$\|f - B_n(f)_\infty\| \geqslant \frac{1}{2n\sqrt{e}}\|\sum_{i=1}^n \epsilon_j\|_2 \geqslant \frac{1}{2\sqrt{n}\sqrt{e}} \geqslant \frac{1}{2\sqrt{e}}\omega\left(\frac{1}{\sqrt{n}}\right). \tag{38.566}$$

□

Chapitre 39

Statistiques

39.1 Notations et hypothèses

Nous notons X le caractère à étudier, et Ω l'ensemble des individus. Le caractère à étudier est vu comme une fonction sur Ω :

$$X : \Omega \to \mathbb{R}, \mathbb{N}, \{0, 1\}, \ldots \tag{39.1}$$

Les **statistiques descriptives** sont les techniques pour présenter et résumer les données : diagrammes, graphiques, indicateurs numériques : moyenne, écart-type, médiane, ...

Nous faisons les hypothèses suivantes :

(1) Chaque observation x_i est la réalisation de la variable aléatoire X qui sera de loi inconnue μ.
(2) Le n-uple $(x_1, \ldots, x_n)$ est la réalisation de $(X_1, \ldots, X_n)$ qui est l'échantillon de taille n.
(3) Les variables aléatoires X_i sont indépendantes et identiquement distribuées, de loi commune μ. La loi μ est la **loi parente** de l'échantillon.

Exemple 39.1
Un échantillon de taille 1 consisterait à tirer au sort une personne dans une population et mesurer sa taille. △

Exemple 39.2
Une échantillon de taille n consisterait à tirer au sort n personnes dans une population et de mesurer leurs tailles. △

L'**inférence statistique** est l'art de dégager des informations sur la population à partir d'informations partielles : intervalles de confiance, estimateurs, test d'hypothèses, ...

En théorie des probabilités, nous connaissons la loi de la variable aléatoire X et nous en déduisons des informations sur le réalisations de X : valeur la plus probable, moyenne, intervalle dans lequel $X(\omega)$ a le plus de chance d'appartenir. En statistique, au contraire, la loi est inconnues et nous cherchons des informations sur la loi à partir d'un échantillon de données numériques observées.

39.2 Modèle statistique

Une **modèle statistique** est un triplet

$$\mathcal{S} = \Big[(\Omega, \mathcal{F}, P), (X_\theta)_{\theta\in\Theta}, (\mu_\theta)_{\theta\in\Theta}\Big] \tag{39.2}$$

où $(\Omega, \mathcal{F}, P)$ est un espace probabilisé, (X_θ) est une famille de variables aléatoires définies sur Ω et telles que pour tout $\theta \in \Theta$, la variable aléatoire X_θ suive la loi μ_θ. Les μ_θ sont des mesures sur les boréliens de $\mathbb{R}$ et pour tout $B \in \mathcal{B}\text{or}(\mathbb{R})$ nous avons

$$P(X_\theta \in B) = \mu_\theta(B). \tag{39.3}$$

Remarque 39.3.
D'une certaine manière, l'introduction de μ_θ dans la définition est redondante parce que ces mesures sont déjà contenues dans la données des variables aléatoires X_θ.

Exemple 39.4(Modèle statistique gaussien)
Si nous savons que les variables aléatoires X_i suivent une loi gaussienne, alors nous considérons $\Theta = \mathbb{R} \times \mathbb{R}^+$ et $\theta = (m, \sigma^2 j)$. Dans ce cas, $\mu_\theta = \mathcal{N}(m, \sigma^2)$ et le but de la statistique est de déterminer la valeur de θ qui correspond à une population en partant de l'observation d'un échantillon. △

Définition 39.5.
Si $\Theta \subset \mathbb{R}^k$, nous disons que le modèle statistique est un modèle ***paramétrique****.*

Le modèle gaussien est un modèle paramétrique : dès que m et σ^2 sont déterminés, la loi du phénomène X est connue.

Définition 39.6.
Pour chaque $\theta \in \theta$, nous disons qu'un ***échantillon*** *de taille n associé à un modèle statistique $\big[(\Omega, \mathcal{F}, P), (X_\theta), (\mu_\theta)\big]$ est un vecteur $\big(X_{\theta,1}, \dots, X_{\theta,n}\big)$ de taille n de variables aléatoires indépendantes et identiquement distribuées de la même loi que la variable aléatoire X_θ. La loi μ_θ est la* ***loi parente*** *de l'échantillon.*

Définition 39.7.
Un ***modèle d'échantillonnage*** *sur le modèle statistique $\mathcal{S}$ est une famille $(X_{\theta,1}, \dots, X_{\theta,n})_{\theta\in\Theta}$ d'échantillons de taille $n \geqslant 1$.*

Nous noterons souvent $(X_1, \dots, X_n)$ à la place de $(X_{\theta,1}, \dots, X_{\theta,n})$ un échantillon, mais il faut se souvenir que les X_i suivent toujours la même loi donnée par θ. La loi du vecteur $(X_1, \dots, X_n)$ est $\mu_\theta \otimes \dots \otimes \mu_\theta$ et est définie sur l'espace $(\Omega^n, \mathcal{F} \otimes \dots \otimes \mathcal{F}, P^{\otimes n})$.

Remarque 39.8.
Le travail du statisticien est de proposer un modèle statistique $\mathcal{S}$ a priori. Si nous étudions la taille d'une population, nous allons choisir un modèle gaussien. Plus le modèle est précis, plus l'espace Θ est petit mais plus il y a de risques que le vérité soit hors de l'ensemble considéré.

Exemple 39.9
Soit X une variable aléatoire de carré intégrable que l'on sait simuler. Afin d'évaluer la moyenne μ de X, nous pouvons considérer la moyenne empirique des simulations : $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ où les variables aléatoires X_i sont indépendantes, identiquement distribuées et de même loi que X.

La loi des grands nombres nous enseigne que $\bar{X}_n \to \mu$. De plus,

$$\lim_{n\to\infty} P\left(\mu \in \left[\bar{X}_n - \frac{a\sigma}{\sqrt{n}}, \bar{X}_n + \frac{a\sigma}{\sqrt{n}}\right]\right) = \int_{-a}^{a} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}. \tag{39.4}$$

En effet, la condition sur μ est équivalente à

$$-a \leqslant \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant a, \tag{39.5}$$

tandis que le théorème central limite nous enseigne que la variable aléatoire $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$ se comporte comme $\mathcal{N}(0,1)$ lorsque n est grand. Dans ce cas, nous avons que

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \in [-a, a]\right) = \frac{1}{\sqrt{2\pi}}\int_{-a}^{a} e^{-x^2/2}dx. \tag{39.6}$$

Notons que dans ce calcul nous avons utilisé le fait que $\mu = E(X_1)$.

Montrons que la suite

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \tag{39.7}$$

converge presque sûrement vers σ^2. Le théorème central limite implique que

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p.s.} E(X_1^2) \tag{39.8}$$

et que

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \xrightarrow{p.s.} E(X_1)^2. \tag{39.9}$$

La différence converge donc presque sûrement vers σ^2 en vertu de la proposition 38.26.

Nous avons également $E(\sigma_n^2) = \sigma^2$. En effet, sachant que $E(X_i) = E(X_1) = \mu$ et que $E(X_i^2) = E(X_1^2) = \mu^2 + \sigma^2$,

$$E(\sigma_n^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - \frac{1}{n^2}\left(\sum_{i=1}^{n} E(X_i^2)\right) + \sum_{i\neq j} E(X_iX_j)) \tag{39.10a}$$

$$= \sigma^2 + \mu^2 - \frac{1}{n^2}\big(n(\sigma^2+\mu^2) + (n^2-n)\mu^2\big) \tag{39.10b}$$

$$= \sigma^2 - \frac{1}{n}\sigma^2, \tag{39.10c}$$

dont la limite $n \to \infty$ donne bien σ^2.

Nous voudrions à présent montrer que

$$\frac{\bar{X}_n - \mu}{\sigma_n/\sqrt{n}} \xrightarrow{\mathscr{L}} \mathscr{N}(0,1). \tag{39.11}$$

Vu que le théorème central limite donne une convergence en loi, nous pouvons utiliser le lemme de Slutsky pour montrer que

$$\left(\frac{\bar{X}_n - \mu}{1/\sqrt{n}}, \sigma_n^2\right) \xrightarrow{\mathscr{L}} (\sigma Z, \sigma^2) \tag{39.12}$$

où $Z \sim \mathscr{N}(0,1)$. En vertu de la proposition 38.75 appliqué à la fonction $f\colon \mathbb{R}^2 \to \mathbb{R}$,

$$f(x,y) = \begin{cases} \frac{x}{\sqrt{y}} & \text{si } y \neq 0 \\ 0 & \text{sinon} \end{cases} \tag{39.13}$$

nous avons la convergence en loi

$$f\left(\frac{\bar{X}_n - \mu}{1/\sqrt{n}}, \sigma_n^2\right) \xrightarrow{\mathscr{L}} f(\sigma Z, \sigma^2), \tag{39.14}$$

c'est à dire

$$\frac{\bar{X}_n - \mu}{\sigma_n/\sqrt{n}} \xrightarrow{\mathscr{L}} Z. \tag{39.15}$$

Afin d'être complet, précisons que

$$P\big((\sigma Z, \sigma) \in \mathbb{R} \times \{0\}\big) = 0. \tag{39.16}$$

△

Proposition 39.10.
Soit X_i des variables aléatoires indépendantes et identiquement distribuées de loi parente $\mathscr{B}(n,p)$. Alors si $\bar{X}_n$ désigne la moyenne empirique,

$$\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \xrightarrow{\mathscr{L}} Z \sim \mathscr{N}(0,1). \tag{39.17}$$

Démonstration. Cela est une application de la loi des grands nombres, tu théorème central limite, du lemme de Slutsky et de la proposition 38.75.

D'abord, la loi des grands nombres nous indique que $\bar{X}_n \to p$ parce que p est l'espérance de Bernoulli. Ensuite nous avons

$$\frac{\bar{X}_n - E(X)}{\sqrt{\text{Var}(X)}/\sqrt{n}} = \sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \tag{39.18}$$

parce que la variance d'une loi de Bernoulli est $p(1-p)$. Le théorème central limite nous indique par conséquent que

$$\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{\mathscr{L}} Z \sim \mathscr{N}(0,1). \tag{39.19}$$

Le lemme de Slutsky implique alors la convergence du couple :

$$\left(\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}}, \bar{X}_n\right) \xrightarrow{\mathscr{L}} (Z,p). \tag{39.20}$$

Nous appliquons maintenant la proposition 38.75 avec la fonction

$$f(x,y) = \frac{\sqrt{p(1-p)}x}{\sqrt{y(1-y)}} \tag{39.21}$$

qui est une fonction dont l'ensemble des points de discontinuité est $C = \{0\}$. Étant donné que $P(\bar{X}_n = 0) = 0$, la proposition s'applique et nous avons

$$f\left(\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}}, \bar{X}_n\right) \xrightarrow{\mathscr{L}} f(Z,p), \tag{39.22}$$

c'est à dire

$$\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \xrightarrow{\mathscr{L}} Z \sim \mathscr{N}(0,1). \tag{39.23}$$

□

39.3 Modèles d'échantillonnages

Soit X, une variable aléatoire sur $(\Omega, \mathcal{F}, P)$. Un **échantillon** de taille n pour X est une suite de n variables aléatoires $(X_1, \dots, X_n)$ définies sur $(\Omega, \mathcal{F}, P)$ indépendantes et de même loi que X. Nous disons que la loi de X est la **loi parente** de la suite X_i.

Définition 39.11.
Soit

$$\mathcal{S} = \Big[(\Omega, \mathcal{F}, P), (X_\theta), (\mu_\theta)\Big]_{\theta\in\Theta}, \tag{39.24}$$

*un modèle statistique. Un **modèle d'échantillonnage** de taille n associée au modèle statistique $\mathcal{S}$ est la donnée d'une famille de n-échantillons $(X_{\theta,1}), \dots, X_{\theta,n}$ telle que pour tout $\theta \in \Theta$, l'échantillon $(X_{\theta,i})$ soit de variable parente X_θ.*

La **moyenne empirique** du n-échantillon (X_i) est la variable aléatoire

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{39.25}$$

La proposition suivante signifie que la moyenne empirique est une « bonne » façon d'approcher la variable aléatoire.

Proposition 39.12.
Soit X, une variable aléatoire dans $L^2(\Omega)$ (c'est à dire $E(X^2) < \infty$) d'espérance m et de variance σ^2. Alors

(1) $E(\bar{X}_n) = m$ et $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$.

(2) Nous avons les convergences

$$\bar{X}_n \xrightarrow{p.s.} m \tag{39.26a}$$

$$\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \xrightarrow{\mathscr{L}} Z \sim \mathscr{N}(0,1). \tag{39.26b}$$

(3) Si X est de loi $\mathscr{N}(m,\sigma^2)$, alors $\bar{X}_n \sim \mathscr{N}(m, \frac{\sigma^2}{n})$, c'est à dire

$$\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \sim \mathscr{N}(0,1). \tag{39.27}$$

Remarque 39.13.
L'intérêt de cette proposition en statistique descriptive expérimentale est le suivant. La taille moyenne des français est un nombre m qui existe, mais qui est largement hors de portée de l'expérience (mesurer $65 \cdot 10^6$ personnes risque de prendre un sacré temps). Si on mesure seulement n personnes dont les tailles sont $(x_i)_{i=1,\dots,n}$ (ici x_i est un *nombre expérimental*, pas une variable aléatoire), alors on peut calculer la moyenne $\bar{x}_n$ de ces n personnes-là. La proposition indique que si n est assez grand, alors $\bar{x}_n$ donne une bonne idée de m.

Ne pas confondre X_n qui est une variable aléatoire, c'est à dire une application mesurable, qui nous sert à démontrer des théorèmes en mathématique, avec x_n qui est un nombre mesuré sur le terrain, qui a une existence *physique* bien définie, mais aucun status mathématique.

Si on croit que toute cette histoire de variables aléatoires, de tribu et de mesures décrit effectivement la réalité, alors on peut croire que le comportement de la suite $\bar{X}_n$ décrit bien le comportement de la suite $\bar{x}_n$ (cette dernière n'étant même pas une suite parce qu'on n'a jamais qu'un nombre fini de mesures expérimentales).

La **variance empirique** d'un échantillon est la variable aléatoire

$$V_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2. \tag{39.28}$$

La **variance empirique corrigée** est la variable aléatoire

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2. \tag{39.29}$$

Lemme 39.14.
La variance corrigée et la variance empirique ont comme espérances :

$$E(V_n^2) = \frac{n-1}{n}\sigma^2 \tag{39.30a}$$

$$E(S_n^2) = \sigma^2. \tag{39.30b}$$

Démonstration. Nous commençons par calculer l'espérance de la variance non corrigée. La première étape est de la récrire sous la forme

$$\begin{aligned} V_n^2 &= \frac{1}{n}\sum_k (X_k^2 - 2X_k\bar{X}_n + \bar{X}_n^2) \\ &= \frac{1}{n}\sum_k X_k^2 - \frac{2}{n}\bar{X}_n \underbrace{\sum_k X_k}_{=n\bar{X}_n} + \frac{1}{n}\underbrace{\sum_k \bar{X}_n^2}_{=n\bar{X}_n^2} \\ &= \frac{1}{n}\sum_k X_k^2 - 2\bar{X}_n^2 + \bar{X}_n^2 \\ &= \frac{1}{n}\sum_k X_k^2 - \bar{X}_n^2. \end{aligned} \tag{39.31}$$

Nous calculons séparément l'espérance de ces deux termes. Si X est la loi parente des X_i, en utilisant l'indépendance des X_i nous trouvons

$$\begin{aligned} E\left(\frac{1}{n}\sum_k X_k^2\right) &= \frac{1}{n}\sum_{k=1}^n E(X_k^2) \\ &= E(X^2) \\ &= E(X)^2 - \operatorname{Var}(X). \end{aligned} \tag{39.32}$$

Nous devons à présent calculer l'espérance de $\bar{X}_n^2$:

$$E(\bar{X}_n^2) = E(\bar{X}_n)^2 + \operatorname{Var}(\bar{X}_n). \tag{39.33}$$

En utilisant le lemme 38.23,

$$\operatorname{Var}(\bar{X}_n) = \operatorname{Var}\left(\frac{1}{n}\sum_k X_k\right) \tag{39.34a}$$

$$= \frac{1}{n^2}\sum_k \operatorname{Var}(X_k) \tag{39.34b}$$

$$= \frac{1}{n}\operatorname{Var}(X). \tag{39.34c}$$

Par conséquent

$$E(V_n^2) = \operatorname{Var}(X)\left(1 - \frac{1}{n}\right) = \frac{n-1}{n}\operatorname{Var}(X). \tag{39.35}$$

En ce qui concerne la variance corrigée,

$$S_n^2 = \frac{1}{n-1}\sum_k (X_k - \bar{X}_n)^2 = \frac{n}{n-1}V_n^2, \tag{39.36}$$

par conséquent $E(S_n^2) = \frac{n}{n-1}E(V_n^2) = \operatorname{Var}(X)$. □

Théorème 39.15 (Théorème de Cochran[195])**.**
Soient (X_i) des variables aléatoires gaussiennes indépendantes de loi $\mathcal{N}(m, \sigma^2)$ avec $\sigma > 0$. Alors

(1) $\bar{X}_n \sim \mathcal{N}(m, \frac{\sigma^2}{n})$,

(2) $\left(\frac{n-1}{\sigma^2}\right) S_n^2 = \left(\frac{n}{\sigma^2}\right) \bar{V}_n^2 \sim \chi^2(n-1)$,

(3) les variables aléatoires $\bar{X}_n$ et $\bar{V}_n$ sont indépendantes et

$$\frac{\bar{X}_n - m}{S_n/\sqrt{n}} = \frac{\bar{X}_n - m}{\sqrt{\bar{V}_n/(n-1)}} \sim \mathcal{T}(n-1). \tag{39.37}$$

Nous pouvons aussi écrire le dernier résultat en termes de la variance corrigée S_n, l'estimateur sans biais de la variance parce que

$$\sqrt{\frac{\bar{V}_n}{n-1}} = \frac{1}{\sqrt{n}} S_n \tag{39.38}$$

en vertu de la définition (39.29).

Proposition 39.16.
Soit X une variable aléatoire de variance $\mathrm{Var}(X) = \sigma$. Si $E(X^4) < \infty$, alors

(1) $S_n^2 \xrightarrow{p.s.} \sigma^2$.

(2)

$$\frac{S_n^2 - \sigma^2}{\sqrt{\frac{\mu^4 - \sigma^4}{n}}} \xrightarrow{\mathscr{L}} Z \sim \mathscr{N}(0,1) \tag{39.39}$$

où $\mu^4 = E(X^4)$ est le moment d'ordre 4 de X.

Théorème 39.17.
Si $(X_1, \dots, X_n)$ est un n-échantillon de loi parente $\mathscr{N}(m, \sigma^2)$, alors

(1) Les variables aléatoires

$$\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \quad et \quad (n-1)\frac{S_n^2}{\sigma^2} \tag{39.40}$$

sont indépendantes.

(2) La loi de $(n-1)\frac{S_n^2}{\sigma^2}$ est $\chi^2(n-1)$.

Si un échantillon vérifie ces deux propriétés, alors les X_i sont de loi $\mathscr{N}(m, \sigma^2)$.

L'inégalité de Markov donne une borne supérieure à la probabilité qu'une variable aléatoire positive soit plus grande ou égale à une constante.

Théorème 39.18 (Inégalité de Markov)**.**
Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$ et $\varphi\colon \mathbb{R}^d \to [0, \infty[$. Alors

$$P\big(\varphi(X) \geqslant a\big) \leqslant \frac{E\big(\varphi(X)\big)}{a} \tag{39.41}$$

pour tout $a > 0$.

Démonstration. Calculons le second membre :

$$\begin{aligned}
\frac{E\big(\varphi(X)\big)}{a} &= \int_\Omega \frac{\varphi(X)}{a} dP \\
&= \int_{\varphi(X)\geqslant a} \underbrace{\frac{\varphi(X)}{a}}_{\leqslant 1} dP + \int_{\varphi(X)<a} \frac{\varphi(X)}{a} dP \\
&\geqslant \int_{\varphi(X)\leqslant a} dP \\
&= P\big(\varphi(X) \leqslant a\big).
\end{aligned} \tag{39.42}$$

D'où l'inégalité voulue. □

39.4 Estimation ponctuelle

Nous considérons un modèle statistique

$$\mathcal{S} = \big[(\Omega, \mathcal{F}, P), (X_\theta), (\mu_\theta)\big]_{\theta\in\Theta} \tag{39.43}$$

et pour tout θ nous notons $X = (X_{\theta,1}, \dots, X_{\theta,n})$ un échantillon de loi parente μ_θ. Tant que nous travaillerons avec un θ fixé, nous écrirons $X = (X_1, \dots, X_n)$ sans expliciter la paramètre θ. Nous noterons

$$E_\theta\big(\varphi(X_1, \dots, X_n)\big) = \int_{\mathbb{R}^n} \varphi(x_1, \dots, x_n) d\mu_\theta^{\otimes n}(x_1, \dots, x_n). \tag{39.44}$$

Dans cette notation nous plaçons le θ sur l'espérance, tandis qu'en réalité le θ devrait être sur chaque X_1. Tant qu'aucune confusion n'est possible nous ferons toujours cet abus d'écriture.

Le but de la théorie de l'estimation est de déduire la valeur de θ (et donc la loi μ_θ) à partir d'un échantillon de loi parente θ.

Nous posons les hypothèses suivantes.

(1) Le modèle statistique $\mathcal{S}$ est paramétré, c'est à dire que $\Theta \subset \mathbb{R}^d$ avec le plus souvent $d = 1, 2$. Typiquement les paramètres seront la moyenne et la variance.

(2) Le modèle statistique est **identifiable**, c'est à dire que pour tout couple $(\theta_1, \theta_2) \in \Theta^2$, si $\theta_1 \neq \theta_2$, alors $\mu_{\theta_1} \neq \mu_{\theta_2}$.

(3) Le modèle $\mathcal{S}$ est **dominé** par la mesure de Lebesgue si les lois μ_θ sont continues et par la mesure de comptage si les lois μ_θ sont discrètes.

Exemple 39.19
Quelque familles identifiables :

— La famille des lois exponentielles $\big(\mathscr{E}(\lambda)\big)_{\lambda>0}$ est identifiable.

— Les lois gaussiennes $\big(\mathscr{N}(m, \sigma^2)\big)_{m\in\mathbb{R},\sigma^2>0}$ sont également identifiables.

En réalité il est assez compliqué de trouver un exemple de modèle non identifiable à moins de la faire exprès. Par exemple en paramétrant les lois exponentielles de la façon suivante : $\big(\mathscr{E}(\sin(\lambda))\big)_{\lambda\in\mathbb{R}}$. Cette famille n'est pas identifiable. △

Le corollaire 16.152 ainsi que l'hypothèse de modèle dominé implique que les lois ont des densités. Si la loi μ_θ est discrète, nous notons

$$p(x, \theta) = \mu_\theta(\{x\}) \tag{39.45}$$

la densité de μ_θ par rapport à la mesure de comptage. Si La loi μ_θ est continue, nous notons

$$p(x, \theta) = f_\theta(x) \tag{39.46}$$

la densité par rapport à la mesure de Lebesgue.

Si μ_θ est une loi discrète et si $(X_1, \dots, X_n)$ est un échantillon de taille n, alors pour tout $(x_1, \dots, x_n) \in \mathbb{R}^n$ nous avons

$$p_n(x_1, \dots, x_n; \theta) = \mu_\theta^{\otimes n}\big(\{x_1, \dots, x_n\}\big) = \mu_\theta(\{x_1\}) \dots \mu_\theta(\{x_n\}) = p(x_1, \theta) \dots p(x_n, \theta). \tag{39.47}$$

La première et la dernière égalité sont des notations ; la seconde est une conséquence de l'indépendance des X_i contenues dans l'échantillon. Pour une loi continues, nous adoptons la même notation. Le vecteur aléatoire $(X_1, \dots, X_n)$ admet la densité

$$(x_1, \dots, x_n) \mapsto p_n(x_1, \dots, x_n; \theta) = f_\theta(x_1) \dots f_\theta(x_n) = p(x_1, \theta) \dots p(x_n, \theta). \tag{39.48}$$

Exemple 39.20
Soit $(X_1, \dots, X_n)$ un n-échantillon de loi $\mathscr{B}(1, \theta)$ avec $\theta \in]0, 1[$. C'est une loi discrète portée par l'ensemble $\{0, 1\}$. Nous avons

$$p(x, \theta) = \begin{cases} 0 & \text{si } 0 \neq x \neq 1 \\ 1 - \theta & \text{si } x = 0 \\ \theta & \text{si } x = 1. \end{cases} \tag{39.49}$$

De façon plus condensée nous pouvons écrire

$$p(x,\theta)=\theta^x(1-\theta)^{1-x}\mathbb{1}_{\{0,1\}}(x). \tag{39.50}$$

Pour tout $(x_1,\dots,x_n)\in\mathbb{R}^p$, la densité du n-échantillon est donnée par

$$p_n(x_1,\dots,x_n;\theta)=\theta^{x_1+\cdots+x_n}(1-p)^{1-\sum_i(1-x_i)}\mathbb{1}_{\{0,1\}^n}(x_1,\dots,x_n). \tag{39.51}$$

△

39.5 Statistiques et estimateurs

Définition 39.21.
*Une **statistique** sur un modèle d'échantillonnage est une variable aléatoire fonction de l'échantillon $(X_1,\dots,X_n)$ ne dépendant pas de θ*[1]*. C'est à dire une application borélienne $T\colon\mathbb{R}^n\to\mathbb{R}$ ne dépendant pas de θ. La statistique associée à cette application est $S=T(X_1,\dots,X_n)$.*

Les fonctions $T(X_1,\dots,X_n)$ données par $\sum_i X_i$, $e^{\sum X_i}$ sont des statistiques. La constante $\frac{1}{2}$ est également une statistique (mais elle est moins intéressante).

Un **estimateur** est une statistique qui prend ses valeurs dans Θ. Nous la noterons

$$\hat{\theta}_n=\theta(X_1,\dots,X_n). \tag{39.52}$$

La fonction $\hat{\theta}_n$ est borélienne à valeurs dans Θ.

Exemple 39.22
Soit un n-échantillon de loi $\mathscr{B}(1,\theta)_{\theta\in[0,1]}$. Les fonctions $\hat{\theta}_n=\frac{1}{n}\sum_{i=1}^n X_i$ et $\hat{\varphi}_n=\frac{1}{2}$ sont des estimateurs. Cependant nous devinons que la première va être plus intéressante que la seconde.
△

Pour la suite, nous travaillerons avec des estimateurs de carré intégrable, c'est à dire que

$$E_\theta\big(|\hat{\theta}_n(X_1,\dots,X_n)|^2\big)<\infty \tag{39.53}$$

pour tout $\theta\in\Theta$.

39.5.1 Qualité des estimateurs

Définition 39.23.
*Une estimateur est **convergent** ou **consistant** si pour tout $\theta\in\Theta$, la suite de variables aléatoires $\hat{\theta}_n(X_1,\dots,X_n)$ converge en probabilité vers θ.*

En d'autres termes, l'estimateur $\hat{\theta}_n$ est convergeant si pour tout $\theta\in\Theta$ et pour tout $\eta>0$,

$$\lim_{n\to\infty}P\big(|\hat{\theta}_n(X_1,\dots,X_n)-\theta|>\eta\big)=0. \tag{39.54}$$

La probabilité dans le membre de gauche est donnée par

$$\mu_\theta^{\otimes n}\Big(\big\{(x_1,\dots,x_n)\in\mathbb{R}^n\text{ tel que }|\hat{\theta}_n(x_1,\dots,x_n)-\theta|>\eta\big\}\Big). \tag{39.55}$$

Soit $\hat{\theta}_n$ un estimateur pour θ. Nous cherchons à minimiser l'erreur commise en remplaçant θ par $\hat{\theta}_n$. Nous introduisons donc le **risque quadratique** de l'estimateur $\hat{\theta}_n$ par

$$R(\hat{\theta}_n,\theta)=E_\theta\big((\hat{\theta}_n-\theta)^2\big). \tag{39.56}$$

Nous disons qu'un estimateur $\hat{\theta}_{n,1}$ est préférable à $\hat{\theta}_{n,2}$ si pour tout $\theta\in\Theta$ nous avons

$$R(\hat{\theta}_{n,1},\theta)<R(\hat{\theta}_{n,2},\theta). \tag{39.57}$$

1. Parce que d'habitude c'est ce qu'on cherche à estimer.

Lemme 39.24.
Une formule alternative pour le risque quadratique :

$$R(\hat{\theta}_n, \theta) = \mathrm{Var}(\hat{\theta}_n) + \big(E_\theta(\hat{\theta}_n) - \theta\big)^2 \tag{39.58}$$

Démonstration. Nous avons

$$E_\theta\Big((\hat{\theta}_n - \theta)^2\Big) = \mathrm{Var}(\hat{\theta}_n - \theta) - E_\theta(\hat{\theta}_n - \theta)^2. \tag{39.59}$$

D'une part $\mathrm{Var}(\hat{\theta}_n - \theta) = \mathrm{Var}(\hat{\theta}_n)$ et d'autre part $E_\theta(\hat{\theta}_n - \theta)^2 = \big[E(\hat{\theta}_n) - \theta\big]^2$. Par conséquent

$$R(\hat{\theta}_n, \theta) = \mathrm{Var}(\hat{\theta}_n) - \Big(E(\hat{\theta}_n) - \theta\Big)^2. \tag{39.60}$$

□

Le **biais** de l'estimateur $\hat{\theta}_n$ est la quantité

$$E_\theta(\hat{\theta}_n) - \theta. \tag{39.61}$$

À ce niveau, nous rappelons que nous écrivons E_θ l'espérance calculée en supposant la valeur θ pour le paramètre des différentes variables aléatoires entrant dans le calcul. Voir la discussion autour de la définition (39.44).

Exemple 39.25
Dans le cadre de la proposition 38.103, nous voulons savoir si $\frac{N_t}{t}$ est un estimateur sans biais de λ. Pour ce faire nous calculons

$$E_\lambda\left(\frac{N_t}{t}\right) = \frac{1}{t}E_\lambda(N_t) = \lambda \tag{39.62}$$

parce que $E(N_t) = \lambda t$. Ici nous avons calculé $E(N_t)$ en prenant λ pour valeur du paramètre du processus de Poisson, alors que en principe c'est justement le paramètre que nous voulons estimer. △

Exemple 39.26
La moyenne empirique $\bar{X}_n$ est un estimateur sans billet de la moyenne. L'estimateur

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_i)^2 \tag{39.63}$$

est un estimateur sans billet de la variance. △

Un estimateur sans biais n'est pas toujours de meilleur qualité qu'un estimateur avec biais. En effet ce que nous voulons est de se donner un (petit) intervalle I autour de la bonne valeur de θ et de maximiser $P(\hat{\theta}_n \in I)$. Sur la figure 39.1, c'est l'estimateur biaisé rouge tombe plus souvent sur le bon intervalle que l'estimateur non biaisé bleu.

Nous allons maintenant étudier quelques manières de construire des estimateurs convergeant [2]. Il vont évidemment s'appuyer sur la loi des grands nombres.

39.5.2 Méthode des moments

Sans surprises, un bon estimateur pour la moyenne est

$$\hat{\theta}_n(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{39.64}$$

2. Là je ne suis vraiment pas sûr de l'orthographe. Comme partout où j'utilise le verbe « converger » d'ailleurs.

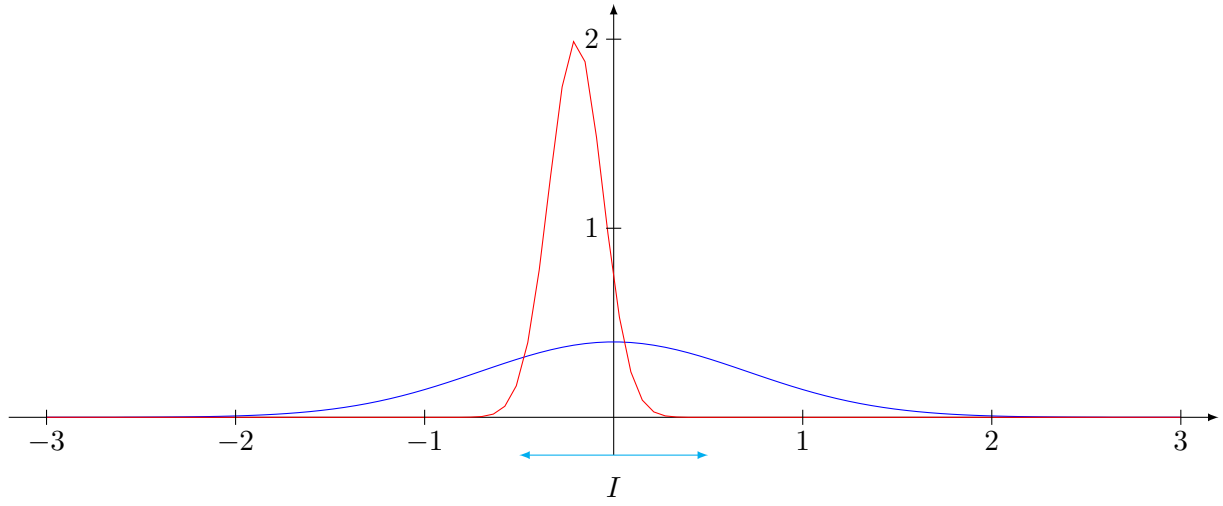


FIGURE 39.1 – Un estimateur sans biais et un avec biais.

Plus généralement, nous supposons qu'il existe une fonction borélienne[3] $M\colon \mathbb{R}\to\mathbb{R}$ telle que

$$E_\theta\big(|M(X)|\big)<\infty \tag{39.65}$$

où X est la loi parente de l'échantillon. Supposons également que la fonction

$$h(\theta)=E_\theta\big(M(X)\big) \tag{39.66}$$

soit inversible et continue sur Θ. Dans ce cas, pour estimer le paramètre θ, nous considérons l'estimateur

$$\hat{\theta}_n=h^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}M(X_i)\right). \tag{39.67}$$

Cela est un estimateur convergeant. En effet, la loi des grands nombres dit que

$$\frac{1}{n}\sum_i M(X_i)\xrightarrow{p.s.}E_\theta\big(M(X)\big). \tag{39.68}$$

En composant avec la fonction h, nous avons

$$\hat{\theta}_n\xrightarrow{p.s.}h^{-1}\Big(E_\theta\big(M(X)\big)\Big)=\theta. \tag{39.69}$$

Dans cette construction, $M(X)$ est le moment de X que l'on souhaite déterminer.

Exemple 39.27
Soit $(X_1,\dots,X_n)$ un échantillon de loi $\mathscr{E}(\lambda)$. Construire $\hat{\lambda}_n$. Pour une loi exponentielle,

$$E(X)=\frac{1}{\lambda}. \tag{39.70}$$

Nous devons donc déterminer le moment d'ordre 1 de X (c'est à dire sa moyenne). Nous considérons donc la fonction $M(x)=x$; par conséquent

$$h(\lambda)=E(X)=\frac{1}{\lambda} \tag{39.71}$$

et $h^{-1}(\theta)=1/\theta$. L'estimateur que nous considérons pour λ est finalement

$$\hat{\theta}_n=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}X_i}. \tag{39.72}$$

△

3. Définition 16.64.

39.5.3 Méthode de substitution

Supposons que nous connaissions un estimateur convergeant $\hat{\theta}_n \to \theta$. Si $g\colon \mathbb{R}^d \to \mathbb{R}$ est une fonction continue, alors

$$g(\hat{\theta}_n) \to g(\theta). \tag{39.73}$$

39.5.4 Méthode du maximum de vraisemblance

Exemple 39.28
Nous désirons contrôler la qualité d'une chaîne de production ; pour cela nous prélevons un échantillon de 10 pièces, et nous en trouvons 3 défectueuses. Que dire de la proportion de pièces défectueuses ?

Évidemment, le plus probable est que la proportion de pièces défectueuses soit de 1/3. Analysons en détail comment nous arrivons à ce résultat. Nous considérons que le fait de tirer 10 pièces revient à une expérience binomiale de paramètres 10 et de probabilité p inconnue. Dans ce cas, la probabilité d'observer exactement 3 pièces défectueuses est de

$$L(p) = P(X = 3) = \binom{10}{3} p^3 (1-p)^7. \tag{39.74}$$

Le maximum de $L(p)$ est $p = 3/10$.

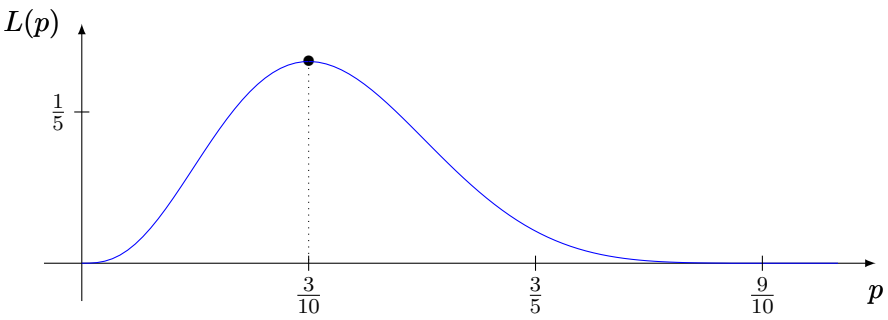


FIGURE 39.2 – La fonction de vraisemblance de l'exemple 39.28.

△

Soit $(x_1, \dots, x_n)$, une réalisation de l'échantillon $(X_1, \dots, X_n)$. L'application

$$\theta \mapsto p_n(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} p(x_i, \theta) \tag{39.75}$$

est la **vraisemblance** de l'échantillon. Nous définissons $\hat{\theta}_n$ par

$$p_n(x_1, \dots, x_n; \hat{\theta}_n) = \sup_{\theta \in \Theta} p_n(x_1, \dots, x_n; \theta). \tag{39.76}$$

Remarque 39.29.
Nous passons sous le silence le fait que la fonction sup soit une fonction mesurable, et que par conséquent $\hat{\theta}_n$ soit bien une variable aléatoire.

La variable aléatoire $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ est l'**estimateur de maximum de vraisemblance** de θ.

39.5.5 Exemples sous forme d'exercices

Exemple 39.30
Soit $(X_1, \dots, X_n)$ un échantillon de loi $\mathscr{B}(1, \theta)$ avec $\theta \in [0, 1]$. Trouver l'estimateur de maximum de vraisemblance de θ.

En utilisant la densité de la loi multinomiale,

$$p_n(x_1, \dots, x_n; \theta) = \prod_i p(x_i, \theta) = \prod_i \theta^{x_i}(1-\theta)^{1-x_i} \mathbb{1}_{\{0,1\}^n}(x_1, \dots, x_n). \tag{39.77}$$

En passant au logarithme, si $x_i \in \{0, 1\}$,

$$L_n(\theta) = \sum_i x_i \ln(\theta) + (1 - x_i) \ln(1 - \theta) \tag{39.78}$$

En passant à la dérivée, nous trouvons que l'extremum est donné par

$$\theta = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{39.79}$$

△

Exemple 39.31

Soit $(X_1, \dots, X_n)$ un échantillon de loi $\mathscr{E}(\lambda)$ avec $\theta > 0$. Trouver l'estimateur de maximum de vraisemblance de θ.

Nous avons

$$p_n(x_1, \dots, x_n) = \prod_{i=1}^{n} \theta_i e^{-\theta x_i}. \tag{39.80}$$

En passant au logarithme la fonction à minimiser est

$$L(\theta) = n \ln(\theta) - \theta \sum_i x_i. \tag{39.81}$$

La minimisation donne

$$\theta = \frac{n}{\sum_i x_i}, \tag{39.82}$$

c'est à dire

$$\hat{\theta}_n(X_1, \dots, X_n) = \frac{n}{\sum_i X_i} = \frac{1}{\bar{X}_n}. \tag{39.83}$$

Ce résultat n'est pas étonnant vu que le paramètre λ de la loi exponentielle est l'inverse de la moyenne.

△

Exemple 39.32

Soit $(X_1, \dots, X_n)$, un échantillon de loi parente uniforme sur $[0, \theta]$ avec $\theta > 0$.

(1) Montrer que la fonction de vraisemblance est donnée par

$$L(x_1, \dots, x_n; \theta) = \frac{1}{\theta^n} \mathbb{1}_{[0,\infty[}\big(\min(x_1, \dots, x_n)\big) \mathbb{1}_{[\max(x_1,\dots,x_n),\infty[}(\theta). \tag{39.84}$$

(2) Déterminer l'estimateur du maximum de vraisemblance de θ.

Lorsque nous parlons de paramètres qui peuvent prendre un spectre continu de valeurs, il est inutile de calculer la *probabilité* parce qu'elle est nulle. Le système de maximum de vraisemblance fonctionne avec les densités. Dans notre cas, la fonction de vraisemblance est le produit des densités :

$$L(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} p(x_i;\theta) \tag{39.85a}$$
$$= \prod_i \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_i) \tag{39.85b}$$
$$= \frac{1}{\theta^n}\mathbb{1}_{[0,\theta]^n}(x_1,\ldots,x_n). \tag{39.85c}$$

De cette expression nous voyons que $\min\{x_1,\ldots,x_n\}$ doit être positif en même temps que le maximum doit être plus petit que θ. Cette seconde condition peut s'écrire $\mathbb{1}_{[\max\{x_1,\ldots,x_n\},\infty[}(\theta)$. Au final nous avons

$$L(x_1,\ldots,x_n;\theta) = \frac{1}{\theta^n}\mathbb{1}_{[0,\infty[}\big(\min(x_1,\ldots,x_n)\big)\mathbb{1}_{[\max\{x_1,\ldots,x_n\},\infty[}(\theta). \tag{39.86}$$

Il n'est évidemment pas possible de dériver explicitement cette expression. Par contre pour que cette fonction soit non nulle, il faut obligatoirement $\theta \geqslant \max\{x_1,\ldots,x_n\}$. Par conséquent elle prend son maximum pour $\theta = \max\{x_1,\ldots,x_n\}$.

La conclusion est que l'estimateur de maximum de vraisemblance de θ est $\hat{\theta}_n = \max_i\{X_i\}$. △

Exemple 39.33

Déterminer l'estimateur de maximum de vraisemblance de la moyenne pour un modèle statistique dans laquelle la famille de probabilités est

$$(\mu_\theta) = \{\mathscr{N}(\theta,\sigma^2) \text{ tel que } \theta \in\in \mathbb{R}\}. \tag{39.87}$$

Nous supposons que σ est connu.

La densité de la variable aléatoire conjointe $(X_1,\ldots,X_n)$ au point $(x_1,\ldots,x_n)$ est le produit des densités, donc

$$\prod_{i=1}^{n}\gamma_{m,\sigma}(x_1,\ldots,x_n) = \frac{1}{\sigma^n(2\pi)^{n/2}}\exp -\frac{1}{2}\left(\sum_i\left(\frac{x_i-m}{\sigma}\right)^2\right). \tag{39.88}$$

Étant donné que le but est de minimiser, nous pouvons oublier le facteur et passer au logarithme :

$$L_0(\bar{x}) = -\frac{1}{2}\sum_{i=0}^{n}\left(\frac{x_i-m}{\sigma}\right)^2. \tag{39.89}$$

Nous pouvons également supprimer le $\frac{1}{2}$ et le $1/\sigma^2$. La fonction à minimiser devient

$$L(x_1,\ldots,x_n) = -\sum_i (x_i-m)^2, \tag{39.90}$$

dont la dérivée vaut $2nm - 2\sum_i x_i$. Par conséquent nous avons un minimum avec

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{39.91}$$

△

39.5.6 Estimation d'une fonction de répartition

Théorème 39.34 (Glivenko-Cantelli[427]).
Soient X_i des variables aléatoires indépendantes et identiquement distribuées suivant une loi dont la fonction de distribution est F (inconnue). Nous définissons l'estimateur

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leqslant x}. \tag{39.92}$$

Alors

$$P\big(\lim_{n\to\infty} \|F_n - F\|_\infty = 0\big) = 1. \tag{39.93}$$

Autrement dit, pour presque tout $\omega \in \Omega$,

$$\lim_{n\to\infty} \|F_n(\omega, .) - F\|_\infty = 0. \tag{39.94}$$

C'est à dire qu'il y a presque certainement convergence en probabilité.

Notons que de façon générale lorsqu'on parle d'estimateurs, partout où il y a un « n » dans une variable aléatoire, il y a une dépendance sous-entendue en ω.

Proposition 39.35.
Pour presque tout x, l'estimateur $F_n(x)$ est sans biais par rapport à $F(x)$:

$$E\big(F_n(x)\big) = F(x). \tag{39.95}$$

Démonstration. C'est juste un calcul :

$$E\big(F_n(x)\big) = \frac{1}{n}\sum_{i=1}^{n} E\big(\mathbb{1}_{\{X_i \leqslant x\}}\big) = \frac{1}{n}\sum_{i=1}^{n} F(x) = F(x). \tag{39.96}$$

□

39.5.7 Exemples sous forme d'exercices

Exemple 39.36
Dans cet exercice nous construisons un estimateur biaisé qui présente un risque quadratique inférieur à un estimateur non biaisé.

Soit un modèle statistique dont la famille de lois est $\{\mathcal{N}(m, \sigma^2)\}_{\theta\in]0,\infty[}$ où m est un paramètre réel connu. En ce qui concerne la variance nous considérons

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2. \tag{39.97}$$

(1) Montrer que $\hat{\sigma}_n^2$ est un estimateur sans biais de θ.
(2) Montrer que le risque quadratique de $\hat{\sigma}_n^2$ est

$$R(\hat{\sigma}_n^2, \theta) = \frac{2\theta^2}{n}. \tag{39.98}$$

(3) Considérer la famille d'estimateurs $c\hat{\sigma}_n^2$ avec $c > 0$. Déterminer la valeur de c qui minimise le risque quadratique de l'estimateur.

(1) Nous calculons

$$\begin{aligned} E(\hat{\sigma}_n^2) &= \frac{1}{n}\sum_i E\big[(X_i - m)^2\big] \\ &= \frac{1}{n}\sum_i \mathrm{Var}(X_i - m) - \frac{1}{n}\sum_i E(X_i - mj^2) \\ &= \frac{1}{n}\sum_i \mathrm{Var}(X_i) \\ &= \sigma^2. \end{aligned} \tag{39.99}$$

(2) Par la formule (39.60) nous savons que le risque d'un estimateur sans biais est donné par sa variance :

$$R(\hat{\theta}_n, \theta) = \text{Var}(|\hat{\sigma}_n^2 - \theta|) = \text{Var}(\hat{\sigma}_n^2). \tag{39.100}$$

Étant donné que les variables aléatoires X_i sont indépendantes et identiquement distribuées, le lemme 38.23 nous enseigne que la variance de la somme est la somme des variances. Nous avons donc à calculer

$$\begin{aligned}\text{Var}(|\hat{\sigma}_n^2 - \theta|) &= \text{Var}(\hat{\sigma}_n^2) \\ &= \frac{1}{n^2}\sum_i \text{Var}\left[(X_i - m)^2\right] \\ &= \frac{1}{n^2}\sum_i E\big[(X_i - m)^4\big] - E\big[(X_i - m)^2\big]^2.\end{aligned} \tag{39.101}$$

Ces espérances ne sont pas très compliquées à calculer en utilisant la fonction caractéristique donnée par la proposition 38.105 :

$$\Phi_{X-m}(t) = E\big(e^{it(X-m)}\big) = e^{-itm}E(e^{itX}) = e^{-itm}\Phi_X(t) = \exp\left(-\frac{\sigma^2 t^2}{2}\right). \tag{39.102}$$

Nous avons $\Phi^{(4)}(0) = 3\sigma^4$ et $\Phi''(0) = -\sigma^2$. Par conséquent

$$\text{Var}(\hat{\sigma}_n^2) = \frac{1}{n^2}\sum_i 2\theta^2 = \frac{2\theta^2}{n}. \tag{39.103}$$

Notez ici que $\theta = \sigma^2$.

(3) En tenant compte du fait que $\text{Var}(\hat{\sigma}_n^2) = \frac{2\theta^2}{n}$ et $E(\hat{\sigma}_n^2) = \theta$, nous avons

$$E\big((c\hat{\sigma}_n^2 - \theta)^2\big) = \text{Var}(c\hat{\sigma}_n^2 - \theta) + E(c\hat{\sigma}_n^2 - \theta)^2 \tag{39.104a}$$

$$= 2\frac{c^2\theta^2}{n} + (c-1)^2\theta^2. \tag{39.104b}$$

La dérivée (par rapport à c) de cela s'annule pour $c_0 = \frac{n}{n+2}$. Notons que nous n'avons pas tout à fait démontré que cela est bien un minimum. Calculons cependant le risque quadratique de notre estimateur pour cette valeur de c. Pour cela nous reportons $c = c_0$ dans l'expression (39.104b) :

$$E\big(\frac{n}{n+2}\hat{\sigma}_n^2\big) = \frac{2\theta^2}{n+2}. \tag{39.105}$$

Cela est effectivement plus petit que $R(\hat{\sigma}_n^2, \theta)$.

Nous avons ainsi construit un estimateur biaisé qui a un risque quadratique plus petit que l'estimateur non biaisé. △

Exemple 39.37

Nous considérons la famille de probabilités $\mu_\theta = \mathcal{N}(\theta)$ où $\theta = (m, \sigma^2) \in \mathbb{R} \times]0, \infty[$. Déterminer l'estimateur de maximum de vraisemblance du paramètre $\theta = (m, \sigma^2)$.

Pour chaque observation x_i nous avons une densité gaussienne. Le produit donne

$$p\big(x_1, \dots, x_n; (m, \sigma^2)\big) = \frac{1}{\sigma^2(2\pi)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i - m)^2\right]. \tag{39.106}$$

En passant au logarithme et en supprimant des facteurs inutiles à la minimisation,

$$L(m, \sigma) = -n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_i (x_i - m)^2. \tag{39.107}$$

L'annulation de la dérivée par rapport à m donne immédiatement $m = \frac{1}{n}\sum_i x_i$. L'annulation de la dérivée par rapport à σ donne

$$-n\sigma^2 + \sum_i (x_i - m)^2 = 0 \tag{39.108}$$

et donc

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \bar{x}_n). \tag{39.109}$$

L'estimateur de maximum de vraisemblance du couple $\theta = (m, \sigma^2)$ est donc

$$\hat{\theta}_n = \left(\frac{1}{n}\sum_i X_i, \frac{1}{n}\sum_i (X_i - \bar{X}_n)^2 \right). \tag{39.110}$$

△

39.5.8 Espérance et variance d'un estimateur

Soit $T_n = T_(X_1, \dots, X_n)$ un estimateur du paramètre θ dans un modèle d'échantillonnage. Les moyennes et variances de l'estimateur sont les variables aléatoires

$$m_{\theta,n} = E_\theta(T_n) = \int_{\mathbb{R}^n} T_n(x_1, \dots, x_n) d\mu_\theta^{\otimes n}(x1, \dots, x_n), \tag{39.111a}$$

$$\mathrm{Var}_\theta(T_n) = \int_{\mathbb{R}^n} [T_n(x_1, \dots, x_n) - m_{\theta,n}]^2 d\mu_\theta^{\otimes n}(x1, \dots, x_n), \tag{39.111b}$$

Lemme 39.38.
Si l'estimateur T_n satisfait

$$\lim_{n\to\infty} E_\theta(T_n) = \theta \tag{39.112a}$$

$$\lim_{n\to\infty} \mathrm{Var}_\theta(T_n) = 0, \tag{39.112b}$$

alors il est convergeant.

Démonstration. Nous utilisons l'inégalité de Markov (théorème 39.18) et nous introduisons l'espérance de l'estimateur :

$$P\big(|T_n(X_1, \dots, X_n) - \theta| > \epsilon\big) \leqslant \frac{1}{\epsilon} E\big(T_n(X_1, \dots, X_n) - \theta\big) \tag{39.113a}$$

$$\leqslant \frac{1}{\epsilon} E\big(|T_n - m_{n,\theta}|\big) + \frac{1}{\epsilon} E\big(|m_{n,\theta} - \theta|\big) \tag{39.113b}$$

$$\tag{39.113c}$$

Le second terme est l'espérance d'une constante. Nous majorons le premier terme en utilisant le fait que $\|.\|_1 \leqslant \|.\|_2$ (voir la remarque 28.24 après l'inégalité de Hölder) :

$$P\big(|T_n(X_1, \dots, X_n) - \theta| > \epsilon\big) \leqslant \frac{1}{\epsilon} E\big(|T_n - m_{n,\theta}|^2\big)^{1/2} + \frac{1}{\epsilon}|E_\theta(T_n) - \theta| \tag{39.114a}$$

$$= \frac{1}{\epsilon}\mathrm{Var}(T_n)^{1/2} + \frac{1}{\epsilon}|E_\theta(T_n) - \theta|. \tag{39.114b}$$

Les deux termes tendent séparément vers zéro par hypothèse. Nous avons par conséquent la convergence en probabilité $T_n \to \theta$. □

39.6 Estimation par intervalle de confiance

Nous voudrions estimer la proportion d'individus dans une population ayant un certain caractère déterminé par une variable booléenne : chaque individu a ou non le caractère étudié. L'échantillon sera donc une suite de 0 et de 1.

Pour tout $i \in \{1, \dots, n\}$ nous notons

$$x_i = \begin{cases} 1 & \text{si le } i\text{ème individu a la caractère} \\ 0 & \text{sinon.} \end{cases} \tag{39.115}$$

et $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Notre modèle statistique sera

$$\mathcal{S} = \Big[(\Omega, \mathcal{F}, P), (X_\theta), B(1, \theta)\Big] \tag{39.116}$$

où Ω est l'ensemble des individus étudiés, P est la manière de choisir les individus lors du sondage (essentiellement c'est une loi uniforme) et X_θ est la variable aléatoire

$$X(\omega) = \begin{cases} 1 & \text{si } \omega \text{ a le caractère} \\ 0 & \text{sinon} \end{cases} \tag{39.117}$$

Cela est une variable aléatoire de distribution $\mathscr{B}(1, p)$ où p est inconnu. Ici, $\Theta = [0, 1]$ est l'ensemble des p possibles.

Remarque 39.39.
Nous supposons que Ω est la population entière et que la variable aléatoire est l'opinion de la personne ω. En cela, nous considérons que le tirage de l'échantillon est sans remise. Le fait que nous modélisions par une variable aléatoire de Bernoulli signifie que nous considérons l'approximation dans laquelle la population globale est grande.

Nous supposons que nous ayons un échantillon $(X_1, \dots, X_n)$ dont nous avons observé une réalisation $(x_1, \dots, x_n)$ de fréquence $\bar{x}_n$. Nous voudrions déterminer un intervalle dans lequel $\bar{X}_n$ a de fortes chances de se trouver. Plus précisément nous considérons un petit α et nous cherchons ϵ tel que

$$P\big(p \in [\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]\big) = 1 - \alpha. \tag{39.118}$$

Typiquement, $\alpha = 5\%$. Le nombre α est le **niveau de confiance** que nous nous fixons a priori.

Si nous trouvons un intervalle I tel que $P(p \in I) = 1 - \alpha$, nous disons que l'intervalle est **exact**, si nous avons $P(p \in I) \geqslant 1 - \alpha$, nous disons que l'intervalle est **par excès**.

Il y a deux points de départs pour trouver l'intervalle. Le plus simple est d'utiliser le théorème central limite et considérer

$$\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{\mathscr{L}} \mathscr{N}(0, 1). \tag{39.119}$$

La seconde est d'utiliser la loi exacte : $n\bar{X}_n = \sum_i X_i \sim \mathscr{B}(n, p)$.

Bien entendu la seconde donne lieu à des calculs plus compliquée.

Remarque 39.40.
Dans certaines vraies vies (par exemple en médecine), la taille des échantillons est très réduite. Dans ce cas le théorème central limite n'a aucun sens et les calculs exact s'imposent.

De plus dans de nombreux cas de la vraie vie, nous avons un ordinateur à disposition pour calculer avec la loi exacte. L'utilisation du théorème central limite dans le but de produire un intervalle de confiance semble donc de plus en plus être une survivance du passé.

Dans la suite, nous allons supposer que n est suffisamment grand pour justifier l'approximation normale du théorème central limite 38.89. Si Z est une variable aléatoire normale centrée réduite,

notre premier essai est de faire

$$1-\alpha = P\big(p\in[\bar{X}_n-\epsilon,\bar{X}_n+\epsilon]\big) \tag{39.120a}$$
$$= P\left(\frac{-\epsilon\sqrt{n}}{\sqrt{p(1-p)}}\leqslant\sqrt{n}\frac{\bar{X}_n-p}{\sqrt{p(1-p)}}\leqslant\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \tag{39.120b}$$
$$\simeq P\left(-\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}\leqslant Z\leqslant\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \tag{39.120c}$$
$$= 2P\left(Z\leqslant\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)-1. \tag{39.120d}$$

La dernière ligne utilise la symétrie de la distribution $\mathcal{N}(0,1)$. Le nombre ϵ que nous cherchons vérifie donc

$$P\left(Z\leqslant\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)=1-\frac{\alpha}{2}. \tag{39.121}$$

De nos jours, les ordinateurs donnent la loi de répartition inverse des normales. Cela nous fournit un nombre t_α tel que

$$\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}=t_\alpha \tag{39.122}$$

où t_α est le nombre tel que $P(Z\leqslant t_\alpha)=1-\alpha/2$.

Conclusions de ce premier essai :

(1) Le problème est que nous ne pouvons pas déduire ϵ de l'équation (39.122) parce que p est inconnu.

(2) Cela ruine notre premier essai et nous demande de trouver mieux.

(3) L'astuce est évidemment de remplacer p par $\bar{X}_n$, mais il faut le justifier.

Méthode Slutsky Le point de départ du premier essai infructueux était la convergence

$$\sqrt{n}\frac{\bar{X}_n-p}{\sqrt{p(1-p)}}\xrightarrow{\mathscr{L}}N\sim\mathcal{N}(0,1) \tag{39.123}$$

donnée par le théorème central limite 38.89. Ce que nous voudrions en réalité est la convergence

$$\sqrt{n}\frac{\bar{X}_n-p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\xrightarrow{\mathscr{L}}N \tag{39.124}$$

La loi des grands nombres nous donne

$$\bar{X}_n(1-\bar{X}_n)\xrightarrow{p.s.}p(1-p). \tag{39.125}$$

Par conséquent le lemme de Slutsky implique la convergence en loi du couple :

$$\left(\sqrt{n}\frac{\bar{X}_n-p}{\sqrt{p(1-p)}},\sqrt{\bar{X}_n(1-\bar{X}_n)}\right)\xrightarrow{\mathscr{L}}\big(N,\sqrt{p(1-p)}\big). \tag{39.126}$$

À ce point des opérations nous pouvons utiliser la proposition 38.75 au couple avec la fonction

$$f(x,y)=\frac{x}{y}\sqrt{p(1-p)} \tag{39.127}$$

dont la probabilité d'être continue est 1 ($y=0$ est de mesure nulle dans $\mathbb{R}^2$). La conclusion du théorème est que

$$\sqrt{n}\frac{\bar{X}_n-p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}}\xrightarrow{\mathscr{L}}\mathcal{N}(0,1). \tag{39.128}$$

C'est à partie de là que nous pouvons construire notre intervalle de confiance :

$$1-\alpha = P(\bar{X}_n - \epsilon \leqslant p \leqslant \bar{X}_n + \epsilon) \tag{39.129a}$$

$$= P\left(\frac{\sqrt{n}\epsilon}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \geqslant \sqrt{n}\frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \geqslant \frac{-\sqrt{n}\epsilon}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \right)) \tag{39.129b}$$

$$\simeq P\left(\frac{\sqrt{n}\epsilon}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \geqslant N \geqslant \frac{-\sqrt{n}\epsilon}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \right)). \tag{39.129c}$$

Nous cherchons maintenant dans les tables le ξ qui fait

$$P(-\xi \leqslant N \leqslant \xi) = 1-\alpha \tag{39.130}$$

puis nous cherchons ϵ de telle sorte à avoir

$$\frac{\sqrt{n}\epsilon}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} = \xi. \tag{39.131}$$

Dans cette équation tout est connu à part le ϵ qui se découvre.

Méthode piétonne Nous remarquons que l'événement

$$-\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}} \leqslant \sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \leqslant \frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}} \tag{39.132}$$

est le même que l'événement

$$\left| \sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \right| \leqslant t_\alpha. \tag{39.133}$$

Vu que t_α est positif, cela est encore le même événement que

$$n\frac{(\bar{X}_n - p)^2}{p(1-p)} \leqslant t_\alpha^2 \tag{39.134}$$

ou encore

$$p^2(n+t_\alpha^2) - p(2n\bar{X}_n + t_\alpha^2) + n\bar{X}_n^2 \leqslant 0. \tag{39.135}$$

Les racines du polynôme du membre de gauche sont

$$p_\pm = \frac{2n\bar{X}_n + t_\alpha^2 \pm \sqrt{(2n\bar{X}_n + t_\alpha)^2 - 4(n+t_\alpha^2)n\bar{X}_n^2}}{2(n+t_\alpha^2)}. \tag{39.136}$$

Le but étant d'effectuer une limite $n \to \infty$, nous factorisons d'abord n. Après simplification

$$p_\pm = \frac{\bar{X}_n + \frac{t_\alpha}{2n} \pm t_\alpha\sqrt{\frac{t_\alpha^2}{4n^2} + \frac{\bar{X}_n(1-\bar{X}_n)}{n}}}{1+\frac{t_\alpha}{n}}. \tag{39.137}$$

Étant donné que nous considérons que n est grand, nous allons négliger les termes en $\frac{1}{n}$ en faisant attention à ce que le terme en $\frac{1}{n}$ sous la racine est en réalité $1/\sqrt{n}$ et ne doit pas être négligé. Nous trouvons, à cette approximation, que

$$p \in \Big[\bar{X}_n - t_\alpha\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + t_\alpha\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \Big] \tag{39.138}$$

avec une probabilité $1-\alpha$.

39.6.1 Région de confiance

Soit un n-échantillon $(X_1, \dots, X_n)$ de loi parente μ_θ. Nous supposons $\Theta \subset \mathbb{R}$ avec Θ ouvert. Soit $\alpha \in [0,1]$ un intervalle de confiance et une application mesurable

$$\begin{aligned} \Lambda\colon \mathbb{R}^n &\to \mathcal{B}\text{or}(\Theta) \\ (x_1, \dots, x_n) &\mapsto \Lambda(x_1, \dots, x_n). \end{aligned} \tag{39.139}$$

On appelle **région de confiance exact** au niveau de confiance $1-\alpha$ une région aléatoire $\Lambda(x_1, \dots, x_n)$ telle que

$$P\big(\theta \in \Lambda(x_1, \dots, x_n)\big) = 1 - \alpha. \tag{39.140}$$

Si $d = 1$, la région $\Lambda(x_1, \dots, x_n)$ est un intervalle.

39.6.2 Fonction pivotale

Soit $\hat{\theta}_n$, un estimateur de θ. Une fonction v sur $\Theta \times \Theta$ est **pivotale** pour θ si la loi de la variable aléatoire $v(\hat{\theta}_n, \theta)$ ne dépend pas de θ. Elle est **asymptotiquement pivotale** si

$$v(\hat{\theta}_n, \theta) \xrightarrow{\mathscr{L}} \xi \tag{39.141}$$

où ξ est une variable aléatoire indépendante de θ.

En pratique, nous essayons de faire apparaître une variable aléatoire de loi connue qui ne dépend pas du paramètre que l'on recherche. Si la variance est connue et si l'échantillon est grand, le théorème central limite nous permet d'introduire une loi normale centrée réduite.

Exemple 39.41
Soit $X_1, \dots, X_n$ des variables aléatoires de loi parente $\mathscr{N}(m, \sigma^2)$. Une fonction asymptotiquement pivotale pour m est

$$v(z_1, z_2) = \frac{z_1 - z_2}{\sigma/\sqrt{n}} \tag{39.142}$$

parce que la variable aléatoire

$$v(\bar{X}_n, m) = \frac{\bar{X}_n - m}{\alpha/\sqrt{n}} \tag{39.143}$$

tend vers $\mathscr{N}(0,1)$ qui ne dépend pas de m. △

Exemple 39.42
Si (X_n) est une suite de variables aléatoires indépendantes et identiquement distribuées de moyenne m et d'écart type σ que nous supposons inconnus. Le fonction suivante est asymptotiquement pivotale pour m :

$$v(\bar{X}_n, m) = \frac{\bar{X}_n - m}{\sigma/\sqrt{n}}. \tag{39.144}$$

△

Soit $(X_1, \dots, X_n)$ un échantillon de loi $\mathscr{N}(m, \sigma_0^2)$ avec σ_0^2 connu. Nous cherchons un intervalle de confiance $1-\alpha$ pour m. Pour cela nous allons utiliser une fonction asymptotiquement pivotale, à savoir

$$\frac{\bar{X}_n - m}{\sigma_0^2/\sqrt{n}} \sim \mathscr{N}(0,1). \tag{39.145}$$

Nous devrions chercher des valeurs z_+ de z_- telles que

$$P\left(z_- \leqslant \frac{\bar{X}_n - m}{\sigma_0/\sqrt{n}} \leqslant z_+\right) = 1 - \alpha. \tag{39.146}$$

Pour des raisons de symétries (de la courbe gaussienne), nous allons chercher un intervalle symétrique : $z_- = -z_+$. Le nombre à chercher est donc le z_α tel que

$$P\big(|Z| \leqslant z_\alpha\big) = 1 - \alpha. \tag{39.147}$$

Si nous demandons $\alpha = 5\%$, la réponse est $z_\alpha = 1.96$, c'est à dire que

$$P\left(-1.96 \leqslant \frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \leqslant 1.96\right) = 0.95. \tag{39.148}$$

Nous avons donc

$$P\left(m \in \big[\bar{X}_n - \frac{1.96\sigma}{\sqrt{n}}, \bar{X}_n + \frac{1.96\sigma}{\sqrt{n}}\big]\right) = 0.95. \tag{39.149}$$

Supposons maintenant que nous avons observé 100 valeurs numériques avec $\bar{x}_n = 12$ et $\sigma = 1$. La réalisation de l'intervalle de confiance pour m au niveau de confiance 0.95 est :

$$\big[12 - 0.196, 12 + 0.196\big]. \tag{39.150}$$

Cet intervalle est à interpréter de la façon suivante : si nous recommençons un grand nombre de fois le sondage, la moyenne tombera 95% des fois dans l'intervalle ainsi calculé. Mais il faut bien comprendre que la probabilité

$$P\left(m \in \big[12 - 0.196, 12 + 0.196\big]\right) \tag{39.151}$$

vaut zéro ou un.

Exemple 39.43

Soit un échantillon $(X_1, \dots, X_n)$ de loi parente $\mathscr{N}(m, \sigma^2)$ où m et σ^2 sont inconnus. Déterminer un intervalle de confiance exact symétrique au niveau de confiance $1 - \alpha$.

Déterminer un intervalle de confiance pour σ^2.

Nous savons que la moyenne empirique est un estimateur de la moyenne :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{39.152}$$

Nous cherchons un intervalle du type $I = [\bar{X}_n - \epsilon, \bar{X}_n + \epsilon]$ pour lequel $P(m \in I) = 1 - \alpha$. Nous savons que la variable aléatoire

$$\frac{\bar{X}_n - m}{\alpha/\sqrt{n}} \tag{39.153}$$

suit une loi $\mathscr{N}(0, 1)$, mais la variance est inconnue. La subtilité à savoir est que la variable aléatoire

$$Z = \frac{\bar{X}_n - m}{S_n/\sqrt{\sigma}} \tag{39.154}$$

où $S_n^2 = \sum_i (X_i - \bar{X}_n)^2/(n-1)$ suit une loi de Student à n degrés de liberté $\mathscr{T}(n-1)$ en vertu du théorème de Cochran 39.15. Comme il est usuel de le faire, nous inversons l'intervalle :

$$1 - \alpha = P\left(-\epsilon \leqslant \bar{X}_n - m \leqslant \epsilon\right) \tag{39.155a}$$

$$= P\left(-\frac{\epsilon\sqrt{n}}{S_n} \leqslant Z \leqslant \frac{\epsilon\sqrt{n}}{S_n}\right). \tag{39.155b}$$

Les valeurs se trouvent dans des tables ; par exemple pour $n = 10$ et $\alpha = 5\%$ nous trouvons

$$\frac{\epsilon\sqrt{n}}{S_n} = 2.262. \tag{39.156}$$

Plus généralement nous notons $t_{n-1,1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathscr{T}(n-1)$, c'est à dire le nombre tel que

$$P(Z \leqslant t_{n-1,1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} \tag{39.157}$$

si $Z \sim \mathscr{T}(n-1)$. L'intervalle de confiance est alors donné par

$$I = \left[\bar{X}_n - \frac{t_{n-1,1-\frac{\alpha}{2}} S_n}{\sqrt{n}}, \bar{X}_n + \frac{t_{n-1,1-\frac{\alpha}{2}} S_n}{\sqrt{n}} \right]. \tag{39.158}$$

Cela est un intervalle exact pour m au niveau de confiance $1-\alpha$.

Nous pouvons aussi trouver un intervalle asymptotique en utilisant le théorème central limite :

$$\frac{\bar{X}_n - m}{S_n/\sqrt{n}} \xrightarrow{\mathscr{L}} T \tag{39.159}$$

avec $T \sim \mathscr{N}(0,1)$.

Rappel : dire que I_n est un **intervalle de confiance asymptotique** signifie que

$$\lim_{n\to\infty} P(m \in I_n) = 1 - \alpha. \tag{39.160}$$

En ce qui concerne la variance σ^2, l'intervalle de confiance se construit en utilisant la partie (2) du théorème de Cochran 39.15. Nous introduisons la variable aléatoire pivot

$$Z = (n-1)\frac{S_n^2}{\sigma^2} \tag{39.161}$$

qui suit une loi $\chi^2(n-1)$. Cette loi n'étant pas symétrique (voir figure 38.1), nous n'allons pas chercher un intervalle de confiance symétrique. Nous cherchons c_1 et c_2 tels que

$$\begin{cases} P\big(\sigma^2 \in [c_1, c_2]\big) = 1 - \alpha & \text{(39.162a)} \\ P\big(\sigma^2 \in [0, c_1]\big) = \dfrac{\alpha}{2} & \text{(39.162b)} \\ P\big(\sigma^2 \in [c_2, \infty[\big) = \dfrac{\alpha}{2} & \text{(39.162c)} \end{cases}$$

La situation est représentée à la figure 39.3.

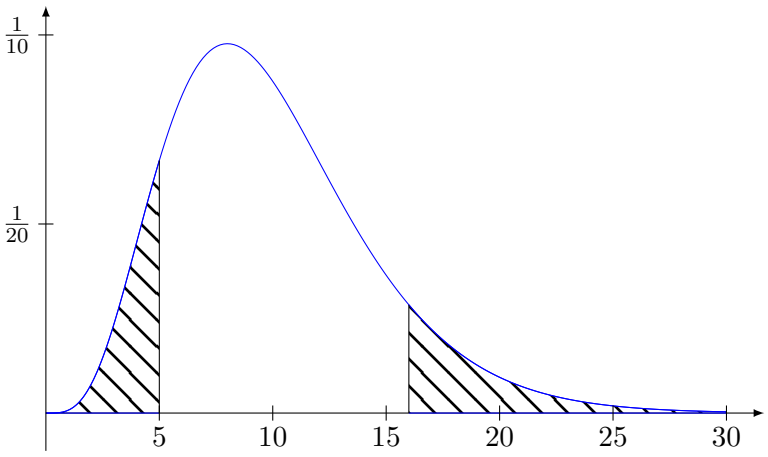


FIGURE 39.3 – L'intervalle de confiance pour une χ^2.

Le construction des nombres c_1 et c_2 passe par la relation

$$P(c_1 \leqslant \sigma^2 \leqslant c_2) = P\left(\frac{(n-1)S_n^2}{c_2} \leqslant Z \leqslant \frac{(n-1)S_n^2}{c_1} \right). \tag{39.163}$$

Nous notons $t_{n-1,\frac{\alpha}{2}}$ et $t_{n-1,1-\frac{\alpha}{2}}$ les quantiles donnés sur la figure 39.3, c'est à dire

$$P(Z \leqslant t_{n-1,\frac{\alpha}{2}}) = \frac{\alpha}{2} P(Z \geqslant t_{n-1,1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}. \tag{39.164a}$$

Ce que nous obtenons est

$$\frac{(n-1)S_n^2}{c_2} = t_{n-1,\frac{\alpha}{2}} \tag{39.165a}$$

$$\frac{(n-1)S_n^2}{c_1} = t_{n-1,1-\frac{\alpha}{2}}, \tag{39.165b}$$

et par conséquent l'intervalle de confiance pour σ^2 est

$$I = \left[\frac{(n-1)S_n^2}{t_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S_n^2}{t_{n-1,1-\frac{\alpha}{2}}} \right] \tag{39.166}$$

avec $P(\sigma^2 \in I) = 95\%$.

Remarque 39.44.
Il n'est pas clair a priori que la longueur de l'intervalle I décroisse avec n parce qu'il y a n dans les t au numérateur.

△

39.6.3 Sondage de proportion

Une utilisation classique des statistiques est d'interpréter une proportion donnée par un sondage. Nous considérons une élection avec deux candidats A et B. Nous avons interrogés $n = 2500$ personnes et nous avons obtenus 51% pour le candidat A et 49% pour le candidat B. Que peut on dire ?

La modélisation de cette situation est que nous avons des variables aléatoires $X_i \sim \mathscr{B}(p_A)$ et que nous en avons observés n avec une moyenne

$$\bar{x}_n = 0.51. \tag{39.167}$$

La loi de $\bar{X}_n$ est une binomiale. Sa densité n'est pas symétrique, mais si n est grand, elle le devient. Nous cherchons un intervalle

$$I = [\bar{X}_n - \epsilon, \bar{X}_n + \epsilon] \tag{39.168}$$

tel que $P(p_A \in I) = 1 - \alpha$. Pour cela nous considérons le fait que $n = 2500$ est grand et nous utilisons la limite de la proposition 39.10 :

$$Z_n = \sqrt{n} \frac{\bar{X}_n - p_A}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \xrightarrow{\mathscr{L}} Z \sim \mathscr{N}(0,1). \tag{39.169}$$

La variable aléatoire Z_n est asymptotiquement pivotale et normale centrée réduite. Nous cherchons donc un intervalle symétrique pour $\bar{X}_n - p_A$:

$$1 - \alpha = P(-\epsilon \leqslant \bar{X}_n - p_A \leqslant \epsilon), \tag{39.170}$$

c'est à dire, si n est grand,

$$1 - \alpha = P\left(-\epsilon \frac{\sqrt{n}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \leqslant Z_n \leqslant \epsilon \frac{\sqrt{n}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \right) \tag{39.171}$$

où Z_n est une normale centrée réduite. Nous trouvons ainsi, via les tables que

$$\frac{\epsilon\sqrt{n}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} = 1.96 \tag{39.172}$$

si nous voulons un intervalle à 5%. Par conséquent nous avons $\epsilon = 1.96\frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{nj}}$ et l'intervalle de confiance est

$$I_C = \left[\bar{X}_n - \frac{1.96\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}, \bar{X}_n + \frac{2.96\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n}}\right]. \tag{39.173}$$

La propriété de cet intervalle est que

$$\lim_{n\to\infty} P(p_A \in I_c) = 1-\alpha. \tag{39.174}$$

Remarque 39.45.
À quel moment avons nous fait une hypothèse sur la taille de la population globale ? En modélisant les sondés par des variables de Bernoulli et leur somme par une binomiale, nous supposons que le sondage est *avec remise*, sinon, elles ne seraient pas indépendantes. En supposant les sondés indépendants, nous avons donc fait comme si la population totale était infinie.

39.7 Estimer une densité lorsqu'on ne sait rien

Nous supposons avoir une série d'observations issues d'un processus complexe dont nous n'avons aucune idée de la loi parente, et nous voudrions nous faire une idée de la densité de cette loi inconnue.

Nous observons une suite de réalisations que nous modélisons comme étant des variables aléatoires $(X_1,\dots,X_n)$ de loi parente (inconnue) μ. Notre but est de trouver un estimateur $\hat{\mu}$ de μ. Par simplicité nous allons nous restreindre aux lois admettant une densité par rapport à la mesure de Lebesgue. C'est à dire que nous allons estimer μ par une suite de lois $\hat{\mu}_n$ qui sont toutes des lois acceptant une densité par rapport à la mesure de Lebesgue.

39.7.1 Distance entre des mesures

Si ν_1 et ν_2 sont deux mesures de densité sur $\mathbb{R}$, la **distance** entre ν_1 et ν_2 est définie par

$$d(\nu_1,\nu_2) = \sup_{A\in\mathcal{B}\text{or}(\mathbb{R})} \left|\nu_1(A)-\nu_2(A)\right| \tag{39.175}$$

où $\mathcal{B}\text{or}(\mathbb{R})$ désigne l'ensemble des boréliens de $\mathbb{R}$.

Théorème 39.46 (de Scheffé[428])**.**
Si f_1 et f_2 sont les densités de ν_1 et ν_2 par rapport à la mesure de Lebesgue, alors

$$d(\nu_1,\nu_2) = \int_{\mathbb{R}} (f_1-f_2)_+ = \frac{1}{2}\int_{\mathbb{R}} |f_1-f_2| = \frac{1}{2}\|f_1-f_2\|_1 \tag{39.176}$$

où f_+ est la partie positive de f (pour la décomposition $f(x) = f_+(x) - f_-(x)$).

Démonstration. La dernière égalité est simplement une notation usuelle ; nous devons seulement prouver les deux premières. Pour la première nous commençons par prouver que le borélien réalisant le supremum est

$$B = \{x\in\mathbb{R} \text{ tel que } f_1(x) \geqslant f_2(x)\}. \tag{39.177}$$

En effet si A est un borélien nous avons

$$\nu_1(A)-\nu_2(A) = \int_A f_1-f_2 \leqslant \int_{A\cap B} f_1-f_2 \leqslant \int_B f_1-f_2 = \int_B (f_2-f_2)_+ = \int_{\mathbb{R}} (f_1-f_2)_+ \tag{39.178}$$

Justifications :

— $f_1 - f_2$ négative sur $A \cap \complement B$.
— Vu que $f_1 - f_2 \geqslant 0$ sur B, l'intégrale augmente si on augmente le domaine.
— Sur B nous avons $f_1 - f_2 = (f_1 - f_2)_+$.

Donc pour tout borélien A nous avons

$$d(\nu_1, \nu_2) \leqslant \int_{\mathbb{R}} (f_1 - f_2)_+. \tag{39.179}$$

Mais pour $A = B$ nous avons égalité :

$$\nu_1(B) - \nu_2(B) = \int_B f_1 - f_2 = \int_B (f_1 - f_2)_+ = \int_{\mathbb{R}} (f_1 - f_2). \tag{39.180}$$

Pour la seconde égalité nous savons que f_1 et f_2 s'intègrent toutes deux à 1 (parce que ce sont des densités de probabilité), donc

$$\int_{\mathbb{R}} f_1 - f_2 = 0. \tag{39.181}$$

En particulier nous avons

$$\int_{\mathbb{R}} (f_1 - f_2)_+ = \int_{\mathbb{R}} (f_1 - f_2)_-, \tag{39.182}$$

ce qui donne bien

$$\int_{\mathbb{R}} (f_1 - f_2)_+ = \frac{1}{2} \int_{\mathbb{R}} |f_1 - f_2|. \tag{39.183}$$

□

39.7.2 Estimateur par fenêtres glissantes

Nous considérons les estimations suivantes de la fonction de répartition :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leqslant x\}}, \tag{39.184}$$

et un nombre h_n qui sera la taille de la fenêtre glissante. Nous avons en tête de faire $\lim_{n\to\infty} h_n = 0$. Nous considérons ceci comme estimateur de la densité inconnue f des variables aléatoires X_i :

$$\hat{f}_n(x) = \frac{F_n(x + h_n) - F_n(x - h_n)}{2h_n}. \tag{39.185}$$

L'idée sous-jacente est de prendre la dérivée de la fonction de répartition comme densité.

Lemme 39.47 ([428]).
Pour tout $h_n > 0$, l'estimateur $\hat{f}_n$ est une densité de probabilité.

Démonstration. D'abord $\hat{f}_n$ est bien à valeurs positives ou nulle. Ensuite devons parler de son intégrale. Pour le numérateur nous avons

$$F_n(x + h_n) - F_n(x - h_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \in B(x, h_n)}. \tag{39.186}$$

En réalité cette égalité est valable seulement presque partout parce qu'elle n'est pas valable en $x = x \pm h_n$, mais cela ne va pas nous ennuyer dans la mesure où nous avons dans l'idée d'intégrer cela sur $\mathbb{R}$. Avant de nous lancer dans l'intégrale nous remarquons que $X_i \in B(x, h_n)$ est la même chose que $x \in B(X_i, h_n)$, c'est à dire que

$$\{X_i \in B(x, h_n)\} = \{x \in B(X_i, h_n)\}. \tag{39.187}$$

Donc

$$\int_{\mathbb{R}} \hat{f}_n = \frac{1}{2h_n} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\int_{\mathbb{R}} \mathbb{1}_{B(X_i, h_n)}}_{2h_n} = \frac{1}{2nh_n} \sum_{i=1}^{n} 2h_n = 1. \tag{39.188}$$

□

Lemme 39.48 ([428]).
L'estimateur $\hat{f}_n$ est déjà pas mal parce que

$$\lim_{h_n\to 0} E\big(\hat{f}_n(x)\big) \to f(x) \tag{39.189}$$

pour presque tout $x \in \mathbb{R}$.

Démonstration. Nous commençons par nous rappeler le fait que $F_n(x)$ est un estimateur sans biais de $F(x)$ (proposition 39.35). Donc

$$E\big(\hat{f}_n(x)\big) = \frac{F(x+h_n) - F(x-h_n)}{2h_n}. \tag{39.190}$$

Nous devons prendre la limite de cela lorsque $h_n \to 0$, c'est à dire considérer la dérivée de F. Attention : rien ne dit que F soit dérivable, si ce n'est la proposition 19.33 qui indique qu'elle est dérivable presque partout avec f comme dérivée.

La limite $h_n \to 0$ dans (39.190) nous donne donc bien presque partout

$$\lim_{h_n\to 0} E\big(\hat{f}_n(x)\big) = f(x). \tag{39.191}$$

□

Proposition 39.49 ([428]).
Si la suite (h_n) est telle que $h_n \to 0$ et $nh_n \to \infty$, alors pour presque tout $x \in \mathbb{R}$ nous avons les convergences

$$\hat{f}_n(x) \xrightarrow{L^2(P)} f(x) \tag{39.192}$$

et

$$\hat{f}_n(x) \xrightarrow{P} f(x). \tag{39.193}$$

Démonstration. Il faut d'abord comprendre ce que signifie la convergence $L^2(P)$ pour presque tout x. Pour cela il faut comprendre que $\hat{f}_n(x)$ est en soi une variable aléatoire et est en réalité une fonction $\omega \mapsto \hat{f}_n(x,\omega)$. Ce que nous allons montrer est que pour presque tout x (maintenant fixé), cette variable aléatoire converge vers une constante (par rapport à ω) et que cette constante est $f(x)$.

La convergence $X_n \xrightarrow{L^2(P)} X$ signifie $E\big(|X_n - X|^2\big) \to 0$, c'est à dire

$$\int_\Omega \big|X_n(\omega) - X(\omega)\big|^2 dP(\omega) \to 0. \tag{39.194}$$

En faisant une décomposition biais-variance nous devons donc étudier

$$E\Big[\big(\hat{f}_n(x) - f(x)\big)^2\Big] = E\big[\hat{f}_n(x) - f(x)\big]^2 + \operatorname{Var}\big(\hat{f}_n(x) - f(x)\big) \tag{39.195}$$

Ici $f(x)$ doit être vue comme la variable aléatoire constante sur Ω. Par le lemme 39.48 et la proposition 38.25 le terme de biais converge vers zéro lorsque $n \to \infty$.

Pour traiter le terme de biais, nous savons déjà que

$$2nh_n\hat{f}_n(x) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in B(x,h_n)\}}, \tag{39.196}$$

où le membre de droite (et donc aussi celui de gauche) est est une variable aléatoire binomiale comptant le nombre de succès de l'expérience $X_i \in B(x,h_n)$ en n essais. Nous notons $p_{x,n} = P\big(X_i \in B(x,h_n)\big)$. Si μ est la loi parente des X_i, alors

$$p_{n,x} = P\big(X_i \in B(x,h_n)\big) = \mu\big(B(x,h_n)\big) = F(x+h_n) - F(x-h_n) \tag{39.197}$$

où F est la fonction de répartition (parente) des X_i.

Alors la variance de ladite binomiale est donnée par (38.327), c'est à dire $np_{x,n}(1-p_{x,n})$. Nous avons alors

$$\text{Var}\left(2nh_n\hat{f}_n(x)\right) = bp_{x,n}(1-p_{x,n}) \tag{39.198}$$

et

$$\text{Var}\left(\hat{f}_n(x)\right) = \frac{1}{4n^2h_n^2}np_{x,n}(1-p_{x,n}). \tag{39.199}$$

Nous pouvons faire la majoration $t(1-t) \leqslant t$ qui est valable pour tout t et écrire

$$\text{Var}(\hat{f}_n(x)) \leqslant \frac{1}{4nh_n}\frac{p_{x,n}}{h_n}. \tag{39.200}$$

Le premier facteur tend vers zéro parce que nous avons supposé que $nh_n \to \infty$. Pour le second facteur, il faut remarquer que l'expression (39.197) nous donne presque partout

$$\lim_{h_n\to 0}\frac{p_{n,x}}{h_n} = 2f(x), \tag{39.201}$$

qui est constant et certainement borné.

Nous avons maintenant prouvé que pour presque tout x nous avons $\hat{f}_n(x) \overset{L^2(P)}{\to} f(x)$. Montrons que cela implique la convergence en loi, c'est à dire que pour tout $\eta > 0$, nous avons la limite

$$P\big(|\hat{f}_n(x) - f(x)| > \eta\big) \to 0. \tag{39.202}$$

Si cela n'était pas vrai, nous aurions un nombre $\eta_0 > 0$ et $\epsilon > 0$ tel que pour tout n à partir d'une certaine taille,

$$P\Big(|\hat{f}_n(x) - f(x)|^2 > \eta_0\Big) > \epsilon, \tag{39.203}$$

et en particulier en notant A l'événement $|\hat{f}_n(x) - f(x)|^2 > \eta_0^2$, $P(A) > \epsilon$. Alors

$$\int_\Omega \big|\hat{f}_n(x,\omega) - f(x,\omega)\big|^2 dP(\omega) \geqslant \int_A \big|\hat{f}_n(x,\omega) - f(x,\omega)\big|^2 dP(\omega) \geqslant \int_A \eta_0^2 = \eta_0^2 P(A). \tag{39.204}$$

Cela signifie que

$$\|\hat{f}_n(x) - f(x)\|_{L^2(P)} \geqslant \eta_0 P(A), \tag{39.205}$$

ce qui contredit la première convergence démontrée. □

Note : l'hypothèse $nh_n \to \infty$ revient à dire que nous voulons que chaque boîte contienne de plus en plus d'observations. Si nous avions $nh_n \to 0$, alors avec n qui augmente, la majorité des boîtes deviendraient vides, ce qui reviendrait à une perte d'information.

39.8 Test d'hypothèses, prise de décision

39.8.1 Exemple : qualité des pièces d'usine

Une usine fabrique des composantes électronique garantis un an. Le constructeur ne veut pas accepter que plus de 5% des pièces tombent en panne avant un an.

Nous supposons que la durée de vie T d'une pièce soit une variable aléatoire suivant une loi exponentielle de paramètre λ (qui est l'inverse de la moyenne : $\theta = 1/\lambda$). Le fabriquant veut donc s'assurer que

$$0.95 \leqslant P(T \geqslant 1), \tag{39.206}$$

ou encore

$$P(T \geqslant 1) = \int_0^\infty \frac{1}{\theta}e^{-x/\theta}dx = e^{-1/\theta} \geqslant 0.95, \tag{39.207}$$

donc le fabriquant doit s'assurer que

$$\theta \geqslant \frac{1}{\ln\left(\frac{1}{0.95}\right)}. \tag{39.208}$$

Nous posons donc $\theta_0 = 19.5$ et nous prenons comme modèle de décision que si $\theta < \theta_0$, alors la chaîne de production doit être revue, et si $\theta > \theta_0$, alors l'usine peut continuer son travail.

Ce dont nous disposons n'est pas de θ, mais d'une estimation de θ à partir d'un échantillon. Cela étant il faudra aussi pouvoir estimer la probabilité de faire un mauvais choix.

39.8.2 Exemple : la résistance d'un fil

Un artisan a besoin d'un fil qui a une résistance à une traction de 100 g en moyenne. Si la résistance est trop faible, le fil casse ; si elle est trop grande, c'est trop rigide et ça ne convient pas.

Remarque 39.50.
Dans l'exemple précédent, avoir $\theta > \theta_0$ ne dérange pas. Si la durée de vie moyenne est de 2 ans, le directeur de l'usine ne sera pas malheureux. Ici par contre l'artisan cherche une valeur précise et a donc une borne vers le haut et vers le bas.

L'artisan reçoit un lot de fils et souhaite savoir s'il est conforme. Pour cela, il prend 4 fils au hasard et mesure une moyenne de 112 g. Est-ce que cela est cohérent avec une moyenne de 100 g ?

Nous faisons l'hypothèse que ma résistance des fils suit une loi normale $\mathscr{N}(m, \sigma^2)$ avec m inconnu. Pour la simplicité nous supposons que σ est connu et vaut 10.

Nous devons prendre une décision entre deux hypothèses. L'hypothèse H_0 sera de dire que le lot a une résistance de 100 g et l'hypothèse alternative sera que le lot a une résistance différente.

Les 4 observations sont quatre variables aléatoires $(X_i)_{i=1,2,3,4}$, et le nombre 112 est une réalisation de la variable aléatoire

$$\bar{X}_4 = \frac{1}{4}(X_1 + X_2 + X_3 + X_4). \tag{39.209}$$

Nous supposons que H_0 est vraie, et nous calculons quelle est l'intervalle autour de $m = 100$ qui a 95% de chances de contenir la moyenne observée. Si 112 est dedans, nous acceptons H_0 et si 112 est hors de cet intervalle, nous refusons H_0.

Compte tenue de l'hypothèse H_0, nous avons

$$\frac{\bar{X}_4 - 100}{\frac{10}{\sqrt{4}}} = \frac{\bar{X}_4 - 100}{5} \sim \mathscr{N}(0, 1). \tag{39.210}$$

Nous commençons à connaître par cœur l'intervalle de confiance à 95% d'une loi normale centrée réduite ; le quantile est à 1.96, c'est à dire

$$P\left(-1.96 \leqslant \frac{\bar{X}_4 - 100}{5} \leqslant 1.96\right) = 0.95, \tag{39.211}$$

ou encore

$$P\big(\bar{X}_4 \in [90.2, 109.8]\big). \tag{39.212}$$

Il y a donc moins de 5% de chances que la moyenne de ces quatre fils tombent en dehors de l'intervalle $[90.2, 109.8]$. L'artisan doit donc rejeter l'hypothèse et considérer que le lot est mauvais.

La région

$$]-\infty, 90.2] \cup [109.8, \infty[\tag{39.213}$$

est la **région de rejet**, ou **région critique**.

Ici, le nombre 5% représente le risque de refuser H_0 alors qu'elle était vraie. Notons que nous ne pouvons pas calculer le risque d'accepter H_0 alors qu'elle est fausse. En effet, si H_0 est fausse, nous ne savons pas quelles sont les valeurs de $\bar{X}_4$ acceptables parce qu'il y a une infinité de possibilités pour m qui soient alternatifs à $m = 100$.

Évidemment si la vraie moyenne est $100+10^{-7}$, l'hypothèse H_0 sera acceptée, mais nous n'avons aucun moyen de savoir si elle est vraie ou non.

39.8.3 Vocabulaire et théorie

Nous avons un modèle d'échantillonnage paramétrique $(X_{\theta,1},\dots,X_{\theta,n})$ de taille n et de paramètre inconnu θ, de loi parente μ_θ appartenant à une famille paramétrique de lois $(\mu_\theta)_{\theta\in\Theta}$.

Soient H_0 et H_1 deux ensembles disjoints tels que $\Theta = H_0 \cup H_1$. L'ensemble H_0 sera nommé **hypothèse nulle** et l'ensemble H_1 sera l'**hypothèse alternative**.

Pour l'exemple des fils, nous avions $H_0 = \{100\}$ et $H_1 = \mathbb{R}\backslash\{100\}$. Si une hypothèse est réduite à un singleton, nous parlons d'hypothèse **simple** et sinon c'est une hypothèse **composite** ou **multiple**. Faire un tests consiste à déterminer une région critique.

Définition 39.51.
*Un **test** est une application mesurable δ qui à $(x_1,\dots,x_n)\in\mathbb{R}^n$ associe*

$$\delta(x_1,\dots,x_n)\in\{0,1\}. \tag{39.214}$$

Si $\delta(x_1,\dots,x_n)=0$ on accepte l'hypothèse H_0 pour l'échantillon $x_1,\dots,x_n$, et si $\delta(x_1,\dots,x_n)=1$, alors on rejette H_0 et on choisit H_1. L'ensemble

$$W=\{(x_1,\dots,x_n)\in\mathbb{R}^n \text{ tel que } \delta(x_1,\dots,x_n)=1\} \tag{39.215}$$

*est la **région de rejet** ou la **région critique**.*

L'ensemble $W=\delta^{-1}(1)$ est un borélien de $\mathbb{R}^n$ parce que δ est mesurable. L'événement auquel nous sommes intéressés est l'événement

$$R=\{(X_{\theta,1},\dots,X_{\theta,n})\in W\}. \tag{39.216}$$

Exemple 39.52
Pour l'exemple de 39.8.2 nous avions

$$\delta(x_1,\dots,x_4)=\mathbb{1}_{\complement[90.2,109.8]}(\bar{x}_4). \tag{39.217}$$

△

39.8.4 Risque de première et seconde espèce

Le modèle de décision que nous avons introduit comprend deux façons de se tromper. Soit nous rejetons H_0 alors qu'elle est vraie (c'est le **risque de première espèce**), soit nous acceptions H_0 alors qu'elle est fausse (risque de **seconde espèce**). Nous pouvons formaliser ces concepts de la façon suivante.

Nous considérons un test de région critique W. Le risque de première espèce, noté α est la fonction

$$\begin{aligned}\alpha\colon H_0 &\to [0,1]\\ \theta &\mapsto P\big((X_{\theta,1},\dots,X_{\theta,n})\in W\big).\end{aligned} \tag{39.218}$$

Il s'agit de la probabilité de rejeter H_0 alors qu'elle es vraie. Le risque de seconde espèce est la fonction

$$\begin{aligned}\beta\colon H_1 &\to [0,1]\\ \theta &\mapsto P\big((X_{\theta,1},\dots,X_{\theta,n})\notin W\big).\end{aligned} \tag{39.219}$$

C'est la probabilité d'accepter H_0 alors qu'elle est fausse.

Définition 39.53.
*Soit δ un test de région critique W. La **puissance** du test est la fonction*

$$\begin{aligned}\eta\colon H_1 &\to [0,1]\\ \theta &\mapsto P\big((X_{\theta,1},\dots,X_{\theta,n})\in W\big).\end{aligned} \tag{39.220}$$

*La **courbe d'efficacité** du test est la fonction*

$$\begin{aligned} h\colon \Theta &\to [0,1] \\ \theta &\mapsto P\big((X_{\theta,1},\dots,X_{\theta,n}) \notin W\big). \end{aligned} \tag{39.221}$$

La puissance d'un test est la probabilité de rejeter H_0 lorsque H_1 est vraie. Plus la puissance est grande, mieux c'est. La courbe d'efficacité du test est la probabilité d'accepter H_0 pour une certaine valeur de θ.

Soit un test δ. Une statistique $T_\theta = T_n(X_{\theta,1},\dots,X_{\theta,n})$ est une **variable de décision** pour δ si $\complement W$ peut s'écrire d'une des façons suivantes

$$\complement W = \begin{cases} \{(x_1,\dots,x_n) \in \mathbb{R}^n \text{ tel que } T_n(x_1,\dots,x_n) < c\} & \text{test unilatéral à droite} \\ \{(x_1,\dots,x_n) \in \mathbb{R}^n \text{ tel que } T_n(x_1,\dots,x_n) > c\} & \text{test unilatéral à gauche} \\ \{(x_1,\dots,x_n) \in \mathbb{R}^n \text{ tel que } c_1 \leqslant T_n(x_1,\dots,x_n) < c_2\} & \text{test bilatéral.} \end{cases} \tag{39.222}$$

Le plus souvent la variable de décision sera la moyenne : $T_n(x_1,\dots,x_n) = \frac{1}{n}\sum_{i=1}^n x_i$. Les valeurs c, c_1, c_2 sont des **valeurs critiques**.

En ce qui concerne les notations, ici T_n représente la valeur mesurée sur un n-échantillon (d'où l'indice n) alors que T_θ est la valeur *théorique* de T lorsque θ est la vraie valeur du paramètre qu'on veut estimer.

Pour un test unilatéral à gauche, nous fixons la valeur critique c de telle manière a avoir

$$P(T_\theta > c) \leqslant \alpha. \tag{39.223}$$

Pour un test unilatéral à gauche, nous fixons c de telle manière a avoir

$$P(T_\theta < c) \leqslant \alpha \tag{39.224}$$

et pour un test bilatéral nous fixons c_1 et c_2 de telle façon à avoir

$$P(T_\theta > c_2) = P(T_\theta < c_1) \leqslant \frac{\alpha}{2}. \tag{39.225}$$

39.8.5 Modèle paramétrique de loi gaussienne

Soit un modèle statistique paramétrique de lois gaussiennes $\mathscr{N}(m,1)$ de moyenne m inconnue avec $m \in \mathbb{R}^+$. Nous avons $\Theta = [0,\infty[$.

Nous observons un échantillon de taille $n = 36$. Avec un risque de première espèce de 5% nous voulons estimer l'hypothèse $H_0 = \{0\}$ contre l'hypothèse $H_1 =]0,\infty[$. De notre échantillon nous construisons la variable aléatoire

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \tag{39.226}$$

dans laquelle les X_i sont les éléments de l'échantillon, elles sont indépendantes et identiquement distribuées de loi $\mathscr{N}(m,1)$ avec m inconnu.

Si $\bar{X}_n$ est proche de zéro nous acceptons H_0, sinon nous la rejetons. La région de rejet s'écrit donc sous la forme

$$W = \{(x_1,\dots,x_n) \in \mathbb{R}^n \text{ tel que } \bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i > u\} \tag{39.227}$$

dans lequel il faut fixer le u pour satisfaire au risque de première espèce de 5%. La contrainte est d'avoir

$$P(\bar{X}_n > u) = \alpha \tag{39.228}$$

si H_0 est vérifiée. Cela revient à dire que dans 5% des cas où H_0 est correcte, nous la rejetterons. Si H_0 est vraie alors $\bar{X}_n$ est une moyenne de gaussiennes de moyennes m et nous avons

$$\frac{\bar{X}_n - m}{\sigma/\sqrt{n}} \sim \mathscr{N}(0,1) \tag{39.229}$$

avec $m = 0$ et $\sigma = 1$. L'équation (39.228) devient donc

$$\alpha = P\left(\frac{\bar{X}_n}{1/\sqrt{n}} > \frac{u}{1/\sqrt{n}}\right) = P(T > \sqrt{n}u) \tag{39.230}$$

où $T \sim \mathcal{N}(0,1)$. Avec $n = 36$ et $\alpha = 5\%$ nous trouvons

$$u = \frac{1.645}{6} \simeq 0.274 \tag{39.231}$$

La règle de décision est donc la suivante : si $\bar{x}_n > 0.274$ alors nous rejetons H_0, et sinon nous l'acceptons.

Calculons la puissance de ce test (définition 39.53). C'est la fonction donnée par

$$\begin{aligned}\eta\colon H_1 &\to \mathbb{R}\\ m &\mapsto P\big((X_{1,m},\dots,X_{n,m}) \in W\big) = P\left(\frac{1}{n}\sum X_i > u\right).\end{aligned} \tag{39.232}$$

Dans ce calcul, les X_i sont d'une loi normale $\mathcal{N}(m,1)$, et non $\mathcal{N}(0,1)$. En retranchant m et en divisant par $1/\sqrt{n}$ nous trouvons

$$\eta(m) = P\left(\frac{\frac{1}{n}\sum X_i - m}{1/\sqrt{t}} > \frac{u-m}{1/\sqrt{n}}\right) \tag{39.233a}$$

$$= P(T > \sqrt{n}(u-m)) \tag{39.233b}$$

$$= P(T > 16.45 - 6m) \tag{39.233c}$$

$$= 1 - \Phi(1.645 - 6m) \tag{39.233d}$$

où Φ est la fonction de répartition de $\mathcal{N}(0,1)$. La fonction η a les propriétés suivantes :

$$\lim_{m\to-\infty} \eta(m) = 0 \tag{39.234a}$$

$$\lim_{m\to\infty} \eta(m) = 1 \tag{39.234b}$$

$$\eta(0) = \frac{5}{100}. \tag{39.234c}$$

Remarque 39.54.
Si nous regardons $m = 0.001$, le risque de seconde espèce est quasiment de 90%. En effet le risque de seconde espèce est d'accepter H_0 alors qu'il est faux. Lorsque $m = 0.001$, l'hypothèse H_0 est fausse, mais la probabilité qu'on l'accepte est grande. D'ailleurs les conséquences de l'accepter à tort ne sont peut-être pas si grandes que cela.

39.9 Tests paramétriques

La proposition suivante montre le lien entre région de confiance et les tests.

Proposition 39.55.
Soit $\Lambda(X_1,\dots,X_n)$ une région de confiance par excès de niveau de confiance $1-\alpha$. Alors il existe un tests pur de niveau α pour tester $H_0 = \{\theta_0\}$ de région de rejet

$$W_n = \{x = (x_1,\dots,x_n) \in \mathbb{R}^n \text{ tel que } \theta_0 \notin \Lambda(x_1,\dots,x_n)\}. \tag{39.235}$$

Démonstration. L'hypothèse sur Λ signifie qu'avec les observations $(X_1,\dots,X_n)$, il y a une forte probabilité (plus grande que $1-\alpha$) que θ soit dans $\Lambda(X_1,\dots,X_n)$. Avec ou sans H_0 nous avons donc

$$P(\theta \in \Lambda) \geqslant 1-\alpha. \tag{39.236}$$

Supposons maintenant l'hypothèse H_0, alors

$$P\big((X_1,\dots,X_n) \in W_n\big) = P\big(\theta_0 \notin \Lambda(X_1,\dots,X_n)\big) \leqslant \alpha. \tag{39.237}$$

□

Remarque 39.56.
Soit W_n la région de confiance d'un test de niveau α pour tester $H_0 = \{\theta_0\}$. Alors

$$\Lambda = \{x \in \mathbb{R}^n \text{ tel que } x \notin W_n\} \tag{39.238}$$

est une région de confiance $1 - \alpha$ pour θ.

Exemple 39.57

Soit $(X_1, \dots, X_n)$ un échantillon de loi parente $\mathcal{N}(\theta, 1)$ avec $\theta \in \{\theta_0, \theta_1\}$. Nous supposons $\theta_0 < \theta_1$. Nous voulons tester $H_0 = \{\theta_0\}$ contre $H_1 = \{\theta_1\}$. Nous proposons le test suivant. La variable de décision sera $\bar{X}_n$ et la région de rejet sera

$$W = \{(x_1, \dots, x_n) \in \mathbb{R}^n \text{ tel que } \frac{1}{n}\sum_i x_i > \frac{\theta_1 + \theta_2}{2}\}. \tag{39.239}$$

(1) Donner le risque de première espèce de ce test.
(2) Soit $0 < \alpha < 1$. Pour quelle valeur de n le tests a-t-il un risque de première espèce égal à α ?
(3) Donner la puissance du test.

Les réponses peuvent être exprimées en termes de la fonction de répartition F de la loi normale centrée réduite.

Le risque de première espèce est donné par

$$\alpha = P\left(\frac{1}{n}\sum_i X_i > \frac{\theta_0 + \theta_1}{2}\right) \tag{39.240}$$

où les X_i sont des variables aléatoires indépendantes et identiquement distribuées de loi parente $\mathcal{N}(\theta_0, 1)$. Cela est la probabilité d'être dans la région de rejet alors que l'hypothèse H_0 est vraie. La formule (39.240) se transforme en

$$\alpha = P\left(\frac{\frac{1}{n}\sum X_i - \theta_0}{1/\sqrt{n}} > \frac{\frac{\theta_0+\theta_1}{2} - \theta_0}{1/\sqrt{n}}\right) \tag{39.241a}$$

$$= P(T > \sqrt{n}\frac{\theta_1 - \theta_0}{2}). \tag{39.241b}$$

En termes de la fonction de répartition nous avons alors

$$\alpha = 1 - F\big(\sqrt{n}\frac{\theta_1 - \theta_0}{2}\big) \tag{39.242}$$

Il s'agit maintenant de trouver le nombre n qui réalise cette égalité. Pour cela nous utilisons l'inverse F^{-1} de la fonction de répartition de la normale :

$$n = \left(\frac{2}{\theta_1 - \theta_0}F^{-1}(1 - \alpha)\right)^2. \tag{39.243}$$

Le risque de seconde espèce est la possibilité d'accepter H_0 lorsque H_1 est vraie, c'est à dire

$$\beta = P\left(\frac{1}{n}\sum_i X_i < \frac{\theta_0 + \theta_1}{2}\right) \tag{39.244}$$

sous l'hypothèse H_1. Dans le calcul de (39.244) nous prenons donc $X_i \sim \mathcal{N}(\theta_1, 1)$. Le résultat est que

$$\beta = F\big(\sqrt{n}\frac{\theta_0 - \theta_1}{2}\big). \tag{39.245}$$

Remarque 39.58.
L'expression (39.243) diminue lorsque θ_0 et θ_1 s'éloignent, ce qui est normal : plus les nombres à discerner sont éloignés, moins l'échantillon à prendre pour réaliser le travail doit être grand.

Notons aussi que $\theta_0 - \theta_1 < 0$, par conséquent augmenter n diminue la valeur de

$$\beta = F\big(\sqrt{n}\frac{\theta_0 - \theta_1}{2}\big) \tag{39.246}$$

△ <++>

39.10 Tests d'adéquation

Soit $X_1, \dots, X_n$ un échantillon de loi parente X finie prenant ses valeurs dans $\{a_1, \dots, a_k\}$. La loi de X est donnée par les nombres

$$p_i = P(X = a_i) \tag{39.247}$$

pour $i = 1, \dots, k$. Nous introduisons l'**effectif empirique**, la variable aléatoire N_i qui compte le nombre de fois que a_i est observé dans l'échantillon. La **fréquence empirique** est la variable aléatoire

$$F_i = \frac{N_i}{n}. \tag{39.248}$$

Nous savons que la loi de N_i est $\mathscr{B}(n, p_i)$, et la loi des grands nombres dit que

$$F_i \xrightarrow{p.s.} p_i \tag{39.249}$$

pour chaque i. Le théorème central limite nous indique de plus que

$$\frac{N_i - np_i}{\sqrt{np_i(1-p_i)}} \xrightarrow{\mathscr{L}} T \sim \mathscr{N}(0,1). \tag{39.250}$$

Nous considérons un cas où les p_i sont inconnus. Ils peuvent être approchés par $N_i \simeq np_i$. Le théorème de Pearson nous indique comment.

Théorème 39.59 (Théorème de Pearson)**.**
Nous avons

$$\sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \xrightarrow{\mathscr{L}} K \sim \chi^2(k-1) \tag{39.251}$$

où la distribution $\chi^2(l)$ est la somme des carrés de l gaussiennes centrées réduites indépendantes.

Démonstration. Nous commençons par le cas $k = 2$. Dans ce cas nous avons $N_2 = n - N_1$ et $p_1 + p_2 = 1$. La sommes que nous regardons est

$$\frac{(N_1 - np_1)^2}{np_1} + \frac{(N_2 - np_2)^2}{np_2} = \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_1 - np_1)^2}{n(1-p_1)} \tag{39.252a}$$

$$= \frac{(N_1 - np_1)^2}{np_1(1-p_1)}. \tag{39.252b}$$

Étant donné que N_1 est une variable aléatoire binomiale nous avons

$$\frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}} \xrightarrow{\mathscr{L}} T \sim \mathscr{N}(0,1). \tag{39.253}$$

Par conséquent la limite de (39.252b) est

$$\left(\frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}}\right)^2 \xrightarrow{\mathscr{L}} T^2 \simeq \chi^2(1). \tag{39.254}$$

Cela conclut le cas $k = 2$.

Passons à présent au cas général. Le k-uple $(N_1, \dots, N_k)$ est une variable aléatoire multinomiale de loi

$$\mathscr{M}(n; k; p_1, \dots, p_k). \tag{39.255}$$

Nous introduisons les variables aléatoires U_i données par $U_i \colon \Omega \to \mathbb{R}^k$ avec

$$P\big(U_i = (0, \dots, 1, \dots, 0)\big) = p_i; \tag{39.256}$$

c'est le vecteur aléatoire qui prend ses valeurs dans les vecteurs de la base canonique de $\mathbb{R}^k$ et qui prend la valeur e_i avec probabilité p_i. Par construction nous avons

$$(N_1, \dots, N_k) = \sum_{i=1}^{n} U_i. \tag{39.257}$$

Nous allons étudier la fonction caractéristique de $(N_1, \dots, N_k)$ définie par l'équation (38.197) :

$$\begin{aligned} \Phi_{(N_1,\dots,N_k)} \colon \mathbb{R}^k &\to \mathbb{C} \\ e_j &\mapsto E(e^{ie_j \cdot N}) = E(e^{iN_j}). \end{aligned} \tag{39.258}$$

Plus généralement,

$$\Phi_{(N_1,\dots,N_k)}(t_1, \dots, t_k) = E\big(e^{i\langle t, N\rangle_{\mathbb{R}^k}}\big). \tag{39.259}$$

Nous avons

$$e^{i\langle t,N\rangle} = e^{i\sum_j \langle t, U_j\rangle} = \prod_j e^{i\langle t, U_j\rangle} \tag{39.260}$$

et vu que les U_i sont indépendantes et identiquement distribuées nous pouvons écrire U_1 à la place de U_j de façon à avoir

$$\Phi_{(N_1,\dots,N_k)}(t_1, \dots, t_k) = \prod_j E\big(e^{i\langle t, U_1\rangle}\big) \tag{39.261a}$$

$$= \prod_j \sum_l p_l e^{i\langle t, e_l\rangle} \tag{39.261b}$$

$$= \prod_j \sum_l p_l e^{it_l} \tag{39.261c}$$

$$= \left(\sum_{l=1}^{k} p_l e^{it_l} \right)^n. \tag{39.261d}$$

Nous allons montrer que

$$\lim_{n\to\infty} \Phi_{(N_1,\dots,N_n)}(t_1, \dots, t_k) = e^{-\frac{1}{2}\sum_{j=1}^n t_j^2} - \left(\sum_{j=1}^{k} t_j \sqrt{p_j} \right)^2. \tag{39.262}$$

Pour ce faire, nous allons effectuer un développement limité. D'abord nous introduisons les variables aléatoires

$$\alpha_j = \frac{N_j - np_j}{\sqrt{np_j}} \tag{39.263}$$

et nous calculons

$$\Phi_{(\alpha_1,\dots,\alpha_k)}(t_1, \dots, t_k) = E\left[\exp\left(i\langle t, \big(\frac{N_1 - np_1}{\sqrt{np_1}}, \dots, \frac{N_k - np_k}{\sqrt{np_k}}\big)\rangle \right) \right] \tag{39.264}$$

Étant donné que n et p_j sont des variables déterministes, nous pouvons les sortir de l'espérance. Nous avons alors

$$\Phi_{(\alpha_1,\dots,\alpha_k)}(t_1, \dots, t_k) = \exp\left(-i \sum_{j=1}^{k} t_j \sqrt{np_j} \right) \Phi_{(N_1,\dots,N_k)}\left(\frac{t_1}{\sqrt{np_1}}, \dots, \frac{t_k}{\sqrt{np_k}} \right) \tag{39.265}$$

parce que

$$E\left(e^{t_j\frac{N_j-np_j}{\sqrt{np_j}}}\right)=e^{\frac{-t_jnp_j}{\sqrt{np_j}}}E\left(e^{t_jN_j/\sqrt{np_j}}\right) \tag{39.266}$$

En remplaçant (39.261) dans (39.265) nous trouvons

$$\Phi_{(\alpha_1,\ldots,\alpha_k)}(t_1,\ldots,t_k)=\exp\left(-i\sum_{j=1}^{k}t_j\sqrt{np_j}\right)\underbrace{\left(\sum_{j=1}^{k}p_je^{i\frac{t_j}{\sqrt{np_j}}}\right)^n}_{A} \tag{39.267}$$

Nous analysons maintenant le terme A. Nous écrivons l'égalité $A=A+1-1$ en tenant compte de $\sum_j p_j=1$ sous la forme

$$A=\left(1+\sum_{j=1}^{k}p_j\big(\exp(it_j/\sqrt{np_j})-1\big)\right)^n, \tag{39.268}$$

Nous avons alors

$$\ln(A)=n\ln\left[1+\sum_{j=1}^{\infty}p_j\big(e^{it_j/\sqrt{np_j}}-1\big)\right] \tag{39.269}$$

Nous développons l'exponentielle en

$$e^{it_j/\sqrt{np_j}}-1=\frac{it_j}{\sqrt{np_j}}-\frac{t_j^2}{2np_j}+\frac{1}{n}\epsilon(1/n) \tag{39.270}$$

et ensuite le logarithme selon la formule

$$\ln(1+x)=x-\frac{x^2}{2}+x^2\alpha(x^2). \tag{39.271}$$

Nous avons

$$\ln(A)=n\ln\left[1+\sum_j p_j\Big(\frac{it_j}{\sqrt{np_j}}-\frac{t_j^2}{2np_j}+\frac{1}{n}\epsilon(\frac{1}{n})\Big)\right] \tag{39.272a}$$

$$=n\ln\left[1+\sum_j p_j\Big(it_j\sqrt{\frac{p_j}{n}}-\frac{t_j^2}{2n}+\frac{1}{n}\epsilon(\frac{1}{n})\Big)\right] \tag{39.272b}$$

$$=n\underbrace{\sum_{j=1}^{k}\left(it_j\sqrt{\frac{p_j}{n}}-\frac{t_j^2}{2n}+\frac{1}{n}\epsilon(1/n)\right)}_{K} \tag{39.272c}$$

$$-n\frac{1}{2}\left[\sum_j\left(it_j\sqrt{\frac{p_j}{n}}-\frac{t_j^2}{2n}+\frac{1}{n}\epsilon(1/n)\right)^2\right] \tag{39.272d}$$

$$+nK^2\alpha(K) \tag{39.272e}$$

Nous introduisons dans ϵ tous les termes en $1/n^2$ et nous trouvons

$$\ln(A)=\sum_j\left(it_j\sqrt{p_jn}-\frac{t_j^2}{2}\right)-\frac{1}{2}\left(\sum_j it_j\sqrt{p_j}\right)^2+\epsilon(1/n)+K^2\alpha(K). \tag{39.273}$$

En remplaçant dans (39.267) et en passant à la limite pour $n\to\infty$,

$$\Phi_{(\alpha_1,\ldots,\alpha_k)}(t_1,\ldots,t_k)=\exp\left(-\frac{1}{2}\sum_j t_j^2+\frac{1}{2}\big(\sum_j t_j\sqrt{p_j}\big)^2\right). \tag{39.274}$$

Nous reconnaissons des lois gaussiennes dans le premier terme de l'exponentielle. Nous allons maintenant nous atteler à identifier le second terme.

Soit C une matrice orthogonale dont la dernière ligne est $(\sqrt{p_1},\dots,\sqrt{p_k})$. Nous considérons les vecteurs

$$\alpha=\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_k\end{pmatrix},\quad t=\begin{pmatrix}t_1\\ \vdots\\ t_k\end{pmatrix}. \tag{39.275}$$

et ensuite nous notons

$$U=Ct=\begin{pmatrix}u_1\\ \vdots\\ u_k\end{pmatrix}. \tag{39.276}$$

Étant donné que C est orthogonale, nous avons $\sum_{i=1}^k\alpha_i^2=\sum_{i=1}^k\beta_i^2$ et

$$\Phi_{(\beta_1,\dots,\beta_k)}=E\big(e^{i\langle u,\beta\rangle}\big)=E\big(e^{i\langle t,\alpha\rangle}\big)=\Phi_{(\alpha_1,\dots,\alpha_k)}(t_1,\dots,t_k). \tag{39.277}$$

Nous pouvons récrire l'argument de l'exponentielle (39.274) de la façon suivante :

$$\sum_{i=1}^k t_j^2=\sum_j u_j^2 \tag{39.278a}$$

$$\sum_{j=1}^k t_j\sqrt{p_j}=(Ct)_k, \tag{39.278b}$$

Nous avons alors

$$\lim_{n\to\infty}\Phi_{(\beta_1,\dots,\beta_k)}(t_1,\dots,t_k)=\lim_{n\to\infty}\Phi_{(\alpha_1,\dots,\alpha_k)}(u_1,\dots,u_k) \tag{39.279a}$$

$$=\exp\left(\frac{1}{2}\sum_{j=1}^{k-1}u_j^2\right) \tag{39.279b}$$

$$=\Phi_{(Z_1,\dots,Z_{k-1},0)}(u_1,\dots,u_k) \tag{39.279c}$$

où les Z_i sont des variables aléatoires indépendantes et identiquement distribuées de distribution normale centrée réduite. Nous avons donc montré que

$$(\beta_1,\dots,\beta_k)\xrightarrow{\mathscr{L}}(Z_1,\dots,Z_{k-1},)). \tag{39.280}$$

Étant donné que l'application $x\mapsto\|x\|^2$ est continue, nous avons aussi

$$\|(\beta_1,\dots,\beta_k)\|^2\xrightarrow{\mathscr{L}}\|(Z_1,\dots,Z_{k-1},0)\|^2, \tag{39.281}$$

et par conséquent

$$\|(\alpha_1,\dots,\alpha_k)\|^2\xrightarrow{\mathscr{L}}\sum_{j=1}^{k-1}Z_j^2\sim\chi^2(k-1). \tag{39.282}$$

D'après la définition (39.263) nous avions

$$\|(\alpha_1,\dots,\alpha_k)\|^2=\sum_{j=1}^k\frac{(N_j-p_j)^2}{np_j}. \tag{39.283}$$

□

Chapitre 40

Chaînes de Markov à temps discret

Mets tes deux pieds en canard, c'est la chaîne de Markov qui se prépare.

40.1 Généralités

Les chaînes de Markov interviennent pour la description des systèmes dont l'évolution future ne dépend que de l'état présent.

Définition 40.1.
*Soit E un ensemble au plus dénombrable et $(\Omega, \mathcal{F}, P)$ un espace probabilisé. Une **chaîne de Markov** à valeurs dans E est une famille $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires telles que pour tout $x_0, \dots, x_{n+1} \in E$,*

$$P(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0) = P(X_{n+1} = x_{n+1} | X_n = x_n). \tag{40.1}$$

Pour une chaîne de Markov, il n'est pas important de savoir l'historique pour prédire la futur : X_{n+1} est seulement déterminé par X_n.

Remarque 40.2.
Il existe une théorie des chaînes de Markov à temps continu ou avec E non dénombrable, mais ce n'est pas au programme.

Nous notons

$$p_n(x, y) = P(X_{n+1} = y | X_n = x) \tag{40.2}$$

la **probabilité de transition** de la chaîne à l'instant n. Si cette probabilité ne dépend pas de n, nous disons que la chaîne de Markov est **homogène**, et nous notons $p(x, y)$ au lieu de $p_n(x, y)$. Nous notons Q la matrice (éventuellement infinie) de transition

$$Q_{xy} = p(x, y). \tag{40.3}$$

Nous avons

$$\sum_{y\in E} p(x, y) = 1 \tag{40.4}$$

parce que c'est la somme de toutes les transitions possibles en partant de x. Notons aussi que $p(x, y) \geqslant 0$.

Définition 40.3.
*Une matrice dont tous les éléments sont positifs ou nuls et donc la somme de toutes les lignes sont 1 est une **matrice stochastique**.*

Notons que l'ensemble des matrices stochastiques est un fermé dans l'ensemble des matrices.

Lemme 40.4.
Si U est une matrice stochastique[1]*, alors il existe une chaîne de Markov dont la matrice de transition est U.*

Remarque 40.5.
La somme $\sum_{x\in E} p(x,y)$ ne vaut pas spécialement 1. Si les états x_1 et x_2 arrivent tous les deux en y de façon certaine, alors nous avons $\sum_x p(x,y) \geqslant 2$. Il n'y a donc pas de limites aux sommes des colonnes.

Exemple 40.6
Nous considérons une fourmi qui se déplace dans un appartement à trois pièces A, B, C. Supposons qu'à chaque minute, elle a une probabilité 1/3 de rester dans la pièce et une probabilité 2/3 de se déplacer. Le plan de l'appartement est

$$A \longrightarrow B \longrightarrow C \tag{40.5}$$

De la pièce A est est donc uniquement possible d'aller vers la pièce B ; de la B il est possible d'aller en A et en C et de la C il est uniquement possible d'aller en B.

La matrice de transition de cette chaîne de Markov est

$$Q = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 2/3 & 1/3 \end{pmatrix} \tag{40.6}$$

△

Exemple 40.7
Si N_t est un processus de Poisson, alors les variables aléatoires $X_n = N_n$ forment une chaîne de Markov. △

40.2 Chaînes de Markov sur un ensemble fini

Une chaîne de Markov est **finie** si l'ensemble E dans lequel elle prend ses valeurs est fini.

Proposition 40.8 ([429])**.**
Si (X_n) est une chaîne de Markov irréductible sur un ensemble fini, alors pour tout ensemble $A \subset E$ nous avons

$$P(\tau_A < \infty) = \lim_{n\to\infty} P(\tau_A \leqslant n) = 1 \tag{40.7}$$

où $\tau_A = \min\{k \text{ tel que } X_k \in A\}$.

Les proposition à venir vont montrer que

(1) Toute matrice stochastique admet un état stationnaire, proposition 40.9.

(2) Si la chaîne de Markov est irréductible, alors il y a unicité de l'état stationnaire, proposition 40.10. Mais attention : cela ne veut pas encore dire que la chaîne converge effectivement vers cet état.

(3) Si la chaîne est irréductible et apériodique, alors il y a convergence en loi vers l'unique loi invariante, théorème 40.13.

Proposition 40.9 ([430])**.**
Toute matrice stochastique admet un état stationnaire.

1. Définition 40.3.

Proposition 40.10 ([430]).
Soit une chaîne de Markov irréductible finie. Alors il existe une unique loi stationnaire π et de plus nous avons $\pi_i > 0$ pour tout état i de E.

Définition 40.11.
*Une chaîne de Markov finie est **régulière** s'il existe un $n \in \mathbb{N}$ tel que P^n a uniquement des éléments strictement positifs.*

Théorème 40.12 ([429]).
Soit P la matrice de transition d'une chaîne de Markov régulière sur un ensemble E de cardinal N. Alors il existe des nombres $\pi_1, \dots, \pi_N$ tels que

(1) $\pi_i > 0$ *pour tout* $i = 1, \dots, N$

(2) $\pi_1 + \cdots + \pi_N = 1$

(3)

$$\lim_{n\to\infty} P^n = \Pi = \begin{pmatrix} \pi_1 & \pi_2 & \dots & \pi_N \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \dots & \pi_N \end{pmatrix} \tag{40.8}$$

De plus le vecteur $\pi = (\pi_1, \dots, \pi_N)$ est l'unique solution de

$$\pi P = \pi. \tag{40.9}$$

Démonstration. Si la chaîne n'a qu'un seul état c'est évident parce que la probabilité de transition est toujours 1 ; fin de l'histoire.

Hypothèse Sinon nous supposons que P n'a que des éléments positifs, quitte à considérer P^n au lieu de P. Nous notons d le plus petit élément de P ; il vérifie $d \leqslant \frac{1}{2}$ parce que la somme des élément d'une ligne de la matrice P doit être égale à 1.

Les suites min et max Soit x un vecteur quelconque (de composantes positives). Nous notons $m_0 = \min\{x_i\}$ et $M_0 = \max\{x_i\}$. Étant donné que les éléments du vecteur Px sont des moyennes pondérées des éléments de x, si nous posons

$$m_k = \min\{(P^k x)_i\}_{i=1,\dots,N} \tag{40.10a}$$

$$M_k = \max\{(P^k x)_i\}_{i=1,\dots,N}, \tag{40.10b}$$

la suite (m_k) est croissante et la suite (M_k) est décroissante.

Stricte croissance et décroissance Si $M_{k+1} = M_k$, alors toutes les composantes de $P^k x$ sont égales à M_k et le théorème est prouvé. Cela est encore une propriété de la moyenne. Même remarque pour la suite (m_k).

Nous pouvons donc supposer que la suite (m_k) est strictement croissante et que la suite (M_k) est strictement décroissante. Elles sont toutes les deux bornées dans $[m_0, M_0]$. Le lemme 9.17 nous donne la convergence.

Égalité des limites Vu que les éléments de $P^k x$ ne sont pas tous les mêmes et s'étalent de m_k à M_k, pour majorer M_{k+1} nous mettons le plus petit coefficient possible (c'est à dire d) devant m_k et nous supposons que toutes les autres composantes sont M_k ; nous avons alors

$$M_{k+1} \leqslant dm_k + (1-d)M_k \tag{40.11}$$

parce que tous les autres coefficients de la ligne contenant le d (dans P^k) sont plus petits ou égaux à $1-d$. De la même façon nous avons la minoration

$$m_{k+1} \geqslant dM_k + (1-d)m_k. \tag{40.12}$$

En faisant la différence, et en nous souvenant que $0 < 1 - 2d < 1$,

$$M_{k+1} - m_k \leqslant (1-2d)(M_k - m_k), \tag{40.13}$$

ce qui signifie que

$$M_{k+1} - m_k \leqslant (1-2d)^k (M_0 - m_0), \tag{40.14}$$

et donc que les deux limites sont égales.

Conclusion pour la limite Pour tout vecteur x, la suite $P^k x$ tend vers un vecteur dont toutes les composantes sont égales. En particulier pour le vecteur e_i de la base canonique,

$$P^k e_i \to \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_1 \end{pmatrix}. \tag{40.15}$$

Mais $P^k e_i$ est la i^{e} colonne de la matrice P^k. Cela prouve la convergence annoncée $P^k \to \Pi$.

Réglons rapidement le cas des deux autres allégations du théorème. D'abord les matrices P^k sont toutes des matrices stochastiques ; et l'ensemble des matrices stochastiques est fermé, donc la convergence se fait à l'intérieur de l'ensemble des matrices stochastiques. Cela prouve que $\pi_1 + \cdots + \pi_N = 1$.

Ensuite la suite (m_k) étant strictement croissante et m_0 étant égal à 0 dans le cas de e_i nous avons toujours $\pi_i > 0$ (strictement). □

Théorème 40.13 ([430])**.**
Si (X_n) est une chaîne de Markov finie, irréductible et apériodique de loi stationnaire π, alors

(1) La suite de matrices stochastiques P^k converge vers la matrice

$$P^k \to \Pi = \begin{pmatrix} \pi \\ \vdots \\ \pi \end{pmatrix}. \tag{40.16}$$

(2) Nous avons convergence des variables aléatoires au sens où

$$P(X_k = \mu P^k) \to \pi. \tag{40.17}$$

40.3 Marche aléatoire sur $\mathbb{Z}$

Soit (Y_n) une suite de variables aléatoires indépendantes et identiquement distribuées valant -1 avec une probabilité p et 1 avec une probabilité $(1-p)$. La loi est

$$Y_n \sim p\delta_{-1} + (1-p)\delta_1. \tag{40.18}$$

Nous considérons la variable aléatoire

$$X_n = X_0 + \sum_{i=1}^{n} Y_i \tag{40.19}$$

où X_0 est une variable aléatoire indépendante des Y_i à valeurs dans $\mathbb{Z}$. Nous vérifions à présent que X_n est une chaîne de Markov avec comme espace d'états $E = \mathbb{Z}$. Nous devons montrer que

$$P\big(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0\big) = P\big(X_{n+1} = x_{n+1} | X_n = x_n\big). \tag{40.20}$$

Pour ce faire nous allons exprimer tout cela en termes des Y_i au lieu des X_i. D'abord étant donné que nous avons égalité des événements

$$\{X_{n+1} = x_{n+1}\} \cap \{X_n = x_n, \dots, X_0 = x_0\} = \{Y_{n+1} = x_{n+1} - x_n\} \cap \{X_n = x_n, \dots, X_0 = x_0\}, \tag{40.21}$$

nous pouvons, en vertu du principe (38.76), remplacer $X_{n+1} = x_{n+1}$ par $Y_{n+1} = x_{n+1} - x_n$ dans le membre de gauche de (40.20). Nous avons donc déjà

$$P\big(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0\big) = P\big(\underbrace{Y_{n+1} = x_{n+1} - x_n}_{A} \,|\, \underbrace{X_n = x_n, \dots, X_0 = x_0}_{B}\big). \tag{40.22}$$

L'événement B est égal à l'événement

$$\{X_0 = x_0, Y_1 = x_1 - x_0, Y_2 = x_2 - x_1, \dots, Y_n = x_n - x_{n-1}\}, \tag{40.23}$$

qui n'est autre que l'ensemble

$$X_0^{-1}(x_0) \cap Y_1^{-1}(x_1 - x_0) \cap \dots \cap Y_n^{-1}(x_n - x_{n-1}) \tag{40.24}$$

qui est dans la tribu engendrée par les variables aléatoires $X_0, (Y_i)_{i=1,\dots,n}$. Le point délicat du raisonnement est de montrer que les événements A et B donnés par

$$A = \{Y_{n+1} = x_{n+1} - x_n\} \tag{40.25a}$$

$$B = \{X_0 = x_0\} \cap \bigcap_{i=1}^{n} \{Y_i = x_i - x_{i-1}\} \tag{40.25b}$$

sont indépendants. Nous ne pouvons pas montrer directement que $P(A \cap B) = P(A)P(B)$ parce que cela est la formule que nous voulons utiliser pour montrer que la chaîne est de Markov. Nous passons donc par les tribus :

$$A \in \sigma(Y_{n+1}) \tag{40.26a}$$

$$B \in \sigma(X_0, Y_1, \dots, Y_n). \tag{40.26b}$$

Nous utilisons maintenant l'hypothèse d'indépendance des variables aléatoires X_0 et Y_i pour conclure que les deux tribus des équations (40.26) sont indépendantes. Les événements A et B sont par conséquent indépendants.

L'événement A est indépendant de l'événement $\{X_n = x_n\}$. Nous avons donc successivement

$$\begin{aligned}
P\big(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0\big) &= P\big(Y_{n+1} = x_{n+1} - x_n | X_n = x_n, \dots, X_0 = x_0\big) && \text{(40.27a)}\\
&= P\big(Y_{n+1} = x_{n+1} - x_n | Y_i = x_i - x_{i-1}, X_0 = x_0\big) && \text{(40.27b)}\\
&= P(Y_{n+1} = x_{n+1} - x_n) && \text{(40.27c)}\\
&= P(Y_{n+1} = x_{x+1} - x_n | X_n = x_n) && \text{(40.27d)}\\
&= P(Y_{n+1} = x_{n+1} - X_n | X_n = x_n) && \text{(40.27e)}\\
&= P(X_{n+1} = x_{n+1} | X_n = x_n). && \text{(40.27f)}
\end{aligned}$$

Justifications :

— (40.27c) parce que les tribus $\sigma(Y_{n+1})$ et $\sigma(Y_i, X_0)$ sont indépendantes.

— (40.27d) Nous avons

$$\{X_n = x_n\} \in \sigma(X_0, Y_1, \dots, Y_n) \tag{40.28}$$

tandis que

$$\{Y_{n+1} = x_{n+1} - x_n\} \in \sigma(Y_{n+1}); \tag{40.29}$$

ce sont donc deux événements issus de tribus indépendantes. Donc conditionner ou non l'événement $Y_{n+1} = x_{n+1} - x_n$ à l'événement $X_n = x_n$ ne change rien.

— (40.27e) est encore l'utilisation du fait que $P(A|B) = P(K|B)$ dès que $A \cap B = K \cap B$.

La chaîne est par conséquent de Markov.

La matrice de transition de cette chaîne de Markov est une matrice infinie « dans tous les sens » :

$$p(x, y) = \begin{cases} p & \text{si } y = x - 1\\ (1 - p) & \text{si } y = x + 1\\ 0 & \text{sinon.} \end{cases} \tag{40.30}$$

Remarque 40.14.
La plupart du temps lorsqu'il faut démontrer qu'une chaîne est de Markov, il faut suivre la procédure que nous venons de suivre pour la marche aléatoire sur $\mathbb{Z}$.

— Écrire tout en fonction des incréments.
— Dire que les incréments conditionnés sont indépendants des incréments qui conditionnent (via les tribus engendrées).
— Écrire que la probabilité cherchée est égale à l'événement conditionné dans lequel on a juste remplacé l'incrément par sa valeur.
— Conditionner à nouveau par rapport au dernier incrément qui est indépendant.
— Changer la valeur du dernier incrément par la variable aléatoire.

Dans ce raisonnement nous utilisons deux fois le fait que $P(A|B) = P(K|B)$ si $A \cap B = K \cap B$.

40.3.1 Chaînes de Markov homogènes

Proposition 40.15.
Voici quelques propriétés des chaînes de Markov homogènes.

(1) La probabilité d'une trajectoire donnée est

$$P(X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = p(x_{n-1}, x_n) \dots p(x_0, x_1) P(X_0 = x_0). \tag{40.31}$$

(2) La probabilité de transition « en n coups » est donnée par la puissance nième de la matrice de transition :

$$P(X_n = x_n | X_0 = x_0) = Q^n_{x_0, x_n}. \tag{40.32}$$

(3) Si l'espace des états E est fini, l'espérance d'une fonction bornée $f : E \to \mathbb{R}$ de l'état est donnée par

$$E\big(f(X_{n+1}) | X_n = x_n, \dots, X_0 = x_0\big) = E\big(f(X_{n+1}) | X_n = x_n\big) \tag{40.33a}$$
$$= \sum_{y \in E} f(y) p(x_n, y).j \tag{40.33b}$$

Démonstration. Pour (1), étant donné que $P(A \cap B) = P(A|B)P(B)$, nous avons

$$P(X_n = x_n, \dots, X_0 = x_0) = P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_0 = x_0) P(X_{n-1} = x_{n-1}, \dots, X_0 = x_0). \tag{40.34}$$

Par la propriété de Markov, le premier facteur est

$$P(X_n = x_n | X_{n-1} = x_{n-1}) = p(x_{n-1}, x_n). \tag{40.35}$$

Le reste est une récurrence sur n.

Pour (2). Montrons avec $n = 2$. En utilisant les divers points du théorème 38.30, nous avons

$$P(X_2 = x_2 | X_0 = x_0) = \sum_{y \in E} P(X_2 = x_2, X_1 = y | X_0 = x_0) \tag{40.36a}$$
$$= \sum_{y \in E} P(X_2 = x_2 | X_1 = y, X_0 = x_0) P(X_1 = y | X_0 = x_0) \tag{40.36b}$$
$$= \sum_{y \in E} P(X_2 = x_2 | X_1 = y) P(X_1 = y | X_0 = x_0) \tag{40.36c}$$
$$= \sum_{y \in E} p(x_2, y) p(y, x_0) \tag{40.36d}$$
$$= Q^2_{x_2, x_0}. \tag{40.36e}$$

Bien entendu ici la notion de produit matriciel doit être comprise de façon formelle lorsque E est infini.

Remarque 40.16.
Nous avons utilisé l'homogénéité de la chaîne de Markov au moment d'écrire l'expression (40.36d). En principe nous aurions dû écrire $p_2(y, x_2)p_1(x_0, y)$.

Pour (3), ... □

40.3.2 Graphe de transition

Le **graphe de transition** d'une chaîne de Markov est le graphe dont les sommets sont les éléments de l'espace des états de la chaîne et dont les sommets sont reliés par des arrêtes pondérées par la probabilité de transition correspondante.

Définition 40.17.
*Une chaîne de Markov est **irréductible** si pour tout $x, y \in E$, il existe n tel que $p^n(x, y) > 0$ où*

$$p^n(x, y) = P(X_n = y | X_0 = x). \tag{40.37}$$

Le nombre n peut dépendre de x et y.

Lemme 40.18.
Une chaîne de Markov homogène est irréductible si et seulement si son graphe de transition est connexe.

Démonstration. Pour chaque couple $(x, y) \in E^2$ nous avons

$$\begin{aligned} p^n(x, y) &= \sum_{z_i \in E} P(X_n = y, X_{n-1} = z_{n-1}, \dots, X_1 = z_1, X_0 = x) \\ &= \sum_{z_i} p(z_{n-1}, y)p(z_{n-2}, z_{n-1}) \dots p(z_1, z_2)p(x, z_1). \end{aligned} \tag{40.38}$$

La positivité d'un des termes de la somme signifie que le graphe est connexe tandis que la positivité de $p^n(x, y)$ signifie que la chaîne est irréductible. □

40.3.3 Chaîne de Markov définie par récurrence

40.3.3.1 Le cas général

Proposition 40.19.
Soit X_0 une variable aléatoire à valeurs dans E, un ensemble au plus dénombrable. Soit (Y_n) une suite de variables aléatoires réelles indépendantes et identiquement distribuées indépendantes de X_0.

Soit (X_n) la suite de variables aléatoires à valeurs dans E définie par récurrence selon la formule

$$X_{n+1} = G(X_n, Y_{n+1}) \tag{40.39}$$

où $G\colon E \times \mathbb{R} \to E$ est une fonction mesurable. Alors (X_n) est une chaîne de Markov.

Démonstration. Soient $x_0, \dots, x_{n+1}$ des éléments de E. Nous devons calculer la valeur de

$$P(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0). \tag{40.40}$$

Commençons par préciser les espaces sur lesquels nos variable aléatoires sont définies. Nous avons

$$X_0\colon \Omega_0 \to E \tag{40.41}$$

et

$$Y_i\colon \Omega \to \mathbb{R}. \tag{40.42}$$

La variable aléatoire X_1 est donnée par

$$\begin{aligned} X_1 \colon \Omega_0 \times \Omega &\to E \\ (\omega_0, \omega_1) &\mapsto G\big(X_0(\omega_0), Y_1(\omega_1)\big). \end{aligned} \tag{40.43}$$

La variable aléatoire X_2 est

$$\begin{aligned} X_2 \colon \Omega_0 \times \Omega^2 &\to E \\ (\omega_0, \omega_1, \omega_2) &\mapsto G\big(X_1(\omega_0, \omega_1), Y_2(\omega_2)\big) \\ &= G\Big(G\big(X_0(\omega_0), \omega_1\big), Y_2(\omega_2)\Big) \end{aligned} \tag{40.44}$$

et ainsi de suite.

Considérons maintenant l'événement

$$\{X_1 = x_1, X_0 = x_0\} \subset \Omega_0 \times \Omega. \tag{40.45}$$

Il est donné explicitement par

$$\{X_1 = x_1, X_0 = x_0\} = \{(\omega_0, \omega_1) \text{ tel que } G\big(X_0(\omega_0), Y_1(\omega_1) = x_1, X_0(\omega_0) = x_0\big)\} \tag{40.46a}$$
$$= \{(\omega_0, \omega_1) \text{ tel que } G\big(x_0, Y_1(\omega_1) = x_1, X_0(\omega_0) = x_0\big)\} \tag{40.46b}$$
$$= \{\omega_0 \in \Omega_0 \text{ tel que } X_0(\omega_0) = x_0\} \times \{\omega_1 \in \Omega \text{ tel que } G\big(x_0, Y_1(\omega_1) = x_1\big)\}. \tag{40.46c}$$

Le premier terme du produit cartésien est dans $\sigma(X_0)$, tandis que le second est dans $\sigma(Y_1)$. Étant donné la définition des tribus produit (définition 16.157) nous avons

$$\{X_1 = x_1, X_0 = x_0\} \in \sigma(X_0, Y_1). \tag{40.47}$$

Ce raisonnement se généralise immédiatement et nous trouvons que

$$\{X_n = x_n, \dots, X_0 = x_0\} \in \sigma(X_0, Y_1, \dots, Y_n). \tag{40.48}$$

Nous sommes donc à calculer

$$\diamondsuit = P(X_{n+1} = x_{n+1} | X_n = X_n, \dots, X_0 = X_0) \tag{40.49a}$$
$$= P\big(\underbrace{G(x_n, Y_{n+1}) = x_{n+1}}_{\in\sigma(Y_{n+1})} | \underbrace{X_n = x_n, \dots, X_0 = x_0}_{\in\sigma X_0, Y_1, \dots, Y_n}\big). \tag{40.49b}$$

Les tribus $\sigma(Y_{n+1})$ et $\sigma(X_0, Y_1, \dots, Y_n)$ étant indépendantes nous avons

$$\diamondsuit = P\big(G(x_n, Y_{n+1}) = x_{n+1}\big) \tag{40.50a}$$
$$= P\big(G(x_n, Y_{n+1}) = x_{n+1} | X_n = X_n\big) \tag{40.50b}$$
$$= P\big(G(X_n, Y_{n+1}) = x_{n+1} | X_n = x_n\big) \tag{40.50c}$$
$$= P(X_{n+1} = x_{n+1} | X_n = x_n). \tag{40.50d}$$

Pour (40.50b) nous avons utilisé le fait que $\sigma(Y_{n+1})$ est indépendante de $\sigma(X_n)$. Nous avons prouvé que la chaîne était de Markov. □

Les probabilités de transition de la chaîne de Markov définie dans la proposition 40.19 sont

$$P(X_1 = y | X_0 = x) = P\big(G(X_0, Y_1) = y | X_0 = x_0\big) = P\big(G(x_0, Y_1) = y\big). \tag{40.51}$$

40.3.3.2 Exemple : la file de réparation de machines à laver

Nous considérons un magasin de réparation d'électroménager. Durant le jour n, un nombre aléatoire Z_n de machines en panne arrivent au magasin. Une machine est réparée chaque jour (aucune si le magasin est vide). Nous supposons que les Z_n soient indépendantes et identiquement distribuées, et nous posons X_n, le nombre de machines en magasin le jour n.

La loi d'avancement de X_n est

$$X_{n+1} = \begin{cases} X_n + Z_n - 1 & \text{si } X_n \neq 0 \\ Z_n & \text{si } X_n = 0. \end{cases} \tag{40.52}$$

Cela est une chaîne de Markov en vertu de la proposition 40.19. Ici la fonction est

$$G(x, y) = x + y - \mathbb{1}_{x\neq 0}. \tag{40.53}$$

Les probabilités de transitions sont

$$p(x, y) = \begin{cases} 0 & \text{si } x \leqslant y - 2 \\ P(Z = 0) & \text{si } x = y - 1 \\ P(Z = k) & \text{si } x = y + k - 1 \end{cases} \tag{40.54}$$

pour $x \neq 0$.

Exemple 40.20

Soit (X_n) une chaîne de Markov de matrice de transition

$$P = \begin{pmatrix} 0.2 & 0.5 & 0.3 \\ 0.1 & 0.1 & 0.8 \\ 0.5 & 0.2 & 0.3 \end{pmatrix}. \tag{40.55}$$

Calculer $P(X_3 = 1|X_0 = 1)$ et $P(X_7 = 0|X_4 = 0)$.

Déterminer, s'il en existe, une loi stationnaire vers laquelle converge la chaîne.

Nous avons

$$P^3 = \begin{pmatrix} 0.344 & 0.251 & 0.405 \\ 0.283 & 0.307 & 0.41 \\ 0.287 & 0.248 & 0.465 \end{pmatrix}. \tag{40.56}$$

La probabilité d'aller de l'état 1 à l'état 1 en trois étapes est donc 0.307. La chaîne étant de Markov, sans mémoire, les probabilités entre les temps 4 et 7 sont les mêmes qu'entre 0 et 3. Nous avons alors

$$P(X_7 = 0|X_4 = 0) = 0.344. \tag{40.57}$$

La chaîne est irréductible et n'a pas d'états absorbants.

△

40.4 Classification des états

Sauf mention expresse du contraire, nous considérons toujours une chaîne de Markov homogène.

Définition 40.21.
*Un état $x \in E$ est **absorbant** pour la chaine (X_n) si $p(x, x) = 1$.*

Il n'est pas spécialement impossible d'arriver sur un état absorbant, mais il est impossible d'en sortir.

Si $x \in E$, nous notons

$$T(x) = \inf\{k \geqslant 1 \text{ tel que } X_k = x\}, \tag{40.58}$$

le **premier temps d'atteinte** de l'état x. Si $X_0 = x$, alors $T(x)$ est le **temps de retour** en x. Si $p \in \mathbb{N}$ nous notons

$$T_p(x) = \inf\{k \geqslant 1 \text{ tel que } X_{k+p} = x\}. \tag{40.59}$$

C'est le temps mis pour atteindre x à partir de l'instant p.

Proposition 40.22.
La loi de la variable aléatoire $[T_p(x)|X_p = x]$ est la même que celle de la variable aléatoire $[T(x)|X_0 = x]$.

Démonstration. Nous devons montrer que

$$P(T_p(x) = k|X_p = x) = P(T(x) = k|X_0 = x). \tag{40.60}$$

Cela est intuitivement évident du fait qu'une chaîne de Markov soit un processus sans mémoire. Afin de prouver, nous allons sommer sur tous les états intermédiaires possibles :

$$\begin{aligned}
P(T_p(x) = k|X_0 = x) &= P(X_{p+k} = x, X_{p+k-1} \neq x, \dots, X_{p+1} \neq x|X_p = x) && \text{(40.61a)}\\
&= \sum_{z_i \neq x} P(X_{p+k} = x, X_{p+k-1} = z_{k-1}, \dots, X_{p+1} = z_1|X_p = x) && \text{(40.61b)}\\
&= \sum_{z_i} P(X_{p+k} = x, X_{p+k-i} = z_i|X_{p+1} = z_1, X_p = x) \underbrace{P(X_{p+1} = z_1|X_p = x)}_{=p(x,z_1)} && \text{(40.61c)}\\
&= \sum_{z_i} P(X_{p+k} = x, X_{p+k-i} = z_i|X_{p+2} = z_2, X_{p+1} = z_1, X_p = x) && \text{(40.61d)}\\
&\quad \underbrace{P(X_{p+2} = z_2|X_{p+1} = z_1, X_p = x)}_{P(X_{p+2}=z_2|X_{p+1}=z_1)=p(z_1,z_2)} p(x, z_1) && \text{(40.61e)}\\
&= \dots && \text{(40.61f)}\\
&= \sum_{z_i} p(x, z_1)p(z_1, z_2) \dots p(z_{k-1}, z_{k-1})p(z_{k-1}, x). && \text{(40.61g)}
\end{aligned}$$

À ce point ci, nous avons éliminé toute référence à p grâce à l'homogénéité de la chaîne. Nous pouvons refaire le calcul à l'envers pour reconstituer l'expression de départ sans le p :

$$\begin{aligned}
&\sum_{z_i} p(x, z_1)p(z_1, z_2) \dots p(z_{k-1}, z_{k-1})p(z_{k-1}, x) && \text{(40.62a)}\\
&\qquad = P(x_k = x, X_{k-1} \neq x, \dots, X_1 \neq x|X_0 = x) && \text{(40.62b)}\\
&\qquad = P(T(x) = k), && \text{(40.62c)}
\end{aligned}$$

ce qu'il fallait obtenir. □

Définition 40.23.
*Un état x est **récurrent** si $P(T(x) = \infty|X_0 = x) = 0$, c'est à dire si la probabilité de ne jamais retourner en x lorsqu'on y est passé est nulle. L'état x est **transient** ou **transitoire** dans le cas contraire.*

*Si x est un état récurrent, et si $E\big(T(x)|X_0 = x\big) < \infty$, nous disons que x est **récurrent positif**. Si $E\big(T(x)|X_0 = x\big) = \infty$ alors nous disons que est **récurrent nul**.*

Nous introduisons une variable aléatoire qui compte le nombre de fois que la chaîne de Markov passe par l'état x :

$$N_x = \sum_{k=0}^{\infty} \mathbb{1}_{\{X_k=x\}}. \tag{40.63}$$

C'est une variable aléatoire à valeurs dans $\mathbb{N} \cup \{\infty\}$.

Proposition 40.24.
Les deux propriétés suivantes sont équivalentes à dire que x est récurrent :

(1) $P(N_x < \infty | X_0 = x) = 0$

(2) $E(N_x | X_0 = x) = \infty$.

Les deux propriétés suivantes sont équivalentes à dire que x est transient :

(1) $P(N_x < \infty | X_0 = x) = 1$

(2) $E(N_x | X_0 = x) < \infty$.

Démonstration. En tant que événements, nous avons l'égalité

$$N_x < \infty = \bigcup_{n\in\mathbb{N}} \{\underbrace{X_n = x, X_{n+k} \neq x \forall k \geqslant 1}_{F_n}\}. \tag{40.64}$$

Nous avons donc

$$P(N_x < \infty | X_0 = x) = \sum_{n=0}^{\infty} P(F_n | X_0 = x), \tag{40.65}$$

et

$$\begin{aligned}
P(F_n | X_0 = x) &= P(X_{n+k} \neq x, \forall k \geqslant 1, X_n = x | X_0 = x) && \text{(40.66a)}\\
&= P(X_{n+k} \neq x, k \geqslant 1 | X_n = x, X_0 = x) P(X_n = x | X_0 = x) && \text{(40.66b)}\\
&= P(X_{n+k} \neq x, k \geqslant 1 | X_n = x) P(X_n = x | X_0 = x) && \text{(40.66c)}\\
&= P(X_k \neq x, k \geqslant 1 | X_0 = x) P(X_n = x | X_0 = x) && \text{(40.66d)}\\
&= P(T(x) = \infty | X_0 = x) P(X_n = x | X_0 = x) && \text{(40.66e)}
\end{aligned}$$

Justifications :

(1) Pour (40.66c), nous utilisons le fait que la chaîne soit « sans mémoire ».

(2) Pour (40.66d), nous utilisons le fait que la chaîne soit homogène.

(3) Pour (40.66e), l'événement $X_k \neq x$ pour tout $k \geqslant 1$ est exactement l'événement $T(x) = \infty$.

En nous servant de la proposition 16.243 (théorème de Fubini et mesure de comptage), nous permutons l'espérance et la somme dans l'expression

$$\begin{aligned}
\sum_{n=0}^{\infty} P(X_n = x | X_0 = x) &= \sum_{n=0}^{\infty} E(\mathbb{1}_{\{X_n = x\}} | X_0 = x) && \text{(40.67a)}\\
&= E\Big(\sum_{n=0}^{\infty} \mathbb{1}_{\{X_n = x\}} | X_0 = x \Big) && \text{(40.67b)}\\
&= E(N_x | X_0 = x). && \text{(40.67c)}
\end{aligned}$$

Voyons ce passage plus en détail. D'abord, en général nous avons

$$E(Y | X = x_0) = \int_{\{X = x_0\}} Y(\omega) dP(\omega) = \int_{\Omega} \mathbb{1}_{\{X = x_0\}}(\omega) Y(\omega) dP(\omega). \tag{40.68}$$

Dans notre cas,

$$E\big(\mathbb{1}_{\{X_n = x\}} | X_0 = x\big) = \int_{\Omega} \mathbb{1}_{X_0 = x}(\omega) \mathbb{1}_{\{X_n = x\}}(\omega) dP(\omega). \tag{40.69}$$

La fonction qui correspond à la proposiiton 16.243 est

$$f(n, \omega) = f_n(\omega) = \delta_{X_0(\omega), x} \delta_{X_n(\omega), x}, \tag{40.70}$$

qui est bien une fonction positive et mesurable.

Nous reprenons à présent le calcul (40.65) en remplaçant les éléments par leurs valeurs que nous avons calculées :

$$P(N_x < \infty | X_0 = x) = P\big(T(x) = \infty | X_0 = x\big) E(N_x | X_0 = x). \tag{40.71}$$

Si x est récurrent, nous avons $P\big(T(x) = \infty | X_0 = x\big) = 0$, mais la relation (40.71) ne permet pas de conclure que le membre de gauche est nul parce qu'il reste la possibilité que $E(N_x | X_0 = x) = \infty$. Nous devons donc faire un pas en arrière et écrire cette espérance comme la limite des sommes partielles :

$$P(N_x < \infty | X_0 = x) = \lim_{N\to\infty} \sum_{n=0}^{N} P\big(T(x) = \infty | X_0 = x\big) P(X_n = x | X_0 = x) = 0 \tag{40.72}$$

parce que tous les termes de la suite des sommes partielles sont nuls. Nous avons donc bien que $P(N_x < \infty | X_0 = x) = 0$. Il s'ensuit immédiatement que $E(N_x | X_0 = x) = 1$.

Nous devons maintenant démontrer l'implication inverse. Supposons que $P(N_x < \infty | X_0 = x) = 0$. Dans ce cas nous avons immédiatement $P(N_x = \infty | X_0 = x) = 1$ et $E(N_x | X_0 = x) = \infty$. L'équation (40.71) nous indique alors que

$$P\big(T(x) = \infty | X_0 = x\big) = 0, \tag{40.73}$$

c'est à dire que x est récurrent. □

40.4.1 Chaînes irréductibles

Proposition 40.25.
Soit (X_n) une chaîne de Markov irréductible.

(1) Un état x est récurrent si et seulement si tous les états sont récurrents.

(2) Un état x est transient si et seulement si tous les états sont transients.

Démonstration. Soient x et y des états de la chaîne de Markov. Nous devons tester la valeur de $P(X_n = y | X_0 = y)$. Afin d'exploiter l'hypothèse d'irréductibilité, nous considérons $r, s \in \mathbb{N}$ tels que

$$p^r(x, y) > 0 \tag{40.74a}$$
$$p^s(y, x) > 0 \tag{40.74b}$$

et nous calculons majorons en passant par quelques intermédiaires :

$$\begin{aligned} P(X_{n+r+s} = y | X_0 = y) &\geqslant P(X_{n+r+s} = y, X_{n+s} = x, X_s = x | X_0 = y) && (40.75a)\\ &= P(X_{n+r+s} = y | X_{n+s} = x, X_s = x, X_0 = y) && (40.75b)\\ &\quad P(X_{n+s} = x | X_s = x, X_0 = y) P(X_s = x | X_0 = y). \end{aligned}$$

Les deux premiers facteurs se calculent en utilisant la propriété de Markov et l'homogénéité de la chaîne. Pour le premier,

$$P(X_{n+s} = x | X_s = x, X_0 = y) = P(X_{n+s} = x | X_s = x) = P(X_n = x | X_0 = x). \tag{40.76}$$

Nous avons donc

$$\sum_{n\in\mathbb{N}} P(X_{n+r+s} = y | X_0 = y) \geqslant p^r(x, y) p^s(y, x) \sum_{n\in\mathbb{N}} P(X_n = x | X_0 = x). \tag{40.77}$$

En réutilisant Fubini comme dans l'équation (40.67), nous avons

$$\sum_{n\in\mathbb{N}} P(X_{n+r+s} = y | X_0 = y) \geqslant K E(N_x | X_0 = x) \tag{40.78}$$

où K est une constante strictement positive, par hypothèse d'irréductibilité de la chaîne de Markov.

Si x est un état récurrent, alors le membre de gauche est infini par la proposition (40.24) et donc

$$\sum_{n\in\mathbb{N}} P(X_{n+r+s}=y|X_0=y)=\infty. \tag{40.79}$$

Aux $r+s$ premiers termes près (qui ne changent pas la somme), nous avons

$$\sum_{n\in\mathbb{N}} P(X_n=y|X_0=y)=\infty, \tag{40.80}$$

ce qui signifie que y est récurrent. □

Nous rappelons que $T(x)$ est le temps que première atteinte de l'état x. Nous notons

$$\pi(x)=\frac{1}{E\big(T(x)|X_0=x\big)}. \tag{40.81}$$

Étant donné que $T(x)$ est un entier positif ou nul nous avons $E\big(T(x)|X_0=x\big)\in[1,\infty]$ et donc $\pi(x)\in[0,1]$.

Si x est un état transient, alors $T(x)=\infty$ lorsque $X_0=x$ et donc $E\big(T(x)|X_0=x\big)=0$ et $\pi(x)=0$. Si x est récurrent par contre, $P\big(T(x)<\infty|X_0=x\big)=1$ et il n'y a pas de garanties sur la valeur de $E\big(T(x)|X_0=x\big)$.

Corollaire 40.26.
Un état récurrent est récurrent positif si et seulement si $\pi(x)>0$. Un état récurrent est récurrent nul si et seulement si $\pi(x)=0$.

Démonstration. C'est la formule (40.81). □

Proposition 40.27.
Soit (X_n) est une chaîne de Markov irréductible.

(1) Si x est un état récurrent, alors $T(X)<\infty$ presque sûrement.

(2) Nous avons une égalité entre les lois

$$\mathscr{L}\big(X_{k+T(x)}|T(x)<\infty\big)=\mathscr{L}(X_k|X_0=x). \tag{40.82}$$

40.4.2 Nombre de visites

La fonction

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_{\{X_k=x\}} \tag{40.83}$$

est la **fréquence empirique** de la chaîne de Markov.

Soit x un état récurrent, c'est à dire que $P\big(T(x)<\infty|X_0=x\big)=1$. Nous classons les visites de la façon suivante :

$$T_1(x)=T(x)=\inf\{k\geqslant 1 \text{ tel que } X_k=x\} \tag{40.84a}$$
$$T_2(x)=\inf\{k\geqslant 1 \text{ tel que } X_{T_1(x)+k}=x\} \tag{40.84b}$$
$$\vdots \tag{40.84c}$$
$$T_n(x)=\inf\{k\geqslant 1 \text{ tel que } X_{T_{n-1}(x)+k}=x\} \tag{40.84d}$$

La variable aléatoire T_i représente le temps entre la visite numéro $i-1$ et la visite numéro i (si $X_0\neq x$, sinon il faut décaler). Nous définissons l'instant la na visite numéro n :

$$S_n=\sum_{k=1}^{n}T_k(x). \tag{40.85}$$

Lemme 40.28.
Les variables aléatoires T_i sont indépendantes.

Démonstration. Nous choisissons n des T_i et nous calculons la probabilité

$$\spadesuit = P(T_{i_1} = k_1, T_{i_2} = k_2, \dots, T_{i_n} = k_n) \tag{40.86}$$

où nous supposons $i_1 > i_2 > \dots > i_n$. Nous décomposons cette probabilité en sommant sur toutes les histoires de la chaîne de Markov compatibles avec les nombres k_i donnés :

$$\spadesuit = \sum_{\substack{\{z_j\}\\ \text{compatibles}}} P(X_j = z_j, j = 1, \dots, N). \tag{40.87}$$

Notons qu'ici, le numéro du dernier terme de la somme n'est pas certain parce que tous les T_i ne sont pas fixés. Nous l'avons noté N, mais en réalité il est différent d'un terme à l'autre de la somme. Il est certain que $z_N = x$ et $z_{N-k_1} = x$ et si $N - k_1 < j < N$, alors $z_j \neq x$. Cela est simplement le fait que nous demandions aux z_i de respecter les conditions données par les k_i. Nous avons

$$\spadesuit = \sum_{\{z_j\}} P(X_N = x, X_j = z_j, N - k_1 < j < N | X_j = z_j, j \leqslant N - k_1) P(X_j = z_j, j < N - k_1) \tag{40.88a}$$

$$= \sum_{\{z_j\}} P(X_N = x, X_j = z_j, N - k_1 < j < N | X_{N-k_1} = x) P(X_j = z_j, j < N - k_1) \tag{40.88b}$$

$$\tag{40.88c}$$

Le premier facteur est $P(T_{i_1} = k_1)$ tandis que le second facteur est précisément $P(T_j = k_j, j > 1)$. Nous avons donc montré que

$$P(T_{i_1} = k_1, T_{i_2} = k_2, \dots, T_{i_n} = k_n) = P(T_{i_1} = k_1) P(T_j = k_j, j > 1), \tag{40.89}$$

et donc les T_i sont indépendants. □

Proposition 40.29.
Si (X_n) est une chaîne de Markov irréductible et si $x \in E$ alors

$$\pi(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{X_k = x\}} \tag{40.90}$$

presque sûrement.

Démonstration. Étant donné que la chaîne est irréductible, les états sont soit tous transient soit tous récurrents par la proposition 40.25. Nous commençons par considérer que x est transient.

En comparant la définition (40.63) de N_x et le membre de droite de (40.90), nous avons pour chaque n l'inégalité

$$\frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{X_k = x\}} \leqslant \frac{1}{n} E(N_x). \tag{40.91}$$

Dans le cas d'un élément transient, nous avons $\pi(x) = 0$, donc il serait bon de montrer que $E(N_x) < \infty$, de sorte que prendre la limite $n \to \infty$ dans (40.91) donne zéro.

Nous décomposons le calcul en deux morceaux :

$$E(N_x) = E\big(N_x | T(x) = \infty\big) P\big(T(x) = \infty\big) + E\big(N_x | T(x) < \infty\big) P\big(T(x) < \infty\big). \tag{40.92}$$

Le fait que le premier terme soit fini découle immédiatement du fait que $T(x) = \infty$ implique $X_k \neq x$ pour tout $k \geqslant 1$. Dans ce cas l'espérance de N_x est évidemment finie.

Pour le second terme nous avons

$$E\big(N_x|T(x)<\infty\big)=E\Big(\sum_{k=0}^{\infty}\mathbb{1}_{\{X_k=x\}}|T(x)<\infty\Big) \tag{40.93a}$$

$$=\sum_{k=1}^{\infty}E\big(\mathbb{1}_{\{X_k=x\}}|T(x)<\infty\big). \tag{40.93b}$$

Pour inverser la somme et l'espérance, nous avons utilisé le théorème de théorème de Fubini-Tonelli qui est encore valable pour des fonctions qui prennent la valeur ∞. Le fait d'inverser ne signifie pas que ni la somme ni l'intégrale soit finie. D'ailleurs c'est exactement ce que nous sommes en train de déterminer.

Étant donné que nous voulons seulement savoir si cette somme est finie ou non, nous pouvons nous restreindre à la somme depuis $k=1$ ou oublier le premier terme. D'autre par nous avons

$$\sum_{k=1}^{\infty}\mathbb{1}_{\{X_k=x\}}=\sum_{j=0}^{\infty}\mathbb{1}_{\{X_{j+T(x)}=x\}} \tag{40.94}$$

parce que les $T(x)$ premiers termes sont par définition nuls. Nous regardons donc

$$\sum_{j=0}^{\infty}E\big(\mathbb{1}_{X_{j+T(x)}=x}|T(x)<\infty\big)=\sum_j P\big(X_{j+T(x)}=x|T(x)<\infty\big) \tag{40.95a}$$

$$=\sum_j P(X_j=x|X_0=x) \tag{40.95b}$$

$$=\sum_j E\big(\mathbb{1}_{\{X_j=x\}}|X_0=x\big) \tag{40.95c}$$

$$=E\Big(\sum_j\mathbb{1}_{X_j=x}|X_0=x\Big) \tag{40.95d}$$

$$=E(N_x|X_0=x) \tag{40.95e}$$

$$<\infty \qquad \text{parce que } x \text{ est transient.} \tag{40.95f}$$

L'équation (40.95b) provient de la proposition 40.27 et plus précisément de l'égalité entre les lois (40.82). Nous avons terminé la preuve dans le cas où x est transient.

Nous passons maintenant au cas où x est récurrent, c'est à dire $P(T(x)<\infty|X_0=x)=1$. Les variables aléatoires T_i définies en (40.84) pour $i\geqslant 2$ sont indépendantes et identiquement distribuées et

$$\mathscr{L}\big(T_k(x)\big)\sim\mathscr{L}\big(T(X)|X_0=x\big). \tag{40.96}$$

La loi des grands nombres nous indique que

$$\frac{S_n}{n}=\frac{T_1(x)}{n}+\frac{1}{n}\sum_{k=2}^{n}T_k(x)\xrightarrow{p.s.}E\big(T_2(x)\big) \tag{40.97a}$$

$$=E\big(T(x)|X_0=x\big). \tag{40.97b}$$

Remarque 40.30.
La loi des grands nombres est encore vraie sans l'hypothèse de variables aléatoires dans L^1 pourvu qu'elles soient positives. Alors dans la conclusion de la loi nous devons accepter la possibilité que l'espérance soit infinie.

Nous posons pour $m\in\mathbb{N}$

$$n(m)=\sum_{j=1}^{m}\mathbb{1}_{\{X_j=x\}} \tag{40.98}$$

qui est le nombre de visites de x avant l'instant m. Nous avons évidemment $n(m) \leqslant m$. Mais S_n est l'instant de la $n$ième visite, par conséquent $S_{n(m)}$ est l'instant de la dernière visite avant le moment m. Pour tout m nous avons les inégalités

$$S_{n(m)} \leqslant m < S_{n(m)+1}. \tag{40.99}$$

Nous divisons par $n(m)$ et nous effectuons la limite $m \to \infty$:

$$\frac{S_n(m)}{n(m)} \leqslant \frac{m}{n(m)} \leqslant \frac{S_{n(m)}+1}{n(m)} \tag{40.100}$$

En ce qui concerne la limite de $n(m)$, nous utilisons la définition (40.98) :

$$n(m) \to \sum_{j=1}^{\infty} \mathbb{1}_{\{X_j=x\}} = \tag{40.101}$$

heur...

$$\lim_{m\to\infty} n(m) = \lim_{m\to\infty} \sum_{n=1}^{m} \mathbb{1}_{\{X_j=x\}} \xrightarrow{p.s.} \infty \tag{40.102}$$

par la proposition (40.24). Plus précisément, la limite vaut N_x qui vaut presque sûrement ∞ dans le cas où x est récurrent. Par ailleurs la loi des grands nombres (40.97) nous enseigne en particulier que

$$\frac{S_{n(m)}}{n(m)} \xrightarrow{p.s.} E\big(T(x)|X_0 = x\big). \tag{40.103}$$

Le terme de droite dans (40.100) se traite de façon usuelle :

$$\frac{S_{n(m)+1}}{n(m)} = \frac{S_{n(m)+1}}{n(m)+1}\frac{n(m)+1}{n(m)}. \tag{40.104}$$

Le dernier facteur tend vers 1 et le tout a pour limite $E\big(T(x)|X_0 = x\big)$. Par conséquent nous avons

$$\frac{m}{n(m)} \xrightarrow{p.s.} E\big(T(x)|X_0 = x\big) \tag{40.105}$$

et

$$\frac{n(m)}{n} = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{\{X_j=x\}} \to \frac{1}{E\big(T(x)|X_0 = x\big)} = \pi(x). \tag{40.106}$$

□

Lemme 40.31.
Soit (X_k) une chaîne de Markov dont l'espace des états est noté E. Pour chaque $x \in E$ nous notons

$$T(x) = \inf\{k \geqslant 1 \text{ tel que } X_k = x\} \tag{40.107}$$

et

$$T_p(x) = \inf\{k \geqslant 1 \text{ tel que } X_{k+p} = x\} \tag{40.108}$$

Alors nous avons

$$P(T_p(x) = k|X_p = y) = P(T(x) = k|X_0 = y). \tag{40.109}$$

La proposition suivante nous permet de parler de chaîne de Markov **récurrence positive**.

Proposition 40.32.
Soit (x_n) une chaîne de Markov irréductible.

(1) Un état x est transient si et seulement si tous les états sont transients.

(2) Un état est récurrent positif si et seulement si tous les états sont récurrents positifs.

Démonstration. Nous rappelons (proposition 40.29) que si la chaîne est irréductible

$$\pi(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{[X_k=x]} \tag{40.110}$$

Notons aussi que

$$\sum_{k=1}^{N} \mathbb{1}_{X_k=x} = \begin{cases} 0 & \text{si } N < T(x) \\ \sum_{k=0}^{N-T(x)} \mathbb{1}_{X_{k+T(x)}=x} & \text{si } N > T(x) \end{cases} \tag{40.111}$$

où dans la seconde ligne nous avons effectué le changement de variable de sommation $k' = k+T(x)$. Dans la limite (40.110) nous sommes toujours dans le cas où N est assez grand. Nous pouvons donc écrire

$$\pi(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-T(x)} \mathbb{1}_{X_{k+T(x)}=x}. \tag{40.112}$$

Nous pouvons aussi écrire

$$\frac{1}{N-T(x)} \sum_{k=0}^{N-T(x)} \mathbb{1}_{X_{k+T(x)}=x} = \frac{N}{N-T(x)} \frac{1}{N} \sum_{k=0}^{N-T(x)} \mathbb{1}_{X_{k+T(x)}=x}. \tag{40.113}$$

Dans cette dernière égalité le membre de droite tend vers $\pi(x)$ et nous avons

$$\lim_{N\to\infty} \frac{1}{N-T(x)} \sum_{k=0}^{N-T(x)} \mathbb{1}_{X_{k+T(x)}=x} = \pi(x) \tag{40.114}$$

ou encore

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} \mathbb{1}_{X_{k+T(x)}=x} = \pi(x) \tag{40.115}$$

Étant donné que $\pi(x)$ est une constante nous avons évidemment $E(\pi(x)) = \pi(x)$. Nous pouvons cependant considérer les variables aléatoires

$$Z_n = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{X_{k+T(x)}=x} \tag{40.116}$$

et remarquer que $Z_n \xrightarrow{p.s.} \pi(x)$ avec $0 \leqslant Z_n \leqslant 1$. Le théorème de la convergence dominée (16.140) nous permet d'inverser la limite et l'espérance et écrire

$$\pi(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} E\big(\mathbb{1}_{X_{k+T(x)}=x}\big) \tag{40.117a}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P\big(X_{k+T(x)} = x\big). \tag{40.117b}$$

Par le lemme 40.31 nous avons

$$P(X_{k+T(x)} = x) = P(X_k = k|X_0 = x) \tag{40.118}$$

et $\pi(x)$ prend la forme

$$\pi(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} P(X_k = x|X_0 = x). \tag{40.119}$$

Soit maintenant un état x positif récurrent et y, un autre état. Par définition 40.23 et par corollaire 40.26 nous avons $\pi(x) > 0$. Nous devons prouver que $\pi(y) > 0$.

Étant donné que la chaîne est irréductible il existe r et s tels que

$$\begin{cases} p^r(x,y) = P(X_r = y|X_0 = x) > 0 & \text{(40.120a)} \\ p^s(x,y) = P(X_s = x|X_0 = y) > 0 & \text{(40.120b)} \end{cases}$$

Nous reprenons l'équation (40.77) multipliée par $1/N$:

$$\frac{1}{N}\sum_{n=1}^{N} P(X_{r+s+n=y|X_0=y}) \geqslant \underbrace{p^r(x,y)p^s(y,x)}_{\geqslant 0} \underbrace{\frac{1}{N}\sum_{n=1}^{N} P(X_n = x|X_0 = x)}_{\to\pi(x)} \tag{40.121}$$

et nous prenons la limite lorsque $N \to \infty$. À $r+s$ termes près, nous trouvons à gauche l'expression (40.119) de $\pi(y)$. Par conséquent

$$\pi(y) \geqslant \lim_{N\to\infty} \frac{1}{n}\sum_{n=1}^{N} P(X_{r+s+n} = y|X_0 = y) \geqslant \alpha\pi(x) \tag{40.122}$$

où α est une constante positive. Le nombre $\pi(x)$ étant strictement positif par hypothèse nous avons montré que $\pi(y) > 0$, c'est à dire que y est récurrent positif. □

40.5 Mesure invariante

Définition 40.33.
Une mesure de probabilité μ sur l'espace des états E d'une chaîne de Markov est ***invariante*** *si pour tout $x \in E$*

$$\mu(x) = \sum_{y\in E} p(y,x)\mu(y). \tag{40.123}$$

Remarque 40.34.
Une mesure invariante est une mesure de probabilité et nous noterons par abus $\mu(x)$ pour $\mu(\{x\})$. Si $A \subset E$ nous avons

$$\mu(A) = \sum_{x\in A} \mu(x). \tag{40.124}$$

Remarque 40.35.
Une loi invariante associée à une chaîne de Markov est une loi associée à la matrice de transition de la chaîne, mais pas à la loi de X_0. Par conséquent nous pouvons tester si μ est une mesure invariante pour une certaine chaîne de Markov (X_k) en considérant la chaîne (Y_k) avec $Y_k = X_k$ pour $k > 0$ et Y_0 arbitraire.

L'adjectif *invariant* provient du lemme suivant.

Lemme 40.36.
Soit (X_n) une chaîne de Markov telle que $X_0 \sim \mu$ où μ est une mesure invariante sur l'espace des états. Alors $X_k \sim \mu$ pour tout k.

Démonstration. Par hypothèse, $P(X_0 = x) = \mu(x)$. Ensuite nous avons

$$P(X_1 = y) = \sum_{x\in E} P(X_1 = y|X_0 = x)P(X_0 = x) \tag{40.125a}$$

$$= \sum_x p(x,y)\mu(x) \tag{40.125b}$$

$$= \mu(y). \tag{40.125c}$$

Par conséquent X_1 suit également la loi μ. Par récurrence tous les états suivent cette même loi. □

Si les états d'une chaîne de Markov ont comme loi une mesure invariante, alors nous disons que la chaîne est **stationnaire**.

Remarque 40.37.
Pour une chaîne de Markov stationnaire de loi invariante μ nous avons

$$\mu(x) = \sum_y p(y,x)\mu(y) \tag{40.126}$$

et si l'ensemble E est fini cette équation signifie

$$\mu = Q\mu \tag{40.127}$$

où Q est la matrice de transition de la chaîne de Markov.

Théorème 40.38 (Théorème ergodique)**.**
Une chaîne de Markov irréductible est positive récurrente si et seulement si elle accepte une mesure invariante. Cette mesure est invariante est alors unique et vérifie $\mu = Q\mu$ où Q est la matrice de transition.

Démonstration. Nous allons seulement prouver le théorème ergodique dans le cas où E est fini. Soit (X_n) une chaîne de Markov récurrente positive ; nous avons $\pi(x) > 0$ pour tout $x \in E$. Nous allons montrer que π est une mesure invariante.

Nous commençons par montrer que

$$\sum_{x\in E} \pi(x) = 1. \tag{40.128}$$

Pour cela nous reprenons la propriété de chaîne irréductible pour écrire

$$\pi(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}_{X_k=x} \tag{40.129}$$

Étant donné que E est fini nous pouvons sommer sur $x \in E$ et permuter la somme avec la limite :

$$\sum_{x\in E} \pi(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \underbrace{\sum_{x\in E} \mathbb{1}_{X_k=x}}_{=1}. \tag{40.130}$$

Nous nous retrouvons donc avec $\lim_{N\to\infty} \frac{1}{N} N = 1$. La fonction π définit donc bien une mesure de probabilité sur E.

Nous montrons à présent que cette mesure est invariante, c'est à dire que

$$\pi(x) = \sum_{y\in E} p(y, x)\pi(y). \tag{40.131}$$

Pour cela nous utilisons encore le théorème de la convergence dominée pour permuter la limite et l'intégrale dans

$$\pi(x) = E(\pi(x)) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \underbrace{E\big(\mathbb{1}_{X_k=x}\big)}_{P(X_k=x)} = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} P(X_{k+1} = x). \tag{40.132}$$

La dernière égalité découle du fait que en divisant par N et en faisant tendre N vers l'infini, le fait d'enlever un terme à la somme ne change pas la valeur de la limite. Nous pouvons substituer dans (40.132) la valeur

$$P(X_{k+1} = x) = \sum_{y\in E} p(y, x)P(X_k = y). \tag{40.133}$$

Nous avons alors

$$\pi(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \sum_{y\in E} p(y, x)P(X_k = y) \tag{40.134a}$$

$$= \sum_{y\in E} p(y, x) \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} P(X_k = y) \tag{40.134b}$$

$$= \sum_{y\in E} p(y, x)\pi(y), \tag{40.134c}$$

ce qui signifie que π est une mesure invariante. Notons que nous avons encore utilisé le fait que E soit fini pour permuter avec la limite.

Il nous reste à montrer l'unicité de la mesure invariante sur la chaîne de Markov. Soit μ une mesure invariante pour la chaîne de Markov (X_k). Comme indiqué dans la remarque 40.35 nous pouvons supposer que X_0 suit la loi μ. Par le lemme 40.36 nous avons $P(X_k = x) = \mu(x)$ pour tout k. Par conséquent

$$\pi(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} P(X_k = x) = \mu(x). \tag{40.135}$$

□

Théorème 40.39 (loi des grands nombres pour les chaîne de Markov)**.**
Soit (X_n) une chaîne de Markov irréductible acceptant une mesure invariante. Soit $f : E \to \mathbb{R}$ une fonction dans $L^1(E, \mu)$. Alors nous avons

$$\frac{1}{N} \sum_{k=1}^{N} f(X_k) \xrightarrow{p.s.} \sum_{x\in E} f(x)\mu(x). \tag{40.136}$$

En ce qui concerne les notations, l'hypothèse $f \in L^1(E, \mu)$ signifie

$$\sum_{x\in E} |f(x)|\mu(x) = \int_E |f(x)|d\mu(x) < \infty. \tag{40.137}$$

Démonstration. Nous prouvons le théorème dans le cas où E est fini. Si nous écrivons

$$f(X_k) = \sum_{y\in E} f(y)\mathbb{1}_{X_k=y}, \tag{40.138}$$

alors

$$\frac{1}{N} \sum_{k=1}^{N} f(X_k) = \sum_{y\in E} f(y) \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}_{X_k=y}. \tag{40.139}$$

Étant donné que E est fini nous pouvons permuter les sommes et prendre la limite $N \to \infty$:

$$\lim_{N\to\infty} \frac{1}{N} \sum_k f(X_k) = \sum_{y\in E} f(y) \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}_{X_k=y} = \sum_{y\in E} f(y)\pi(y). \tag{40.140}$$

□

40.6 Convergence vers l'équilibre

Nous voudrions savoir sous quelles conditions la variable aléatoire X_n converge en loi vers quelque chose lorsque $n \to \infty$. Une telle loi limite doit dépendre de la loi initiale[2] comme le montre l'exemple de la chaîne de Markov

$$1 \circlearrowleft A \xleftarrow{1/2} C \xrightarrow{1/2} B \circlearrowright 1 \tag{40.141}$$

Si $X_0 = C$, alors la loi limite est

$$\frac{1}{2}(\delta_A + \delta_B). \tag{40.142}$$

Si par contre $X_0 = B$, la loi limite est δ_B. Notons que la chaîne de Markov proposée ici est irréductible.

2. Lorsque la loi limite ne dépend pas de la loi initiale, nous disons que la chaîne de Markov est ergodique, nous y reviendrons.

Notons qu'il n'y a pas toujours de lois limite comme le montre l'exemple

$$A \underset{1}{\overset{1}{\rightleftarrows}} B \tag{40.143}$$

avec $X_0 = A$. La loi en est

$$X_k = \begin{cases} \delta_A & \text{si } k \text{ est pair} \\ \delta_B & \text{si } k \text{ est impair.} \end{cases} \tag{40.144}$$

Lemme 40.40.
Si nous avons une loi limite

$$P(X_n = x) \to l(x), \tag{40.145}$$

et que la chaîne est irréductible, alors nous avons $l = \pi$.

Démonstration. D'après la proposition 40.29 nous avons

$$\frac{1}{n} \sum_{k=1}^{n} P(X_k = x) \to \pi(x). \tag{40.146}$$

Par le lemme 12.75 sur la moyenne de Cesaro et l'hypothèse (40.145), nous avons aussi

$$\frac{1}{n} \sum_{k=1}^{n} P(X_k = x) \to l(x). \tag{40.147}$$

Du coup $\pi(x) = l(x)$. □

Lemme 40.41 ([430]).
Si π est une loi stationnaire et si x est un étant transient, alors $\pi(x) = 0$.

Ce lemme (qui peut être prouvé rigoureusement) est principalement dû au fait que la chaîne de Markov ne visite un état transitoire qu'un nombre fini de fois par la proposition 40.24(1).

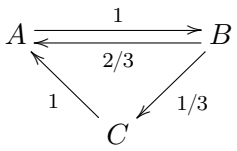
Définition 40.42.
*Un état $x \in E$ est **apériodique** si*

$$\text{pgcd}\{n \geqslant 1 \text{ tel que } p^n(x,x) > 0\} = 1. \tag{40.148}$$

Mettons que tous les n tels que $p^n(x,x) > 0$ ont 2 comme diviseur. L'état n'est alors pas apériodique, mais on voit que si $X_0 = x$, alors les états impairs ne peuvent pas être sur x. Cela est une forme de périodicité.

Si un état est apériodique, il existe p et q premiers entre eux tels que $p^p(x,x)$ et $p^q(x,x)$ sont non nuls. En particulier pour tout $n \in p\mathbb{N}+q\mathbb{N}$, $P(X_n = x) \neq 0$. Par conséquent la proposition 3.35 nous indique qu'à partir d'un certain moment tous les X_k pourraient être x.

L'état C de la chaîne de Markov suivante est apériodique :



(40.149)

En effet $p^3(C,C) \neq 0$ par le chemin $C \to A \to B \to C$ tandis que $p^5(C,C) \neq 0$ également par le chemin $C \to A \to B \to A \to B \to C$. Or $\text{pgcd}\{3,5\} = 1$.

Proposition 40.43 ([430]).
Soit (X_n), une chaîne de Markov irréductible. Un état x est apériodique si et seulement s'il existe N tel que

$$p^k(x,x) = P(X_k = x | X_0 = x) > 0 \tag{40.150}$$

pour tout $k \geqslant N$.

La proposition suivante va nous permettre de parler de **chaîne apériodique** .

Proposition 40.44.
Si une chaîne de Markov est irréductible, alors un état est apériodique si et seulement si tous les états sont apériodiques.

Démonstration. Soit x un état apériodique de la chaîne de Markov $(X_n)_{n\in\mathbb{N}}$. En vertu de la proposition 40.43 il existe N_x tel que $p^k(x,x) \neq 0$ pour tout $k \geqslant N_x$. Soit $y \in E$. Étant donné que la chaîne est irréductible, il existe r et s tels que$p^r(x,y) > 0$ et $p^s(y,x) > 0$. Nous avons

$$p^{k+r+s}(y,y) = P(X_{k+r+s} = y|X_0 = y) \geqslant p^s(x,y)P(X_k = x|X_0 = x)p^r(y,x). \tag{40.151}$$

Si k est assez grand, cette quantité est strictement positive. Donc il suffit de prendre $N_y = N_x+r+s$ pour savoir que y est également apériodique. □

Exemple 40.45
Quelle est la différence entre une chaîne irréductible et une chaîne apériodique ? Une chaîne est irréductible lorsque aucune sous-chaîne ne peut piéger le système. Pour toute paire d'états $x, y \in E$, il existe un n tel qu'il soit possible d'aller de x à y en n pas. Une chaîne est apériodique lorsqu'après un temps suffisamment long, *tous* les états soient possibles en même temps.

Un exemple de chaîne irréductible non apériodique :

$$A \underset{1}{\overset{1}{\rightleftarrows}} B \tag{40.152}$$

Cette chaîne est irréductible parce que le graphe est connexe, par contre il n'est pas apériodique parce que si $X_0 = A$ il n'est pas possible d'être dans l'état A après un nombre impair de pas.

Plus formellement, $p^n(A,A) = 1$ dès que n est pair ; le PGCD de la définition 40.42 n'est donc certainement pas 1. △

Si E est fini et si la chaîne de Markov est irréductible, alors en posant $N = \max_{x\in E} N(x)$, la matrice P^k a des éléments non nuls sur toute la diagonale pour tout $k > N$. Ces éléments diagonaux ne sont autre que les $p^k(x,x)$.

Théorème 40.46 (Convergence en loi des chaîne de Markov)**.**
Si (X_n) est

(1) irréductible,

(2) récurrente positive,

(3) apériodique,

alors X_n converge en loi vers l'unique probabilité invariante π vérifiant

$$\pi(x) = \sum_{u\in E} p(y,x)\pi(y) = \frac{1}{E\big(T(x)|X_0 = x\big)}. \tag{40.153}$$

Cette convergence est indépendante de la loi de X_0 et on a

$$P(X_n = x|X_0 = y) \to_{n\to\infty} \pi(x). \tag{40.154}$$

40.7 Processus de Galton-Watson

Nous considérons une maladie et notons Z_n le nombre de malades à l'instant n. Nous posons $Z_0 = 1$ et

$$Z_{n+1} = \begin{cases} 0 & \text{si } Z_n = 0 \\ \sum_{i=1}^{Z_n} \xi_i^{(n)} & \text{sinon} \end{cases} \tag{40.155}$$

où $\xi_i^{(n)}$ est le nombre de personnes contaminées par le malade i à l'instant n. Nous supposons que ces variables aléatoires sont indépendantes et identiquement distribuées et admettent un moment d'ordre 1.

L'équation de propagation 40.155 signifie que nous supposons qu'une personne malade à l'instant n n'est plus malade à l'instant $n+1$. Par ailleurs les hypothèses d'indépendance signifient qu'à chaque instant, le nombre de personnes contaminées par le malade i est indépendant du nombre de personnes contaminées par le malade j. De plus la façon dont la contamination se passe à l'instant n est indépendant de la façon dont la contamination se passe à l'instant m. Ces hypothèses sont raisonnables tant que le nombre de personnes non contaminées est grand. À partir du moment où presque tout le monde est malade, l'approximation de Galton-Watson ne fonctionne plus.

Nous notons ξ la loi parente des $\xi_i^{(n)}$. Ensuite nous considérons

$$G(s) = E(s^{\xi}) \tag{40.156a}$$
$$m = E(\xi) \tag{40.156b}$$
$$G_n(s) = E(s^{Z_n}). \tag{40.156c}$$

Par le théorème de transfert (proposition 38.60) avec $f(t) = s^t$. Ce que nous avons est

$$G_n(s) = E\big(f(Z_n)\big) = \int_{\mathbb{R}} s^x dP_{Z_n}(x) = \sum_{k=0}^{\infty} s^k P(Z_n = k) \tag{40.157}$$

où l'intégrale s'est transformée en somme parce que la loi de Z_n est discrète : dP_{Z_n} est une somme de masses de Dirac. En particulier nous avons

$$G_n(s) = \sum_{k=0}^{\infty} s^k P(Z_n = k) \tag{40.158a}$$
$$G(0) = P(Z_n = 0) \tag{40.158b}$$

et

$$\eta = \lim_{n\to\infty} P(Z_n = 0) = \lim_{n\to\infty} G_n(0). \tag{40.159}$$

D'où l'intérêt d'étudier G_n.

Lemme 40.47.
Pour tout $n \in \mathbb{N}^*$ et pour tout $s \in [0,1]$, nous avons

$$G_n(s) = \underbrace{G \circ G \circ \ldots \circ G(s)}_{n \text{ fois}}. \tag{40.160}$$

Démonstration. Pour $n = 1$, nous avons $Z_1 = \xi_1^{(1)}$ et donc

$$G_1(s) = E(s^{\xi}) = G(s), \tag{40.161}$$

comme il se doit.

Si $n \neq 1$ nous écrivons

$$G_n(s) = E(s^{Z_n}) \tag{40.162a}$$
$$= E\left(s^{\sum_{i=1}^{Z_{n-1}} \xi_i^{(n-1)}}\right) \tag{40.162b}$$
$$= E\left(\sum_{k=0}^{\infty} \mathbb{1}_{\{Z_{n-1}=k\}} s^{\sum_{i=1}^{k} \xi_i^{(n-1)}}\right). \tag{40.162c}$$

À ce niveau, nous voulons permuter la somme et l'espérance. Étant donné que le lemme est facile à vérifier pour $s = 1$, nous supposons $s < 1$. Du coup

$$s^{\sum_{i=1}^{k} \xi_i^{(n-1)}} < 1 \tag{40.163}$$

et ce qui se trouve dans l'espérance est majoré par

$$\sum_{k=0}^{\infty} \mathbb{1}_{Z_{n-1}=k} = 1. \tag{40.164}$$

La fonction constante 1 est intégrable sur Ω (ici nous utilisons à fond le fait que l'espace Ω soit un espace de probabilité) et nous pouvons utiliser le théorème de convergence dominée de Lebesgue 17.20 pour permuter la somme et l'intégrale. Nous continuons donc le calcul (40.162) :

$$G_n(s) = \sum_{k=0}^{\infty} E\left(\mathbb{1}_{\{Z_{n-1}=k\}} s^{\sum_{i=1}^{k}\xi_i^{(n-1)}}\right). \tag{40.165}$$

La tribu engendrée par la variable aléatoire $\mathbb{1}_{\{Z_{n-1}=k\}}$ est une fonction des variables aléatoires $\xi_i^{(m)}$ avec $m \leqslant n-2$ tandis que la variable aléatoire $s^{\sum_{i=1}^{k}\xi_i^{(n-1)}}$ est une fonction des variables aléatoires $\xi_i^{(n-1)}$. Par conséquent le lemme de regroupement 38.16 nous dit que ces variables aléatoires sont indépendantes, donc

$$G_n(s) = \sum_{k=0}^{\infty} \underbrace{E\big(\mathbb{1}_{\{Z_{n-1}=k\}}\big)}_{=P(Z_{n-1}=k)} E\big(s^{\sum_{i=1}^{k}\xi_i^{(n-1)}}\big). \tag{40.166}$$

Nous avons utilisé le fait que l'espérance d'une fonction indicatrice est la probabilité de l'événement.

En ce qui concerne la puissance de s, les événements ξ_i^{n-1} sont indépendants et suivent tous la même loi ξ, donc

$$s^{\sum_{i=1}^{k}\xi_i^{(n-1)}} = \prod_{i=1}^{k} s^{\xi_i^{(n-1)}} \tag{40.167}$$

et

$$E\big(\prod_{i=1}^{k} s^{\xi}\big) = E(s^{\xi})^k = G(s)^k. \tag{40.168}$$

En mettant tout bout à bout,

$$G_n(s) = \sum_{k=1}^{\infty} P(Z_{n-1}=k)G(s)^j = G_{n-1}\big(G(s)\big). \tag{40.169}$$

□

Théorème 40.48.
La probabilité d'extinction η est donnée par

$$\eta = P\left(\bigcup_{n\geqslant 1}(Z_n=0)\right) = \lim_{n\to\infty} P(Z_n=0). \tag{40.170}$$

Ce nombre est la plus petite solution positive de l'équation $G(s)=s$.

De plus la classification des cas est comme suit.

(1) Si $P(\xi=0)=0$ alors $\eta=0$.

(2) Si $P(\xi=0)\neq 0$ alors

(a) si $m\leqslant 1$ alors $\eta=1$,

(b) si $m>1$ alors $\eta\in\,]0,1[$.

Le cas $m<1$ est dit **sous-critique**, le cas $m=1$ est dit **critique**. Le cas $m>1$ est dit **sur-critique**.

Démonstration. Commençons par prouver que G est une fonction continue. En utilisant la théorème de transfert comme pour l'équation (40.157) nous trouvons que

$$G(s) = E(s^\xi) = \sum_{k=0}^{\infty} p_k s^k \tag{40.171}$$

où nous avons noté $p_k = P(\xi = k)$. Si $r < 1$, alors la suite $p_k r^k$ est bornée, donc le critère d'Abel (17.28) nous indique que la série (40.171) converge absolument et la théorie générale des séries entières conclut que la fonction G est en particulier dérivable terme à terme pour tout $s \in]-1, 1[$.

Le probabilité d'extinction est un point fixe de G En utilisant la continuité de G en 0 nous passons à la limite dans $G_{n+1}(0) = G\big(G_n(0)\big)$ et nous obtenons

$$\eta = G(\eta), \tag{40.172}$$

ce qui signifie que la probabilité d'extinction est un point fixe de G.

η est le plus petit point fixe de G Nous démontrons maintenant que η est plus précisément le plus petit point fixe de G sur $[0, 1]$. Nous allons effectuer cette partie en décomposant selon les valeurs de p_0 et de p_1.

Au vu de l'écriture (40.171), si $p_1 = 1$ alors $G(s) = s$ pour tout $s \in [0, 1]$. Mais dans ce cas nous savons par ailleurs que l'extinction est impossible. Zéro est bien la plus petite solution de $G(s) = s$.

Supposons maintenant que $p_1 < 1$ et $p_0 + p_1 = 1$. Alors $G(s) = p_0 + p_1 s$ et $s = 1$ est l'unique solution. Mais vu que nous savons que η est solution, c'est que $\eta = 1$ et l'extinction est certaine.

Nous passons au cas général : $p_0 + p_1 < 1$. D'abord nous remarquons que $s = 1$ est solution parce que

$$G(1) = p_0 + p_1 + \cdots = 1. \tag{40.173}$$

Remarquons aussi que dans ce cas $s = 0$ n'est plus solution.

La fonction G est strictement convexe sur $[0, 1]$ (parce que $G'' > 0$). Cela se voir en effectuant deux dérivations termes à termes (le rayon de convergence de la dérivée est le même que celui de la fonction). Cette stricte convexité entraine que l'équation $G(s) = s$ a au maximum une autre solution que $s = 1$. Nous nommons s_0 la plus petite solution dans $[0, 1]$. Étant donné que G est croissante on a

$$G(0) \leqslant G(s_0) = s_0. \tag{40.174}$$

En appliquant G à cette équation nous obtenons $G\big(G(s_0)\big) \leqslant G(s_0) = s_0$ et en appliquant n fois,

$$G_n(0) \leqslant s_0. \tag{40.175}$$

En passant à la limite, $\eta \leqslant s_0$ mais η étant solution, nous avons $\eta = s_0$. Nous avons donc prouvé que la probabilité d'extinction η est la plus petite solution de $G(s) = s$.

Classification des cas Nous devons encore discuter les cas. Si $P(\xi = 0) = 0$, alors $p_0 = 0$ et $G(0) = 0$, ce qui signifie que $s_0 = \eta = 0$ et l'extinction est impossible.

Nous passons au cas $p_0 \neq 0$. Si $p_0 + p_1 = 1$, alors $m = p_1 < 1$ et nous avions déjà vu que dans le cas $p_0 + p_1 = 1$, la probabilité d'extinction est $\eta = 1$.

Il nous reste à traiter le cas $p_0 + p_1 < 1$. Encore une fois, la courbe G est strictement convexe sur $[0, 1]$ et elle est en particulier plus grande que sa tangente en $s = 1$, c'est à dire

$$G(s) > G'(1)(s - 1) + G(1). \tag{40.176}$$

Nous savons que $G(1) = 1$. En ce qui concerne $G'(1)$, nous dérivons encore terme à termes :

$$G'(s) = \sum_{k=1}^{\infty} k p_k s^{k-1}, \tag{40.177}$$

donc

$$G'(1) = \sum_{k=1}^{\infty} kp_k = E(\xi) = m. \tag{40.178}$$

Ce que nous avons donc est

$$G(s) > 1 + m(s-1). \tag{40.179}$$

Nous nous particularisons au cas sous-critique ($m \leqslant 1$). En nous rappelant que $s - 1 < 0$,

$$G(s) > 1 + (s-1) = s, \tag{40.180}$$

donc $s = 1$ est la plus petite solution et effectivement nous avons déjà vu que $\eta = 1$ dans ce cas.

Si $m > 1$, alors on a

$$G(s) > 1 + m(s-1). \tag{40.181}$$

Mais dire $m > 1$ revient à dire $G'(1) > 1$ et donc dans un voisinage de $s = 1$ on a

$$\frac{G(s) - G(1)}{s-1} > 1, \tag{40.182}$$

ce qui implique que

$$G(s) < s - 1 + G(1) = s. \tag{40.183}$$

Nous avons donc $G(s) < s$ dans un voisinage de 1. Mais $G(0) - 0 = p_0 > 0$, donc la fonction $f(s) = G(s) - s$ est positive en 0 et négative proche de $s = 1$. Le théorème de la valeur intermédiaire nous indique alors qu'il existe un $s \in]0, 1[$ tel que $f(s) = 0$, c'est à dire tel que $G(s) = s$.

□

Chapitre 41

Martingales

41.1 Convergence de martingales

Définition 41.1.
*Si $\mathcal{A}$ est une tribu, une **filtration** de $\mathcal{A}$ est une suite croissante de sous-tribus $\mathcal{B}_i \subseteq \mathcal{B}_{i+1} \subseteq \mathcal{A}$.*

*Nous disons qu'une suite de variables aléatoires (X_n) est **adaptée** à une filtration $(\mathcal{F}_n)$ si X_i est $\mathcal{F}_i$-mesurable pour tout i.*

Ces définitions impliquent immédiatement que si (X_n) est adapté à $(\mathcal{F}_n)$ alors X_n est $\mathcal{F}_k$-mesurable pour $k \geqslant n$.

Définition 41.2.
*Une **martingale** adaptée à la filtration $(\mathcal{B}_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires M_n telle que*

(1) $M_n \in L^1(\Omega, \mathcal{A}, P)$,

(2) M_n est $\mathcal{B}_n$-mesurable,

(3) $E(M_{n+1}|\mathcal{B}_n) = M_n$.

*Le processus M_n est une **sur-martingale** si $E(M_{n+1}|\mathcal{B}_n) \leqslant M_n$ pour tout n, et c'est une **sous-martingale** si $E(M_n|\mathcal{B}_n) \geqslant M_n$.*

Exemple 41.3
Si $M \in L^1(\Omega, \mathcal{A}, P)$ et si $(\mathcal{B}_n)_{n\in\mathbb{N}}$ est une filtration, nous pouvons considérer la martingale $M_n = E(M|\mathcal{B}_n)$. △

Exemple 41.4
Soit $(X_i)_{i\geqslant 1}$ une suite de variables aléatoires indépendantes et centrées. On pose

$$S_n = X_1 + \cdots + X_n \tag{41.1}$$

et la filtration $\mathcal{B}_n = \sigma(X_1, \ldots, X_n)$. Pour montrer que cela est une martingale, nous commençons par remarquer que

$$E(X_{n+1}|\mathcal{B}_n) = E(X_{n+1}) = 0 \tag{41.2}$$

par indépendance des tribus $\mathcal{B}_n$ et $\sigma(X_{n+1})$. Ici c'est le lemme 38.38 qui joue.

Ensuite nous argumentons que $E(X_1 + \cdots + X_n|\mathcal{B}_n) = X_1 + \cdots + X_n$. En effet d'une part $X_1 + \cdots + X_n$ est $\mathcal{B}_n$-mesurable et évidemment la condition intégrale de l'espérance conditionnelle est satisfaite.

Plus généralement si X est une variable aléatoire et si $\sigma(X) \subset \mathcal{B}$ alors $E(X|\mathcal{B}) = X$. △

Lemme 41.5.
Soit (M_n) une martingales adaptée à la filtration $(\mathcal{F}_n)$ et $n \geqslant k$. Alors

$$E(M_n|\mathcal{F}_k) = M_k \tag{41.3a}$$
$$E(M_k|\mathcal{F}_n) = M_k. \tag{41.3b}$$

Démonstration. La seconde relation revient seulement à dire que M_k est $\mathcal{F}_n$-mesurable, ce qui est évident parce que $\mathcal{F}_k \subset \mathcal{F}_n$.

Nous prouvons la première par récurrence (à l'envers) sur k. D'abord si $k = n$, l'égalité $E(M_n|\mathcal{F}_n) = M_n$. Nous supposons maintenant que $E(M_n|\mathcal{F}_k) = M_k$, et nous prouvons que $E(M_n|\mathcal{F}_{k-1}) = M_{k-1}$. Si $B_{k-1} \in \mathcal{F}_{k-1}$, nous avons

$$\int_{B_{k-1}} M_{k-1} = \int_{B_{k-1}} M_k = \int_{B_{k-1}} M_n. \tag{41.4}$$

La première égalité est la définition d'une martingale, et la seconde est l'hypothèse de récurrence. □

Théorème 41.6 ([431, 432]).
Soit $(M_n)_{n\geqslant 0}$ une martingale bornée dans $L^2(\Omega)$, c'est à dire telle que

$$\alpha = \sup_{n\geqslant 0} E(M_n^2) < \infty. \tag{41.5}$$

Alors la suite M_n converge dans $L^2(\Omega)$.

Démonstration. Nous écrivons M_n en somme télescopique

$$M_n = M_0 + \sum_{k=1}^{n} \Delta_k \tag{41.6}$$

où $\Delta_k = M_k - M_{k-1}$. Nous commençons par monter que les incréments sont orthogonaux au sens où $E(\Delta_n\Delta_k) = 0$. Pour $n > k$, la variable aléatoire $E\big(\Delta_n\Delta_k|\mathcal{F}_{n-1}\big)$ est la variable aléatoire $\mathcal{F}_{n-1}$-mesurable telle que

$$\int_{B_{n-1}} E\big(\Delta_n\Delta_k|\mathcal{F}_{n-1}\big) = \int_{B_{n-1}} \Delta_n\Delta_k \tag{41.7}$$

pour tout $B_{n-1} \in \mathcal{F}_{n-1}$. En particulier avec $B_{n-1} = \Omega$ nous trouvons

$$E\Big(E\big(\Delta_n\Delta_k|\mathcal{F}_{n-1}\big)\Big) = E(\Delta_n\Delta_k) \tag{41.8}$$

par la définition de l'espérance (38.51). Par conséquent, en utilisant le lemme 41.5 nous avons [1]

$$E(\Delta_n\Delta_k) = E\Big(E(\Delta_n\Delta_k|\mathcal{F}_{n-1})\Big) = E\Big(\Delta_k E(\Delta_n|\mathcal{F}_{n-1})\Big) = 0 \tag{41.9}$$

parce que $E(\Delta_n|\mathcal{F}_{n-1}) = E(M_n|\mathcal{F}_{n-1}) - E(M_{n-1}|\mathcal{F}_{n-1}) = 0$.

Utilisant l'orthogonalité des incréments, nous avons

$$E(M_n^2) = E(M_0^2) + \sum_{k=1}^{n} E(\Delta_k^2). \tag{41.10}$$

En prenant le supremum (par rapport à n des deux côtés),

$$E(M_0^2) + \sum_{k=1}^{\infty} E(\Delta_k^2) = \alpha < \infty. \tag{41.11}$$

1. À ce niveau je crois qu'il y a une faute dans [432] qui conditionne par rapport à $\mathcal{F}_n$.

Cela prouve que la suite $\sum_{k=1}^{n}\Delta_k$ converge dans $L^2(\Omega)$. Nous en déduisons immédiatement que (M_n) est de Cauchy dans $L^2(\Omega)$ parce que si $k,l>n$, nous avons (en utilisant encore l'orthogonalité des incréments)

$$E\big(|M_k-M_l|^2\big)=\sum_{i=k+1}^{l}E(\Delta_i^2)\leqslant\sum_{i=k+1}^{\infty}E(\Delta_i^2),\tag{41.12}$$

qui tend vers zéro lorsque $n\to\infty$. □

Le théorème suivant complète la conclusion du théorème 41.6.

Théorème 41.7 ([432]).
Soit $(M_n)_{n\in\mathbb{N}}$ une martingale bornée dans L^2. Alors (M_n) converge dans $L^2(\Omega)$ et presque sûrement vers une même variable aléatoire M_∞ qui vérifie

$$M_n=E(M_\infty|\mathcal{F}_n).\tag{41.13}$$

Notons en particulier que la variable aléatoire M_∞ est presque sûrement finie parce qu'en vertu de (41.13) nous avons

$$\int_\Omega M_\infty=\int_\Omega M_n<\infty.\tag{41.14}$$

Exemple 41.8
Soient des variables aléatoires indépendantes $V_k\sim\mathscr{E}(2^n\lambda)$ et la variable aléatoire somme

$$S_n=\sum_{k=1}^{n}V_k.\tag{41.15}$$

Nous allons montrer que $S_n\xrightarrow{p.s.}X$ où X est une variable aléatoire presque sûrement finie. Nous posons

$$M_n=S_n-\sum_{k=1}^{n}\frac{1}{2^k\lambda}\tag{41.16}$$

Cela est une martingale adaptée à la filtration $\mathcal{F}_n=\sigma(V_1,\dots,V_n)$ en vertu de l'exemple 41.4. Nous montrons à présent qu'elle est bornée dans $L^2(\Omega)$ au sens où $\sum_{n\geqslant1}E(M_n^2)<\infty$. Nous avons

$$E(M_n^2)=E\left(\big[S_n-\sum_k\frac{1}{2^k\lambda}\big]^2\right)=E\left(\big[\sum_k(V_k-\frac{1}{2^k\lambda})\big]^2\right).\tag{41.17}$$

La variable aléatoire $V_k-1/2^k\lambda$ est une variable aléatoire centrée de variance $1/(2^k\lambda)^2$ (voir proposition 38.97). Étant donné que M_n est centrée, $\operatorname{Var}(M_n)=E(M_n^2)$ et nous avons

$$E(M_n^2)=\sum_{k=1}^{n}\operatorname{Var}\left(V_k-\frac{1}{2^k\lambda}\right)=\sum_{k=1}^{n}\frac{1}{(2^k\lambda)^2},\tag{41.18}$$

cette dernière somme étant bornée par $l=\sum_{k=1}^{\infty}\frac{1}{(2^k\lambda)^2}$, nous avons

$$E(M_n^2)\leqslant l\tag{41.19}$$

avec l indépendant de n. C'est pour cela que $(M_n)_{n\in\mathbb{N}}$ est une martingale bornée dans $L^2(\Omega)$. Par le théorème 41.7 nous avons $M_n\to M_\infty$ et en faisant $n\to\infty$ dans

$$S_n=M_n+\sum_{k=1}^{n}\frac{1}{2^k\lambda},\tag{41.20}$$

nous trouvons

$$S_n\to M_\infty+\sum_{k=1}^{\infty}\frac{1}{2^k\lambda}=M_\infty+\frac{1}{\lambda}\tag{41.21}$$

qui est presque sûrement finie. △

41.2 Temps d'arrêt et martingale terminée

Définition 41.9.
*Soit $(\Omega, \mathcal{F}_n, \mathcal{F}, P)$ un espace de probabilité filtré. Une application $T : \Omega \to \bar{\mathbb{N}}$ est un **temps d'arrêt** adapté à la filtration $(\mathcal{F}_n)$ si pour tout $n \in \mathbb{N}$ nous avons $\{T \leqslant n\} \in \mathcal{F}_n$.*

*Le temps d'arrêt T est **borné** s'il existe $k \in \mathbb{N}$ tel que $T(\omega) \leqslant k$ pour presque tout $\omega \in \Omega$.*

Lemme 41.10.
Si T est un temps d'arrêt presque sûrement fini, alors[2]
- $T \wedge n \xrightarrow{p.s.} T$,
- $\lim_{n\to\infty} E(T \wedge n) = E(T)$.

Démonstration. Vu que T est presque sûrement finie, il suffit de prouver que

$$(T \wedge n)(\omega) \to T(\omega) \tag{41.22}$$

pour tout ω tel que $T(\omega) = k$ pour tout $k \in \mathbb{N}$. Soient donc $\omega \in \Omega$ tel que $T(\omega) = k$ et $n > k$. Nous avons

$$(T \wedge n)(\omega) = T(\omega) \wedge n = k = T(\omega). \tag{41.23}$$

En ce qui concerne la seconde assertion, la suite de variables aléatoires $X_x = T \wedge n$ est croissante et positive, donc le théorème de la convergence monotone 16.116 montre que

$$\lim_{n\to\infty} E(T \wedge n) = E(T). \tag{41.24}$$

□

Remarque 41.11.
Notons la différence subtile entre $S_T(\omega)$ et $(S_T)(\omega)$. La première est la variable aléatoire

$$\omega' \mapsto S_{T(\omega')}(\omega) \tag{41.25}$$

et la seconde est le nombre $S_{T(\omega)}(\omega)$.

Théorème 41.12 (Théorème d'arrêt borné[432])**.**
Soit (X_n) une sur-martingale et $S \leqslant T$, deux temps d'arrêts bornés. Alors

(1) les variables aléatoires X_S et X_T sont intégrables,

(2) $E(X_T|\sigma(S)) \leqslant X_S$ presque sûrement.

Si par contre (X_n) est une martingale alors X_S et X_T sont bornées, et

$$E(X_T|\sigma(S)) = X_S. \tag{41.26}$$

Remarque 41.13.
Un cas particulier intéressant de ce théorème 41.12 est le cas $S = 0$ qui est un temps d'arrêt vérifiant $\mathcal{F}_0 = \{\Omega, \varnothing\}$. Si X est n'importe quelle variable aléatoire, la tribu engendrée $\sigma(X)$ est toujours indépendante de la tribu $\{\Omega, \varnothing\}$, donc le résultat $E(X_T|\mathcal{F}_S) = X_S$ donne

$$E(X_T) = X_0. \tag{41.27}$$

Théorème 41.14 (Premier théorème d'arrêt de Doob[433])**.**
Soient (X_n) une martingale et T un temps d'arrêt ; tous deux pour la filtration $(\mathcal{F}_n)$. Nous supposons qu'une des trois propriétés suivantes soit vérifiée :

(1) T est presque sûrement bornée.

2. Dans [4], dans le problème de la ruine du joueur, la seconde assertion est avec une limite sup et non avec une limite normale.

(2) $E(T) < \infty$ *et il existe une constante* c *telle que*

$$E\big(|X_{n+1} - X_n| \,|\mathcal{F}_n\big) \leqslant c \tag{41.28}$$

sur l'événement $\{T \geqslant n\}$.

(3) Il existe une constante c *telle que* $|X_{T\wedge n}| \leqslant c$ *presque sûrement*[3].

Alors X_T *est une variable aléatoire presque sûrement bien définie nous avons*

$$E(T_T) = E(X_0). \tag{41.29}$$

Si (X_n) *est une sur-martingale, alors la conclusion est* $E(X_T) \leqslant E(X_0)$ *et si* (X_n) *est une sous-martingale, la conclusion est* $E(X_T) \geqslant E(X_0)$.

Remarque 41.15.
Sous l'hypothèse (3), il est possible d'avoir $T = \infty$ sur un ensemble de mesure non nulle. Sur cet ensemble, la variable aléatoire X_T doit être définie de façon plus fine.

Problèmes et choses à faire

D'après la page de discussion de l'article sur Wikipédia, il semblerait que la seconde condition soit mal énoncée. Je n'ai pas vérifié.

Définition 41.16.
Nous disons que la martingale $(M_n)_{n\geqslant 1}$ *est* ***terminée*** *s'il existe* $M \in L^1(\Omega, \mathcal{A}, P)$ *telle que* $E(M|\mathcal{A}_n) = M$ *pour tout* $n > 1$.

Définition 41.17.
Un ensemble $H \subset L^1(\Omega, \mu)$ *est* ***équi-intégrable*** *si*

$$\lim_{a\to\infty} \left(\sup_{f\in H} \int_{|f|>a} |f(x)| d\mu(x) \right) = 0. \tag{41.30}$$

Notons dans cette définition que vu que $f \in L^1$ nous avons toujours

$$\lim_{a\to\infty} \int_{|f|>a} |f(x)| d\mu(x) = 0. \tag{41.31}$$

L'équi-intégrabilité donne une sorte d'uniformité en f de cette limite.

Théorème 41.18.
Si (M_n) *est une martingale, nous avons équivalence entre*

(1) (M_n) *converge dans* L^1 ;

(2) (M_n) *est terminée ;*

(3) l'ensemble $\{M_n\}_{n\geqslant 1}$ *est équi-intégrable.*

Attention : en vertu de la proposition 28.15 et surtout de l'exemple 28.16, la convergence L^1 n'implique pas la convergence presque partout.

Théorème 41.19 (Théorème de Doob[195])**.**
À propos de convergence de martingales.

(1) Toute martingale terminée converge presque sûrement et pour la norme L^1.

(2) Toute martingale bornée dans L^2 *converge presque sûrement et pour la norme* L^2.

Proposition 41.20 ([434])**.**
Soient (M_n) *une martingale et* T *un temps d'arrêt (pour la même filtration* $(\mathcal{B}_n)$*). Alors le processus* $V_n = M_{n\wedge T}$ *est une martingale.*

3. Il est d'usage assez classique de noter $a \wedge b$ le minimum de a et b.

Démonstration. Nous décomposons V_n de la façon suivante :

$$V_n = M_{n\wedge T} = M_n \mathbb{1}_{T\geqslant n} + M_T \mathbb{1}_{T<n} = M_n \mathbb{1}_{T\geqslant n} + \sum_{k<n} M_k \mathbb{1}_{T=k}. \tag{41.32}$$

Nous avons, grâce au lemme 16.3,

$$\{T \geqslant n\} = \complement\{T < n\} = \complement\{T \leqslant n-1\} \in \mathcal{B}_{n-1} \tag{41.33}$$

et, si $k \leqslant n$,

$$\{T = k\} = \underbrace{\{T \leqslant k\}}_{\in\mathcal{B}_k} \setminus \underbrace{\{T \leqslant k-1\}}_{\in\mathcal{B}_{k-1}} \in \mathcal{B}_k \subset \mathcal{B}_n. \tag{41.34}$$

La forme (41.32) donne donc manifestement la $\mathcal{B}_n$-mesurabilité de V_n.

En ce qui concerne l'espérance nous devons calculer

$$E(V_{n+1}|\mathcal{B}_n) = E(M_{n+1}\mathbb{1}_{T\geqslant n+1}|\mathcal{B}_n) + \sum_{k<n+1} E(M_k \mathbb{1}_{T=k}|\mathcal{B}_n) \tag{41.35}$$

où nous avons utilisé la proposition 38.25. Étant donné que $\mathbb{1}_{T\geqslant n+1}$ et $\mathbb{1}_{T=k}$ sont des variables aléatoires $\mathcal{B}_n$-mesurables nous pouvons utiliser la proposition 38.41 pour les sortir :

$$E(V_{n+1}|\mathcal{B}_n) = \mathbb{1}_{T\geqslant n+1} M_n + \sum_{k\leqslant n} \mathbb{1}_{T=k} M_k = M_{T\wedge n} = V_n. \tag{41.36}$$

Pour cette ligne, nous avons aussi utilisé les égalités suivantes :

— $E(M_{n+1}|\mathcal{B}_n) = M_n$ parce que (M_n) est une martingale

— $E(M_k|\mathcal{B}_n) = M_k$ parce que M_k est $\mathcal{B}_n$-mesurable.

□

Définition 41.21.
*Si (X_n) est un processus adapté à la filtration $(\mathcal{F}_n)$ et si T est un temps d'arrêt $\mathcal{F}_n$-mesurable alors le **processus arrêté** à l'instant T est le processus $Y_n = X_{n\wedge T}$.*

Nous avons déjà vu par la proposition 41.20 que si (X_n) est une martingale alors son processus arrêté est encore une martingale.

41.3 Décomposition de martingales

Définition 41.22 (Processus croissant prévisible[432])**.**
*Un processus X_n adapté à la filtration $\mathcal{F}_n$ est un processus **croissant prévisible** si*

(1) $A_0 = 0$

(2) $A_n \leqslant A_{n+1}$; c'est cette condition qui correspond à « croissant »,

(3) A_{n+1} est $\mathcal{F}_n$-mesurable ; c'est cette condition qui correspond à « prévisible ».

Proposition 41.23 (Décomposition de Doob pour une sous-martingale[432])**.**
Toute sous-martingale (X_n) s'écrit de façon unique sous la forme

$$X_n = M_n + A_n \tag{41.37}$$

où (M_n) est une martingale et (A_n) est un processus croissant prévisible.

Démonstration. Nous considérons le processus

$$\begin{cases} A_0 = 0 & \text{(41.38a)} \\ A_{n+1} = A_n + E(X_{n+1} - X_n|\mathcal{F}_n). & \text{(41.38b)} \end{cases}$$

Nous vérifions que cela est un processus croissant prévisible. D'abord

$$E(X_{n+1} - X_n|\mathcal{F}_n) = E(X_{n+1}|\mathcal{F}_n) - E(X_n|\mathcal{F}_n). \tag{41.39}$$

Le second terme est égal à X_n parce que cette variable aléatoire est $\mathcal{F}_n$-mesurable tandis que (X_n) étant une sous-martingale nous avons $E(X_{n+1}|\mathcal{F}_n) \geqslant X_n$. Nous avons donc bien $A_{n+1} \geqslant A_n$ et le processus (A_n) est croissant.

En ce qui concerne la prévisibilité nous devons prouver que A_{n+1} est $\mathcal{F}_n$-mesurable. D'une part A_n est $\mathcal{F}_n$-mesurable et d'autre part par définition de l'espérance conditionnelle, la variable aléatoire $E(X_{n+1} - X_n|\mathcal{F}_n)$ est également $\mathcal{F}_n$-mesurable.

Nous posons alors $M_n = X_n - A_n$ et nous devons prouver que cela est une martingale. Nous avons

$$E(M_{n+1} - M_n|\mathcal{F}_n) = E(X_{n+1} - X_n|\mathcal{F}_n) - E(A_{n+1} - A_n|\mathcal{F}_n). \tag{41.40}$$

Le second terme vaut

$$E(A_{n+1} - A_n|\mathcal{F}_n) = E\Big(E(X_{n+1} - X_n|\mathcal{F}_n)|\mathcal{F}_n\Big) = E(X_{n+1} - X_n|\mathcal{F}_n) \tag{41.41}$$

par la proposition 38.36. Le processus (M_n) est donc une martingale. La preuve de l'existence d'une décomposition (41.37) est achevée.

Nous passons maintenant à l'unicité en posant $X_n = M_n + A_n = M'_n + A'_n$. Nous avons $A_0 = A'_0 = 0$ et $A'_n = X_n - M'_n$, donc

$$A'_{n+1} - A'_n = X_{n+1} - X_n + M'_{n+1} - M'_n = X_{n+1} - X_n - (M'_{n+1} - M'_n). \tag{41.42}$$

Nous appliquons $E(.|\mathcal{F}_n)$ des deux côtés de cette égalité :

$$\underbrace{E(A'_{n+1} - A'_n|\mathcal{F}_n)}_{=A'_{n+1}-A'_n} = E(X_{n+1} - X_n|\mathcal{F}_n) - \underbrace{E(M'_{n+1} - M'_n|\mathcal{F}_n)}_{=0}. \tag{41.43}$$

Nous avons utilisé le que que (M_n) étant une martingale, $E(M_{n+1} - M_n\mathcal{F}_n) = 0$, et idem avec (M'_n). Donc

$$A'_{n+1} - A'_n = E(X_{n+1} - X_n|\mathcal{F}_n) = E(M_{n+1} - M_n|\mathcal{F}_n) + E(A_{n+1} - A_n|\mathcal{F}_n) = A_{n+1} - A_n. \tag{41.44}$$

Nous avons donc montré que $A_{n+1} - A_n = A'_{n+1} - A'_n$ et donc que $A_n = A'_n$ pour tout n. Nous en déduisons immédiatement que $M_n = M'_n$ pour tout n et l'unicité de la décomposition. □

Lemme 41.24.
Si (X_n) est une martingale de carré intégrable adaptée à la filtration $(\mathcal{F}_n)$ alors

(1) Le processus (X_n^2) est une sous-martingale.

(2) Si $X_n^2 = M_n + A_n$ est la décomposition de Doob, alors

$$A_n = \sum_{i=1}^{n} \Big(E(X_i^2|\mathcal{A}_{i-1}) - X_{i-1}^2\Big) = \sum_{i=1}^{n} E\Big((X_i - X_{i-1})^2|\mathcal{A}_{i-1}\Big). \tag{41.45}$$

Démonstration. Pour la première assertion, nous utilisons l'inégalité de Jensen 38.52 :

$$E(X_n^2|\mathcal{F}_{n-1}) \geqslant \big(E(X_n|\mathcal{F}_{n-1})\big)^2 = X_n^2 \tag{41.46}$$

parce que $E(X_n|\mathcal{F}_{n-1}) = X_n$ du fait que (X_n) soit une martingale.

En ce qui concerne la seconde assertion, nous nous souvenons que le processus prévisible de la décomposition de Doob d'une sous-martingale est donné par la récurrence (41.38) que nous recopions ici :

$$\begin{cases} A_0 = 0 & \text{(41.47a)} \\ A_{n+1} = A_n + E(X_{n+1}^2 - X_n^2|\mathcal{F}_n) & \text{(41.47b)} \end{cases}$$

Vu que X_n^2 est $\mathcal{F}_n$-mesurable, il peut sortir de l'espérance :

$$A_{n+1} = A_n + E(X_{n+1}^2|\mathcal{F}_n) - X_n^2 \tag{41.48}$$

et donc

$$A_n = \sum_{i=1}^{n} \Big(E(X_i^2 | \mathcal{F}_{i-1}) - X_{i-1}^2 \Big). \tag{41.49}$$

Pour obtenir la dernière partie de (41.45) nous travaillons un peu :

$$\begin{aligned} E\big((X_i - X_{i-1})^2 | \mathcal{F}_{i-1}\big) &= E\big(X_i^2 + X_{i-1}^2 - 2X_iX_{i-1} | \mathcal{F}_{i-1}\big) && (41.50a)\\ &= E(X_i^2|\mathcal{F}_{i-1}) + X_{i-1}^2 - 2E(X_iX_{i-1}|\mathcal{F}_{i-1}) && (41.50b)\\ &= E(X_i^2|\mathcal{F}_{i-1}) + X_{i-1}^2 - 2X_{i-1}E(X_i|\mathcal{F}_{i-1}) && (41.50c)\\ &= E(X_i^2|\mathcal{F}_{i-1}) + X_{i-1}^2 - 2X_{i-1}X_i && (41.50d) \end{aligned}$$

où nous avons utilisé la proposition 38.41 pour obtenir (41.50c). □

41.4 Problème de la ruine du joueur

Nous considérons un joueur compulsif qui joue à un jeu très simple[4] : il joue à pile ou face contre la banque avec une pièce truquée. Si pile sort, la banque donne 1 au joueur et si c'est face, c'est le joueur qui donne 1 à la banque. Nous nommons a la fortune initiale du joueur, b celle de la banque et p la probabilité d'obtenir pile.

Nous supposons que le jeu se poursuit jusqu'à la ruine du joueur ou de la banque. La modélisation est comme suit : nous considérons (Y_n) une suite de variables aléatoires indépendantes et identiquement distribuées de loi

$$Y_n \sim p\delta_1 + (1-p)\delta_{-1}. \tag{41.51}$$

C'est le résultat financier pour le joueur du n^{e} lancé. La fortune du joueur au bout de n lancés est la variable aléatoire

$$S_n = a + \sum_{j=1}^{n} Y_j. \tag{41.52}$$

Nous notons $Y_0 = a$.

Nous considérons la filtration

$$\mathcal{A}_n = \sigma\big(S_i \text{ tel que } 0 \leqslant i \leqslant n\big) = \sigma\big(Y_i \text{ tel que } 0 \leqslant i \leqslant n\big), \tag{41.53}$$

et le temps d'arrêt du jeu :

$$T = \inf\{n \geqslant 1 \text{ tel que } S_n \in \{0, a+b\}\}; \tag{41.54}$$

c'est le temps qu'il faut pour que tout l'argent appartienne soit au joueur soit à la banque.

Nous voulons étudier les paramètres suivants :

(1) $\rho = P(S_T = a+b)$, c'est à dire la probabilité que ce soit le joueur qui gagne contre la banque.

(2) $P(T < \infty)$, c'est à dire la probabilité que le jeu se finisse.

(3) $E(T)$, la durée moyenne du jeu.

Lemme 41.25.
Le processus S_n du problème de la ruine du joueur est vérifie

$$E(S_n|\mathcal{A}_{n-1}) = S_{n-1} + p - q. \tag{41.55}$$

De plus le processus S_n est

(1) une martingale si $p = q = \frac{1}{2}$,

(2) une sous-martingale si $p > q$.

4. Le gros des choses dites à propos de la ruine du joueur provient de [4].

Démonstration. Pour $n \geqslant 1$ nous avons

$$E(S_n|\mathcal{A}_{n-1}) = a + \sum_{j=1}^{n} E(Y_j|\mathcal{A}_{n-1}) = a + \sum_{j=1}^{n-1} E(Y_j|\mathcal{A}_{n-1}) + E(Y_n|\mathcal{A}_{n-1}). \tag{41.56}$$

Si $j \leqslant n-1$ alors $Y_j \in m(\mathcal{A}_{n-1})$. Mais nous savons que si X est $\mathcal{F}$-mesurable, alors $E(X|\mathcal{F}) = X$ (c'est la définition de l'espérance conditionnelle), donc $\sum_{j=1}^{n-1} E(Y_j|\mathcal{A}_{n-1}) = \sum_{j=1}^{n-1} Y_j$.

En ce qui concerne le terme $j = n$ nous utilisons le fait que $\sigma(Y_n)$ soit une tribu indépendante de $\mathcal{A}_{n-1}$; nous avons donc au final pour tout j que $E(Y_j|\mathcal{A}_{n-1}E(Y_j) = p - q$. Nous avons donc

$$E(S_n|\mathcal{A}_{n-1}) = S_{n-1} + p - q. \tag{41.57}$$

Si $p = q = \frac{1}{2}$ alors c'est une martingale, et si $p > q$ c'est une sous-martingale. □

Lemme 41.26.
La variable aléatoire T est un temps d'arrêt.

Démonstration. Par définition $T = \inf\{n \geqslant 1 \text{ tel que } S_n \in \{0, a+b\}\}$. Vu que les variables aléatoires S_i avec $i \leqslant n$ sont $\mathcal{F}_n$-mesurables, les ensembles $\big\{S_k \notin \{0, a+b\}\big\}$ avec $k \leqslant n$ sont $\mathcal{F}_n$-mesurables. Donc les ensembles

$$\{T = n\} = \bigcap_{k \leqslant n} \big\{S_k \notin \{0, a+b\}\big\} \cap \big\{S_n \in \{0, a+b\}\big\} \tag{41.58}$$

sont $\mathcal{F}_n$-mesurables. Nous en concluons que l'ensemble $\{T \leqslant n\}$ est également mesurable. □

41.4.1 Le cas où la pièce est truquée

Nous supposons être dans le cas $p > q$.

41.4.1.1 Introduction d'une martingale

Considérons le processus

$$\begin{cases} A_0 = 0 & \text{(41.59a)} \\ A_n = A_{n-1} + E(S_n - S_{n-1}|\mathcal{A}_{n-1}). & \text{(41.59b)} \end{cases}$$

Vu que $E(S_n|\mathcal{A}_{n-1}) = S_{n-1} + p - q$ (lemme 41.25) et que $E(S_{n-1}|\mathcal{A}_{n-1}) = S_{n-1}$ (parce que S_{n-1} est dans la tribu de $\mathcal{A}_{n-1}$), nous avons $A_n = A_{n-1} + (p-q)$ et donc

$$A_n = n(p-q). \tag{41.60}$$

Ce processus (A_n) est croissant et prévisible. Nous introduisons le processus

$$M_n = S_n - A_n \tag{41.61}$$

et nous montrons que c'est une martingale [5]. Nous conditionnons la définition (41.61) par rapport à $\mathcal{A}_{n-1}$:

$$\begin{aligned} E(M_n\mathcal{A}_{n-1}) &= E(S_n|\mathcal{A}_{n-1}) - \underbrace{E(A_n|\mathcal{A}_{n-1})}_{=A_n} && \text{(41.62a)} \\ &= A_n - A_{n-1} + E(S_{n-1}|\mathcal{A}_{n-1}) - A_n && \text{(41.62b)} \\ &= E(S_{n-1}|\mathcal{A}_{n-1}) - A_{n-1}. && \text{(41.62c)} \end{aligned}$$

Mais S_{n-1} est $\mathcal{A}_{n-1}$-mesurable, donc $E(S_{n-1}|\mathcal{A}_{n-1}) = S_{n-1}$ et

$$E(M_n|\mathcal{A}_{n-1}) = S_{n-1} - A_{n-1} = M_{n-1}, \tag{41.63}$$

ce qui signifie que (M_n) est une martingale.

5. Ceci est un peu le contraire de la décomposition de Doob.

41.4.1.2 Finitude du temps d'arrêt

Nous montrons maintenant, en étudiant $M_{T\wedge n}$ que T est intégrable et nous prouvons que $P(T=\infty)=0$.

Proposition 41.27.
La variable aléatoire T est intégrable, et $P(T=\infty)=0$, c'est à dire que le jeu se termine presque certainement après un temps fini.

Démonstration. Nous rappelons que le lemme 41.26 nous indique que T est un temps d'arrêt. Le temps d'arrêt $T\wedge n$ est borné (par n évidemment) et nous pouvons donc lui appliquer le théorème d'arrêt 41.14 pour dire que

$$E(M_{T\wedge n})=E(M_0). \tag{41.64}$$

Le membre de droite est simple parce que $M_0=S_0-A_0=S_0=a$ parce que c'est l'argent de départ du joueur. Pour l'autre :

$$E(M_{T\wedge n})=E(S_{T\wedge n})-E(A_{T\wedge n}). \tag{41.65}$$

D'une part, $E(A_{T\wedge n})=E\big((T\wedge n)(p-q)\big)$ et d'autre part, $E(S_{T\wedge n})\leqslant a+b$ parce que S_T vaut zéro ou $a+b$ (avec des probabilités encore inconnues[6]). En combinant avec ce qui était dit juste au dessus et remarquant que $(p-q)E(T\wedge n)\geqslant 0$ nous pouvons écrire

$$0\leqslant(p-q)E(T\wedge n)\leqslant b. \tag{41.66}$$

La suite de variables aléatoires $T\wedge n$ est donc croissante, positive et intégrable[7] et donc nous avons du travail pour le théorème de la convergence monotone 16.116. La variable aléatoire T est alors mesurable et

$$\lim_{n\to\infty}E(T\wedge n)=E(T). \tag{41.67}$$

Notons que nous n'avons pas encore prouvé que $E(T)<\infty$, mais en passant à la limite dans (41.66) nous écrivons

$$0\leqslant(p-q)E(T)\leqslant b. \tag{41.68}$$

Maintenant nous avons prouvé que T est intégrable et même L^1. Par conséquent

$$P(T=\infty)=0. \tag{41.69}$$

Le jeu se termine donc presque certainement après un temps fini. □

41.4.1.3 Temps moyen de jeu

Le lemme 41.10 nous indique que $S_{T\wedge n}\xrightarrow{p.s.}S_T$.

Nous avons les bornes $0\leqslant S_{T\wedge n}\leqslant a+b$ et comme $a+b$ est intégrable, $S_{T\wedge n}$ l'est aussi et nous pouvons parler de $E(S_{T\wedge n})$. Repartons de (41.65) :

$$a=E(M_0)=E(M_{T\wedge n})=E(S_{T\wedge n})-E(A_{T\wedge n})=E(S_{T\wedge n})-(p-q)E(T\wedge n). \tag{41.70}$$

La variable aléatoire $S_{T\wedge n}$ est majorée par $a+b$ indépendamment de n ; donc le théorème de la convergence dominée 16.140 donne $\lim_{n\to\infty}E(S_{T\wedge n})=E(S_T)$. En ce qui concerne le second terme, la convergence dominée ne fonctionne pas parce que $T\wedge n$ n'est pas a priori majoré par quelque chose d'indépendant de n, mais le théorème de la convergence monotone donne $\lim_{n\to\infty}E(T\wedge n)=E(T)$. Au final en passant à la limite dans (41.70) nous avons

$$a=E(S_T)-(p-q)E(T). \tag{41.71}$$

6. Mais on y travaille.

7. Je rappelle que les constantes sont des fonctions intégrables sur Ω. Oui, je sais, quand on est habitué à faire de l'analyse sur $\mathbb{R}^n$ c'est un truc qu'on perd toujours un peu de vue.

Étant donné que $T > 0$ et $p - q > 0$ nous pouvons récrire cela sous la forme

$$0 \leqslant (p-q)E(T) = E(S_T) - a. \tag{41.72}$$

Par définition de T nous avons aussi

$$E(S_T) = (a+b)P(S_T = a+b) + 0 \cdot P(S_T = 0) = \rho(a+b). \tag{41.73}$$

Nous déduisons

$$E(T) = \frac{(a+b)\rho - a}{p-q}. \tag{41.74}$$

Ne crions pas victoire trop vite : nous n'avons pas encore d'expression de $\rho = P(S_T = a+b)$. Le temps moyen de jeu n'est donc pas encore tout à fait connu.

41.4.1.4 Probabilité de victoire du joueur

Nous avons besoin d'exprimer ρ en termes de a, b et p. Pour cela nous introduisons la variable aléatoire [8]

$$U_n = \left(\frac{p}{q}\right)^{S_n}. \tag{41.75}$$

Nous commençons par prouver que c'est une martingale en calculant

$$E(U_n|\mathcal{A}_{n-1}) = E\left(\left(\frac{q}{p}\right)^{S_n-1}\left(\frac{q}{p}\right)^{Y_n}|\mathcal{A}_{n-1}\right) \tag{41.76}$$

Nous utilisons la proposition 38.41. Dans notre cas, S_{n-1} et Y_n sont des variables aléatoires $\mathcal{A}_n$-mesurables ; la variable aléatoire Y_n est même $\mathcal{A}_{n-1}$-mesurable et sort donc du conditionnement ; nous avons donc

$$E(U_n|\mathcal{A}_{n-1}) = \left(\frac{q}{p}\right)^{S_{n-1}} E\left(\left(\frac{q}{p}\right)^{Y_n}\right) \tag{41.77}$$

Nous allons utiliser le théorème de transfert 38.60 :

$$E(s^{Y_n}) = \int_\Omega s^{Y_n(\omega)} DP(\omega) = \int_{Y_n=1} s dP(\omega) + \int_{Y_n=-1} \frac{1}{s} dP(\omega). \tag{41.78}$$

Mais nous savons que $P(Y_n = 1) = p$ et $P(Y_n = -1) = 1 - p = q$, donc

$$E(s^{Y_n}) = ps + \frac{1-p}{s} \tag{41.79}$$

et

$$E\left(\left(\frac{p}{q}\right)^{Y_n}\right) = p + q = 1. \tag{41.80}$$

Donc

$$E(U_n|\mathcal{A}_{n-1}) = \left(\frac{q}{p}\right)^{S_n-1} = U_{n-1}, \tag{41.81}$$

ce qui prouve que (U_n) est une martingale.

Par définition nous avons toujours $S_n \geqslant 0$ tant que $n \leqslant T$ [9], donc $U_{T\wedge n} \in [0,1]$. Il est donc évident que si $a \geqslant 1$ nous avons

$$\int_{|U_{T\wedge n}|>a} |U_{T\wedge n}| dP = 0 \tag{41.82}$$

8. Nous dirons un mot sur ce choix dans le « petit complément » plus bas
9. Pour $n > T$ le jeu est terminé, donc on ne se pose pas la question.

parce que le domaine d'intégration est vide. Donc les variables aléatoires $V_n = U_{T\wedge n}$ sont équi-intégrables [10] et le théorème 41.18 montre que la martingale (V_n) est terminée ; par ricochet [11] le théorème de Doob 41.19 montre qu'il existe une variable aléatoire X telle que $V_n \xrightarrow{p.s.} X$. Nous allons prouver que $X = U_T$ presque partout. Nous savions déjà (voir l'équation (41.22) et ses alentours) que

$$S_{n\wedge T} \xrightarrow{p.s.} S_T. \tag{41.83}$$

Nous avons alors (au sens du presque sûrement) :

$$\lim_{n\to\infty} V_n = \lim_{n\to\infty} U_{T\wedge n} = \lim_{n\to\infty} \left(\frac{q}{p}\right)^{S_{T\wedge n}} = \left(\frac{q}{p}\right)^{S_T} = U_T. \tag{41.84}$$

Donc par unicité de la limite presque partout nous avons $X = U_T$ presque partout. Par le théorème de transfert 38.60 nous évaluons

$$E(U_T) = \left(\frac{q}{p}\right)^0 P(S_T = 0) + \left(\frac{q}{p}\right)^{a+b} P(S_T = a+b) = (1-\rho) + \left(\frac{q}{p}\right)^{a+b} \rho. \tag{41.85}$$

La remarque 41.13 nous permet de dire que

$$E(U_{T\wedge n}) = U_0. \tag{41.86}$$

Mais par définition

$$U_0 = \left(\frac{q}{p}\right)^{S_0} = \left(\frac{q}{p}\right)^{a}, \tag{41.87}$$

donc nous avons

$$E(U_{T\wedge n}) = \left(\frac{q}{p}\right)^{a}. \tag{41.88}$$

Nous voudrions passer à la limite $n \to \infty$ dans cette équation. Pour permuter la limite et l'espérance, il faut utiliser le théorème de la convergence dominée 16.140. Vu que nous avons choisi $q > p$, nous avons $q/p > 1$ et donc $U_{T\wedge n} \leqslant (q/p)^{a+b}$, ce qui montre que la fonction $\omega \mapsto (U_{T\wedge n})(\omega)$ est majorée par une constante (qui est une fonction intégrable). Nous pouvons donc permuter la limite et l'espérance :

$$\lim_{n\to\infty} E(U_{T\wedge n}) = E\Big(\lim_{n\to\infty} U_{T\wedge n}\Big). \tag{41.89}$$

Mais nous avion déjà montré que $U_{T\wedge n} \xrightarrow{p.s.} U_T$. Donc

$$E(U_T) = \left(\frac{q}{p}\right)^{a}. \tag{41.90}$$

En égalisant avec l'expression (41.85) de $E(U_T)$ nous trouvons

$$\rho = \frac{\left(\frac{q}{p}\right)^a - 1}{\left(\frac{q}{p}\right)^{a+b} - 1} \tag{41.91}$$

et ensuite nous trouvons $E(T)$ en remettant ce ρ dans l'expression (41.74) donnée plus haut.

41.4.2 Le cas où la pièce est non truquée

Maintenant $p = q = 1/2$.

10. Définition 41.17.
11. Nous rappellons que la convergence L^1 n'implique pas la convergence presque partout.

41.4.2.1 Probabilité de gagner

Le lemme 41.25 nous indique alors que (S_n) est une martingale et le lemme 41.24 nous permet de dire que (S_n^2) est alors une sous-martingale. Le processus croissant prévisible de (S_n^2) est donné par (41.38) qui en adaptant les notations est

$$\begin{cases} B_0 = 0 & \text{(41.92a)} \\ B_n = B_{n-1} + E\Big((S_n - S_{n-1})^2 | \mathcal{A}_{n-1}\Big). & \text{(41.92b)} \end{cases}$$

Nous avons toujours $S_n - S_{n-1} = \pm 1$ parce que soit le joueur gagne soit le joueur perd, mais de toutes façons sa fortune varie de 1 à chaque étape du jeu. Donc (41.92b) nous donne $B_n = B_{n-1} + 1$ et

$$B_n = n. \tag{41.93}$$

Cela nous dit que la variable aléatoire

$$S_n^2 - B_n = S_n^2 - n \tag{41.94}$$

est une martingale (une sur-martingale moins son processus prévisible croissant). Nous lui appliquons le théorème d'arrêt 41.12 avec les temps d'arrêt 0 et $T \wedge n$:

$$E(S_{T\wedge n}^2 - T \wedge n | \mathcal{F}_0) = S_0^2 - 0 \tag{41.95}$$

où $\mathcal{F}_0$ est la tribu engendrée par la variable aléatoire 0, c'est à dire $\{\Omega, \varnothing\}$. Cette tribu est indépendante de toute autre tribu et nous pouvons donc supprimer le conditionnement dans (41.95). Nous avons aussi $S_0 = a$ par définition. Avec tout ça nous avons la majoration

$$E(T \wedge n) = E(S_{T\wedge n}^2) - a^2 \leqslant (a+b)^2 - a^2 \tag{41.96}$$

parce que S_k est toujours positif et entre 0 et $a + b$. En utilisant le lemme 41.10 et en passant à la limite,

$$E(T) \leqslant (a+b)^2 - a^2. \tag{41.97}$$

En particulier, $T \in L^1(\Omega)$ et $P(T < \infty) = 1$.

En suivant exactement les mêmes étapes que dans le lemme 41.1041.10 nous avons aussi

$$\lim_{n\to\infty} S_{T\wedge n} = S_T \tag{41.98}$$

presque partout. De plus nous savons que

$$0 \leqslant S_{T\wedge n}^2 \leqslant (a+b)^2, \tag{41.99}$$

et nous pouvons donc utiliser le théorème de la convergence dominée 16.140 pour dire que

$$\lim_{n\to\infty} E(S_{T\wedge n}^2) = E(S_T^2). \tag{41.100}$$

Nous montrons à présent que $S_{T\wedge n} \xrightarrow{L^2} S_T$. Pour cela nous devons évaluer la limite

$$\lim_{n\to\infty} \int_\Omega |S_{T\wedge n} - S_T|^2. \tag{41.101}$$

La fonction $|S_{T\wedge n} - S_T|^2$ est majorée par $(a+b)^2$ et nous pouvons à nouveau appliquer la convergence dominée :

$$\lim_{n\to\infty} E(|S_{T\wedge n} - S_T|^2) = \lim_{n\to\infty} \int_\Omega |S_{T(\omega)\wedge n}(\omega) - S_{T(\omega)}(\omega)|^2 dP(\omega) = \int_\Omega \lim_{n\to\infty} |S_{T\wedge n} - S_T|^2 = 0. \tag{41.102}$$

La même chose en n'écrivant pas les carrés montre que l'on a aussi $S_{T\wedge n} \xrightarrow{L^1} S_T$.

Il n'y a pas que $n \mapsto S_n^2 - n$ qui est une martingale. Il y a aussi (S_n) lui-même (lemme 41.25). Nous pouvons lui appliquer le théorème d'arrêt 41.12 pour les temps d'arrêts $T \wedge n$ et 0 :

$$E(S_{T\wedge n}) = E(S_0) = a. \tag{41.103}$$

En passant à la limite, $E(S_T) = a$. L'espérance $E(S_T)$ peut par ailleurs être calculée comme

$$E(S_T) = 0 \cdot P(S_T = 0) + (a+b)P(S_T = a+b). \tag{41.104}$$

En égalisant les valeurs (41.103) et (41.104) de $E(S_T)$ nous trouvons

$$\rho = \frac{a}{a+b}. \tag{41.105}$$

Cette formule est assez logique : la probabilité que le joueur gagne est égale à la proportion d'argent en jeu qu'il a amené.

41.4.2.2 Temps moyen de jeu

Nous calculons maintenant l'espérance $E(T)$ du temps de jeu (sans compter les pauses ni les jours de fermeture du casino [12]).

Nous recopions la première égalité de (41.96) sous la forme

$$a^2 = E(S_{T\wedge n}^2 - T \wedge n) \tag{41.106}$$

et nous passons à la limite [13] en sachant que $E(S_T^2) = \rho(a+b)^2$:

$$a^2 = \rho(a+b)^2 - E(T). \tag{41.107}$$

En reprenant la valeur (41.105) de ρ,

$$E(T) = ab. \tag{41.108}$$

Et là, on voit que si le joueur amène 1000 euros contre une banque qui en a un million, et si ils jouent toutes les secondes , on en a pour 32 ans de jeu en moyenne.

Voilà. C'est fini pour la ruine du joueur.

41.4.3 Un petit complément

Nous avons introduit lors de l'équation (41.75) la variable aléatoire $U_n = (p/q)^{S_n}$. Sans aller jusqu'à motiver complètement ce choix, nous nous proposons maintenant de voir que parmi les variables aléatoires $U_n = s^{S_n}$, le choix $s = p/q$ est le seul qui donne une martingale.

Soit donc $U_n = s^{S_n}$ et exprimons le fait que ce soit une martingale. Nous avons

$$E(U_n|\mathcal{A}_{n-1}) = E(s^{S_{n-1}}s^{Y_n}|\mathcal{A}_{n-1}) \tag{41.109a}$$
$$= s^{S_{n-1}}E(s^{Y_n}|\mathcal{A}_{n-1}) \tag{41.109b}$$
$$= s^{S_{n-1}}E(s^{Y_n}). \tag{41.109c}$$

Le passage à (41.109b) se justifie en disant que $s^{S_{n-1}}$ est une variable aléatoire bornée et $\mathcal{A}_{n-1}$-mesurable, et en invoquant proposition 38.41. La variable aléatoire Y_n vaut 1 avec probabilité p et -1 avec probabilité q ; donc l'espérance est vite vue :

$$E(s^{Y_n}) = ps + q\frac{1}{s} \tag{41.110}$$

et nous avons

$$E(U_n|\mathcal{A}_{n-1}) = \left(ps + q\frac{1}{s}\right)s^{S_{n-1}} = (ps + \frac{q}{s})U_{n-1}. \tag{41.111}$$

12. Le joueur est un *vrai* joueur compulsif.
13. Comme il est dit dans La Grande Illusion, à quoi sert un n ? À passer à la limite.

Pour que (U_n) soit une martingale il faut (et il suffit) que

$$ps + \frac{q}{s} = 1. \tag{41.112}$$

Les solutions de cette équation sont $s \in \{1, \frac{p}{q}\}$. C'est évidemment $s = p/q$ qui donne une martingale non triviale. Attention pour être complet, il faut se demander ce qu'il se passe si $s = 0$ séparément parce que manifestement l'équation (41.112) ne traite pas ce cas. Encore une fois, en repartant du début, $s = 0$ ne se révèle pas être une martingale très excitante.

Bref, nous devons poser

$$U_n = \left(\frac{p}{q}\right)^{S_n} \tag{41.113}$$

pour avoir une martingale.

Chapitre 42

Processus de Poisson

42.1 Processus de Poisson

Définition 42.1.
*Une famille de variables aléatoires $(N_t)_{t\geqslant 0}$ est une **processus de Poisson** d'intensité λ s'il existe une suite de variables aléatoires indépendantes et identiquement distribuées $(T_k)_{k\in\mathbb{N}}$ de loi $\mathscr{E}(\lambda)$ telles que*

$$N_t = \sup\{n \geqslant 0 \text{ tel que } \sum_{k=1}^{n} T_k \leqslant t\}. \tag{42.1}$$

Si nous posons $S_n = \sum_{k=1}^{n} T_k$, alors nous avons une expression plus pratique pour N_t :

$$N_t = \sum_{n=1}^{\infty} \mathbb{1}_{\{S_n\leqslant t\}}. \tag{42.2}$$

Nous avons par la proposition 38.103 vu que $N_t \sim \mathscr{P}(\lambda t)$.

Pour chaque $\omega \in \Omega$, la fonction $t \mapsto N_t(\omega)$ est une fonction (pas du tout strictement) croissante à valeurs dans $\mathbb{N}$. Cette fonction part de 0 et fait un saut de taille 1 après des intervalles de temps $T_1(\infty)$, $T_2(\omega)$, etc. Elle est continue à droite.

Nous avons les égalités d'événements suivantes qui sont pratiques :

$$\{s < S_n \leqslant t\} = \{N_t \geqslant n > N_s\} \tag{42.3a}$$
$$\{N_t = n\} = \{S_n \leqslant t \leqslant S_{n+1}\}. \tag{42.3b}$$

Théorème 42.2.
Les variables aléatoires $(N_t)_{t\geqslant 0}$ est un processus de Poisson d'intensité λ si et seulement si elles vérifient les trois propriétés suivantes.

Accroissements indépendants *Pour tout choix $0 < t_0 < t_1 < \dots t_n$, les variables aléatoires $N_{t_{i+1}} - N_{t_i}$ sont indépendantes.*

Accroissements stationnaires *Si $0 < s < t$ et $h > 0$ alors*

$$N_{t+h} - N_{s+h} \overset{\mathcal{L}}{=} N_t - N_s, \tag{42.4}$$

c'est à dire que les accroissements décalés suivent les mêmes lois.

Poisson *Pour tout t nous avons $N_t \sim \mathscr{P}(\lambda t)$.*

Une conséquence des accroissements stationnaires est que $N_t - N_s \overset{\mathcal{L}}{=} N_{t-s} - N_0 = N_{t-s}$ parce que $N_0 = 0$.

Proposition 42.3.
Si (N_t) est un processus de Poisson d'intensité λ, alors

$$\lim_{t\to\infty} N_t = +\infty \tag{42.5}$$

presque sûrement. De plus

$$\lim_{t\to\infty}\frac{N_t}{t}=\lambda \tag{42.6}$$

presque sûrement.

La relation (42.6) est appelée **loi des grands nombres**.

Démonstration. Par définition nous savons que

$$N_t=\sup\{n\geqslant 0 \text{ tel que } S_n\leqslant t\}. \tag{42.7}$$

Évidemment la fonction $t\mapsto N_t$ est croissante, donc la limite

$$\lim_{t\to\infty}N_t(\omega) \tag{42.8}$$

existe dans $[0,\infty]$. Nous pouvons nous restreindre à $t\in\mathbb{N}$ et considérer $L(\omega)=\lim_{n\to\infty}N_n(\omega)$. Par somme télescopique avec $N_0=0$,

$$\frac{N_n}{n}=\frac{\sum_{k=1}^{n}(N_k-N_{k-1})}{n}. \tag{42.9}$$

Étant donné que le processus est de Poisson, les variables aléatoires $(N_k-N_{k-1})_{k=1,\dots,n}$ sont indépendantes et suivent toutes la loi de N_1-N_0, c'est à dire la loi de N_1. Encore par le fait que N_t soit de Poisson nous savons que $N_1\sim\mathscr{P}(\lambda)$. La loi des grands nombres (38.83) appliquée aux variables aléatoires N_k-N_{k-1} nous dit que

$$\frac{N_n}{n}\xrightarrow{p.s.}E(N_1)=\lambda>0. \tag{42.10}$$

Du coup $N_n\to\infty$ et $L(\omega)=\infty$.

Nous démontrons maintenant la loi des grands nombres pour les processus de Poisson. Étant donné que pour les entiers $N_n/n\to\lambda$, pour les réels, si la limite existe, ça ne peut pas être autre chose. Si nous notons $\bar{t}$ la partie entière de $t\in\mathbb{R}^+$,

$$\frac{N_t}{t}=\frac{N_t-N_{\bar{t}}}{t}+\frac{N_{\bar{t}}}{t}. \tag{42.11}$$

Le second terme est relativement simple à traiter :

$$\frac{N_{\bar{t}}}{t}=\underbrace{\frac{N_{\bar{t}}}{\bar{t}}}_{\to\lambda}\cdot\underbrace{\frac{\bar{t}}{t}}_{\to 1}. \tag{42.12}$$

où nous avons utilisé le premier point, $\bar{t}$ étant entier. Pour le premier terme nous savons que $t\mapsto N_t$ est croissante et donc que

$$\frac{N_t-N_{\bar{t}}}{t}\leqslant\frac{N_{\bar{t}+1}-N_{\bar{t}}}{t}=\frac{N_{\bar{t}+1}-N_{\bar{t}}}{\bar{t}+1}\frac{\bar{t}+1}{t}. \tag{42.13}$$

Le second facteur tend vers 1 lorsque $t\to\infty$. Le premier s'écrit

$$\frac{N_n-N_{n-1}}{n} \tag{42.14}$$

et tend vers zéro en tant que terme général de la série (42.9) qui converge. □

Proposition 42.4.
La variable aléatoire N_t/t est un estimateur sans biais de λ. De plus il converge vers λ en moyenne quadratique.

Démonstration. Vu que $N_t/t \to \lambda$ presque sûrement, la variable aléatoire N_t/t est un estimateur de λ. Le fait qu'il soit sans biais a été fait dans l'exemple 39.25.

D'autre part nous avons (voir théorème 38.96)

$$\text{Var}\left(\frac{N_t}{t}\right) = \frac{1}{t^2}\text{Var}(N_t) = \frac{\lambda}{t}. \tag{42.15}$$

En appliquant la formule $\text{Var}(X) = E(X^2) - E(X)^2$ à $X = N_t/t$ nous trouvons

$$E\left(\frac{N_t^2}{t^2}\right) = \frac{\lambda}{t} + \lambda^2. \tag{42.16}$$

Cela montre que $\frac{N_t}{t} \xrightarrow{L^2} \lambda$. □

Pour le théorème central limite d'un processus de Poisson, nous visons un résultat du style de

$$\frac{\frac{1}{n}\sum_i X_i - mn}{\sigma\sqrt{n}} \xrightarrow{\mathscr{L}} \mathscr{N}(0,1). \tag{42.17}$$

Nous écrivons le théorème central limite pour le nombre de sauts que le processus de Poisson a connu en un temps t. Le rôle de la moyenne empirique est joué par N_t. Nous considérons avoir fait *une seule expérience* qui a duré un temps t. Donc le rôle de n est joué par 1 (et non t comme on pourrait le croire). Pour le reste, le nombre de succès en un temps t d'une variable aléatoire exponentielle de paramètre λ est une variable aléatoire de Poisson de paramètre λt, en vertu de ce qui est raconté au point 38.5.8. C'est cela qui motive l'énoncé suivant.

Théorème 42.5 (Théorème central limite pour les processus de Poisson)**.**
Si $(N_t)_{t>0}$ est un processus de Poisson de paramètre λ, alors nous avons

$$\frac{N_t - \lambda t}{\sqrt{\lambda t}} \xrightarrow{\mathscr{L}} \mathscr{N}(0,1). \tag{42.18}$$

Remarque 42.6.
Avant de nous lancer dans la démonstration, remarquons que si nous nous limitons à $t \in \mathbb{N}$, alors nous avons

$$\frac{N_n - \lambda n}{\sqrt{\lambda n}} = \frac{\sum_{k=1}^n (N_k - N_{k-1}) - \lambda n}{\sqrt{\lambda n}} \tag{42.19}$$

or par définition nous avons les égalités de lois

$$N_k - N_{k-1} \sim N_1 \sim \mathscr{P}(\lambda), \tag{42.20}$$

donc

$$\frac{S_n - \lambda n}{\sqrt{\lambda n}} = \frac{\frac{1}{n}S_n - \lambda}{\frac{\sqrt{\lambda n}}{n}} = \frac{\frac{1}{n}S_n - \lambda}{\sqrt{\lambda}/\sqrt{n}}, \tag{42.21}$$

ce qui est exactement le théorème central limite pour une suite de lois de Poisson [1].

Démonstration. Nous écrivons $\bar{t}$ la partie entière de $\bar{t}$ et nous décomposons :

$$\frac{N_t - \lambda t}{\sqrt{\lambda t}} = \underbrace{\frac{N_t - N_{\bar{t}}}{\sqrt{\lambda t}}}_{A} + \underbrace{\frac{N_{\bar{t}} - \lambda \bar{t}}{\sqrt{\lambda t}}}_{B} + \underbrace{\frac{\lambda \bar{t} - \lambda t}{\sqrt{\lambda t}}}_{C}. \tag{42.22}$$

En ce qui concerne le terme B, nous avons

$$B = \sqrt{\frac{\bar{t}}{t}}\frac{N_{\bar{t}} - \lambda \bar{t}}{\sqrt{\lambda \bar{t}}} \to \mathscr{N}(0,1). \tag{42.23}$$

1. Au fait près que nous devrions encore montrer que S_n est de carré intégrable.

Notons que nous utilisons le fait que si $a_n \to 1$ (en tant que suite de nombre) et si $X_n \to \mathcal{N}(0,1)$ (limite en loi), alors $a_n X_n \to \mathcal{N}(0,1)$ en loi.

Le terme C est également facile parce que $\lambda\bar{t} - \lambda t$ est majoré en norme par λ. Du coup

$$-\frac{\lambda}{\sqrt{\lambda t}} \leqslant C \leqslant \frac{\lambda}{\sqrt{\lambda t}}. \tag{42.24}$$

Donc $\lim_{t\to\infty} C = 0$.

Reste à travailler sur A. Vu que $t \mapsto N_t$ est croissante, la différence $N_t - N_{\bar{t}}$ est positive. Soit $\eta > 0$, nous avons

$$P(|A| > \eta) = P(N_t - N_{\bar{t}} > \sqrt{\lambda t}\eta) \leqslant P\big(N_{\bar{t}+1} - N_{\bar{t}} \geqslant \sqrt{\lambda t}\eta\big) = P(N_1 \geqslant \eta\sqrt{\lambda t}) \tag{42.25}$$

parce que nous savons que $N_{\bar{t}+1} - N_{\bar{t}} \sim N_1 \sim \mathcal{P}(\lambda)$. En vertu des propriétés de la loi de Poisson,

$$\lim_{t\to\infty} P(N_1 \geqslant \eta\sqrt{\lambda t}) = 0. \tag{42.26}$$

En effet si Z est une variable aléatoire de Poisson de paramètre λ nous avons

$$P(Z > l) = \sum_{k=l}^{\infty} P(Z = k) = e^{-\lambda} \sum_{k=l}^{\infty} \frac{\lambda^k}{k!}. \tag{42.27}$$

Nous reconnaissons la queue de série de e^{λ}, qui tend donc vers zéro lorsque $l \to \infty$. Nous avons donc prouvé que

$$\lim_{t\to\infty} P\big(|A| > \eta\big) = 0, \tag{42.28}$$

c'est à dire la convergence en probabilité de A vers zéro.

Nous avons montré que

$$B + C \xrightarrow{\mathscr{L}} U \sim \mathcal{N}(0,1) \tag{42.29a}$$

$$A \xrightarrow{P} 0. \tag{42.29b}$$

Le lemme de Slutsky (38.74) nous avons une convergence du couple

$$(A, B + C) \xrightarrow{\mathscr{L}} (0, U). \tag{42.30}$$

Utilisant le corollaire 38.76, nous trouvons la convergence en loi

$$A + (B + C) \xrightarrow{\mathscr{L}} 0 + U, \tag{42.31}$$

ce qu'il fallait. □

42.2 Quelques trucs sur la simulation

Le théorème ergodique dit que

$$\pi(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}_{X_k = x}. \tag{42.32}$$

C'est avec cela qu'on calcule $\pi(x)$ à partir d'une simulation de chaîne de Markov.

42.2.1 Le théorème central limite pour Markov

Théorème 42.7 (Version allégée)**.**
Si (X_n) est irréductible et positive récurrente, alors pour toute fonction f,

$$\frac{1}{\sqrt{N}}\left[\sum_{k=1}^{N} -N\int f d\pi\right] \xrightarrow{\mathscr{L}} \mathscr{N}(0,\sigma^2) \tag{42.33}$$

où σ^2 dépend de la fonction f et de la chaîne de Markov.
Ici, $\int f d\pi = \sum_{x\in E} f(x)\pi(x)$.

Nous allons simuler la variable aléatoire

$$Z = \frac{1}{\sqrt{N}}\left[\sum f(X_k) - N\sum_{x\in E} f(x)\pi(x)\right] \tag{42.34}$$

et puis on va mettre sa réalisation dans un histogramme. Dans le cas où on prend $f(i) = \mathbb{1}_{i=i_0}$, il y a de la simplification dans l'intégrale qui devient

$$Z = \frac{1}{\sqrt{N}}\left[\sum_{i=1}^{N} \mathbb{1}_{X_k=i_0} - N\pi(i_0)\right]. \tag{42.35}$$

42.2.2 Feuille 5

On pose

$$D_n = \sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|. \tag{42.36}$$

On en génère un milliers de fois D_n, on note $D_n^{(k)}$ ces réalisations, et on regarde ce que vaut

$$\frac{1}{1000}\sum_{k=1}^{1000} \mathbb{1}_{D_n^{(k)}\geqslant c}. \tag{42.37}$$

Cela nous donne une approximation de

$$P\big(\sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \geqslant c\big). \tag{42.38}$$

Note que chacun des $D_n^{(k)}$ demande de créer un nouveau vecteur Y_i de lois qu'on veut regarder. Par exemple de loi exponentielle.

42.2.3 Feuille 6

Pour créer une fonction qui renvoie i avec probabilité p_i pour $i = 1, 2, 3$, on peut faire

$$U \sim \mathscr{U}[0,1] \tag{42.39}$$

et puis on a

$$P(U < p_0) = p_0 \tag{42.40a}$$
$$P(p_0 < U < p_0 + p_1) = p_1 \tag{42.40b}$$
$$P(p_0 + p_1 < U < p_2) = p_2. \tag{42.40c}$$

Une façon de faire une loi uniforme $[0,1]$ est de faire `rand`

42.2.4 Feuille 7

L'échantillon est $(Y_1, \dots, Y_n)$ et nous écrivons le vecteur

$$Y = X\beta + \epsilon \tag{42.41}$$

où $Y \sim \mathcal{N}(X\beta, \sigma^2 \mathrm{Id})$ et $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$. Nous utilisons le principe de maximum de vraisemblance. Soit $(y_1, \dots, y_n)$ un échantillon et

$$P_\theta(y_1, \dots, y_n) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_i - X_i^t\beta}{\sigma}\right)^2\right]. \tag{42.42}$$

L'astuce est de faire que $y_i - X_i^t\beta$ est la $i$ième composante du vecteur $Y - X\beta$ et donc la somme qui est dans l'exponentielle devient la norme de $Y - X\beta$:

$$f_\theta(y_1, \dots, y_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\|Y - X\beta\|^2\right]. \tag{42.43}$$

On passe au logarithme et on dérive par rapport à σ^2. Attention : la variable est σ^2, donc la dérivée de σ^2 est 1 et non 2σ. Bref, on trouve

$$\sigma^2 = \frac{1}{2n}\|U + X\beta\|. \tag{42.44}$$

42.2.5 Simuler des lois conditionnelles

Nous voulons générer des couples (X, Y) tels que Y prend les valeurs 0 ou 1 et tels que

$$\begin{cases} P(X|Y=0) \sim \mathscr{E}(\lambda_0) & \text{(42.45a)} \\ P(X|Y=1) \sim \mathscr{E}(\lambda_1). & \text{(42.45b)} \end{cases}$$

Le plus simple est de générer une liste

$$(X_1, 0) \qquad (X_4, 1) \tag{42.46a}$$
$$(X_2, 0) \qquad (X_5, 1) \tag{42.46b}$$
$$(X_3, 0) \qquad (X_6, 1) \tag{42.46c}$$

avec $X_1, X_2, X_3 \sim \mathscr{E}(\lambda_0)$ et $X_4, X_5, X_6 \sim \mathscr{E}(\lambda_1)$.

Avec cette méthode cependant la liste est triée et en plus on a autant de 1 que de 0. On peut faire un peu plus technologique pour corriger cela. Pour créer un couple, on commence par $Y \sim \mathscr{B}(p)$ et puis suivant que $Y = 0$ ou $y = 1$, on génère $X \sim \mathscr{E}(\lambda_0)$ ou $X \sim \mathscr{E}(\lambda_1)$.

Chapitre 43

Utilisation dans les autres sciences

Dans ce chapitre nous donnons des applications de divers théorèmes dans les autres sciences que la mathématique.

43.1 Démystification du MRUA

43.1.1 Preuve de la formule

Nous sommes maintenant en mesure de donner une démonstration complète de la formule du MRUA :

$$x(t) = \frac{at^2}{2} + v_0 t + x_0. \tag{43.1}$$

Au niveau de la physique, nous considérons un mobile qui se déplace avec une accélération constante a. Nous notons par v_0 sa vitesse initiale et par x_0 sa position initiale.

Nous savons que, pour tout mouvement, si $x(t)$ est la position en fonction du temps, et si $v(t)$ et $a(t)$ représentent la vitesse et l'accélération en fonction du temps, alors

$$v(t) = x'(t) \quad \text{et} \quad a(t) = v'(t) = x''(t). \tag{43.2}$$

Afin de trouver $x(t)$ en connaissant $a(t)$, il « suffit » donc de prendre deux fois la primitive. Essayons ça dans le cas facile du MRUA où $a(t) = a$ est constante.

La vitesse $v(t)$ doit être une primitive de la constante a. Il est facile de voir que $v(t) = at$ est une primitive de a. Par le corollaire 14.143(bis),

$$v(t) = at + C_1 \tag{43.3}$$

pour une certaine constante C_1. Afin de fixer C_1, il faut faire appel à la physique : d'après la formule (43.3), la vitesse initiale est $v(0) = C_1$. Donc il faut identifier C_1 à la vitesse initiale : $C_1 = v_0$. Nous avons donc déjà obtenu que

$$v(t) = at + v_0. \tag{43.4}$$

Afin de trouver $x(t)$, il faut trouver une primitive de $v(t)$. Il n'est pas très difficile de voir que $at^2/2 + v_0 t$ fonctionne, donc il existe une constante C_2 telle que

$$x(t) = \frac{at^2}{2} + v_0 t + C_2. \tag{43.5}$$

Encore une fois, regardons la condition initiale : la formule donne comme position initiale $x(0) = C_2$, et donc nous devons identifier C_2 avec la position initiale x_0. En définitive, nous avons bien

$$x(t) = \frac{at^2}{2} + v_0 t + x_0. \tag{43.6}$$

Cette formule est donc maintenant *démontrée* à partir de la seule définition de la vitesse comme dérivée de la position et de l'accélération comme dérivée de la vitesse.

Remarquons cependant que la preuve complète fut *très* longue. En effet, nous avons utilisé les règles de dérivation de la proposition 14.121, pour la démonstration desquels, les résultats 14.22 et 14.21 ont étés utiles. Mais nous avons surtout utilisé le corollaire 14.143(bis) qui repose sur le théorème de Rolle 14.133, qui lui-même demande le théorème de Borel-Lebesgue 14.108 dans lequel la notion d'ensemble compact a été cruciale.

43.1.2 Interprétation graphique

La distance parcourue $x(t)$ en un temps t est la primitive de la vitesse. Nous avons, par ailleurs, que l'opération inverse de la dérivée donnait la surface. Pour reprendre les mêmes notations, nous notons $S_v(t)$ la surface contenue en dessous de la fonction v entre 0 et x. Nous ne serions donc pas étonné que

$$S_v(t) = \frac{at^2}{2} + v_0 t + x_0 \tag{43.7}$$

soit la surface en dessous de la fonction $v(t) = at + v_0$. Nous voyons que la surface totale sous la fonction $v(t) = at + v_0$ est exactement

$$S_v(t) = \frac{at^2}{2} + v_0 t. \tag{43.8}$$

Cela est un bon début, mais hélas nous ne retrouvons pas le terme « $+x_0$ » de la formule (43.7). Cela n'est pas tout à fait étonnant parce que nous savons que la surface sous une fonction était *une* primitive de la fonction, mais nous n'avons pas dit *laquelle*. D'après le fameux corollaire 14.143(bis), la primitive n'est définie qu'à une constante près. Ici, c'est la constante x_0 qu'on a perdue en chemin.

Nous parlerons plus en détail du lien entre les surfaces et les primitives dans la section dédié à l'intégration.

43.2 Relativité en mécanique newtonienne

43.2.1 Relativité du mouvement

Prenons quelqu'un qui cours le cent mètres en onze secondes. Par rapport à un spectateur dans les gradins, il se sera déplacé de cent mètres. Mais si je cours à côté de lui de telle façon à avoir parcouru 80 mètres le temps qu'il en fasse cent, alors par rapport à moi l'athlète ne se sera déplacé que de 20 mètres. Par contre, par rapport à mon chronomètre, il aura également mit onze secondes : ce n'est pas parce que je cours que mon chronomètre s'affole !

Entre moi et les spectateurs, on a donc une loi de transformation

$$x' = x - vt \qquad\qquad t' = t. \tag{43.9}$$

C'est à dire que la distance x' qu'aura parcouru l'athlète par rapport à moi vaut la distance x parcourue par le spectateur moins la vitesse que j'ai courue moi-même, c'est à dire moins vt.

43.2.2 Bob et Alice

Formalisons le concept de changement de repères. Pour cela, prenons deux amoureux, Bob et Alice[1]. Mettons que Bob reste assis sur un banc pendant qu'Alice cours en ligne droite à une vitesse v. Tout deux déclenchent leur chronomètre quand Alice passe devant Bob. À tout moment, Bob et Alice ont leur repères de temps et d'espace. Par exemple si après un temps t, Alice voir une peau de banane à 1 mètre devant elle, elle va dire « Il y a une peau de banane à un mètre. », tandis que Bob va dire « Il y a une peau de banane à $(1 + vt)$ mètres ».

Plus généralement, s'il se passe quelque chose à la position x au temps t pour Bob, ce quelque chose se passera au temps $t' = t$ à l'endroit $x' = x - vt$ pour Alice parce qu'en un temps t, elle aura déjà avancé d'une distance vt.

Ça c'est ce dont tout le monde était persuadé depuis Galilée jusqu'au début du vingtième siècle.

1. C'est plus poétique que dire « soient A et B deux observateurs ».

43.3 Invariance de la vitesse de la lumière

43.3.1 Champ de gravitation et électrique

Nous savons que que la force de gravitation s'écrit :

$$F_{grav} = G\frac{mm'}{r^2},$$

tandis que la force électrique entre deux charges q et q' est donnée par

$$F_{elec} = k\frac{qq'}{r^2}. \tag{43.10}$$

Nous avons aussi fait remarquer que dans le cas de la gravitation, la force a l'air d'être instantanée, et que cela posait quelques problèmes conceptuels. La force électrique a apparemment le même problème. Une différence entre les deux est qu'une charge électrique c'est tout petit et qu'on peut expérimenter à souhait, tandis que pour avoir une masse dont on peut mesurer le champ de gravitation correctement, il faut quelque chose grand comme la Terre [2].

43.3.1.1 Finitude de la vitesse de propagation de la force électrique

Si un micro est placé juste à côté de ton oreille, et qu'il commence à faire biiiiip, tu l'entends directement. Quand il s'arrête, tu ne l'entends plus. Si le micro est placé à 600 m de toi, tu ne commenceras à l'entendre que deux secondes après le commencement du son, et tu continueras à l'entendre deux secondes après qu'il ait fini.

Eh bien, pour la force électrique, on a pu mesurer que c'est la même chose (sauf que ça va beaucoup plus vite). Si on place une charge quelque part, on ne ressent la force (43.10) qu'après qu'elle ait eut le temps d'arriver. Si on déplace la charge électrique, on continue à ressentir la même force pendant un certain temps : il faut que la modification du champ électrique ait le temps d'arriver. Exactement comme quand on fait des remous quelque part dans un étang : il faut du temps que les remous arrivent plus loin.

On a pu faire des dizaines d'expériences de ce type avec l'électricité, le magnétisme et la lumière ; et les résultats sont clairs : il faut du temps pour que ça se déplace. Tout cela provoque des ondes électromagnétiques qui se déplacent à une vitesse finie. On peut produire de telles ondes avec n'importe quel courant électrique alternatif.

43.3.1.2 Pourquoi pas la gravitation ?

La gravitation telle que donnée par Newton pose le même problème de vitesse de propagation que l'électricité. Est-ce qu'en réalité la gravitation se propage également à une vitesse finie ?

Avec la gravitation c'est beaucoup plus compliqué parce qu'elle est beaucoup plus faible, et donc c'est beaucoup plus difficile à détecter. D'après la théorie d'Einstein de la gravitation, la gravitation devrait également produire des ondes gravitationnelles. Seulement, si un simple courant électrique suffit pour mesurer une onde électromagnétique, affin de mesurer une onde gravitationnelle, il faudrait un déplacement de masse de l'ampleur d'une étoile qui explose. Or ça, on ne sait pas produire dans un laboratoire. Les physiciens sont donc pour l'instant (2009) en train d'attendre qu'une étoile explose pas trop loin d'ici affin d'être capable de mesurer une onde gravitationnelle.

L'existence de ces ondes de gravitation ne fait aucun doute dans la tête d'aucun physicien parce qu'elles sont une conséquence logique (et mathématique) de la théorie de la relativité générale, laquelle a déjà eut beaucoup de confirmations expérimentales. Mais comme on est dans le cadre d'une science expérimentale, il faut être patient et attendre d'en avoir effectivement observée une avant de dire avec certitude que ça existe.

2. Une autre différence fondamentale est qu'il existe des charges électriques négatives, mais pas de masses négatives ; de ce fait on ne peut pas construire d'isolant gravitationnel, contrairement aux isolants électriques qui existent. Cela augmente encore la difficulté de faire des expériences avec la gravitation.

43.3.2 Support du champ : pas d'éther

Nous avons dit qu'une onde électromagnétique se propage comme une onde sonore (quoique beaucoup plus vite). Une question se pose alors. En effet, une onde sonore est matérialisé par de l'air qui vibre. Qu'est-ce qui vibre pour une onde électromagnétique ?

Étant donné que les ondes électromagnétiques se propagent dans le vide (c'est pour ça que la radio fonctionne dans l'espace), la question est problématique. Les physiciens ont donc supposé que tout l'univers était rempli d'un fluide invisible appelé **l'éther**. L'électromagnétisme consiste en une vibration de l'éther, exactement comme l'acoustique consiste en une vibration de l'air.

En fait, vérifier cette hypothèse n'est pas très compliqué. En effet il n'y a aucune raison que l'éther suive la Terre dans son mouvement. Or la Terre se déplace à environ 30 km/s autour du Soleil. Donc les ondes électromagnétiques doivent se propager plus vite dans le sens du mouvement de la Terre que dans le sens perpendiculaire. Tout comme le son se propage plus vite dans le sens du vent.

La célèbre expérience de Michelson-Morley a mesuré cet effet ... et ce fut la consternation : il n'y a aucun effet ! Or, la lumière se déplace à 300.000 km/s ; une variation de 30 km/s devrait être détectable !

Mais rien ! On a recommencé les expériences dans tout les sens, à tous les mois de l'année, à tous les endroits de la Terre. On n'a pas observé un poil de variation de la vitesse de la lumière. Et ça, ça pose un gros problème à la physique.

43.3.3 Le problème

Si je joue au football dans un train qui avance à 100 km/h et que je lance une balle à 20 km/h, quelqu'un au sol mesura la vitesse de la balle soit à 120 km/h soit à 80 km/h d'après que l'on ait shooté vers l'avant ou l'arrière du train. Cela paraît logique. Mais ce qu'on vient de voir c'est que ça ne marche pas avec la lumière.

Si un train avance à 100.000 km/s et qu'on y allume une lampe de poche, la lumière avancera à 300.000 km/s par rapport au train et 400.000 km/s par rapport au sol. Non ! Justement pas ! La lumière avancera quand même à 300.000 km/s par rapport au sol.

Là encore, on a fait des dizaines d'expériences partout, sur Terre, dans des avions, dans l'espace avec des atomes, des lampes de poche et des horloges atomique, et dans tous les sens, le sens de déplacement de la Terre, le sens inverse, le sens perpendiculaire, vers le haut, vers le bas : rien ! Personne n'a jamais observé un rayon de lumière se déplacer à une autre vitesse que 300.000 km/s.

Le problème est que le principe d'addition des vitesses est faux pour la lumière. Puisque l'expérience nous force, nous devons faire avec.

Loi numéro 1.
La réalité est que la vitesse de la lumière est la même dans tous les référentiels. On note c cette vitesse. C'est une constante fondamentale de la Nature.

Étant donné que c'est une loi expérimentale, nous n'en pouvons rien. C'est la nature qui est comme ça. En particulier tu ne peux pas en vouloir à ton prof de physique d'avoir inventé une théorie compliquée. Ce n'est pas lui qui l'a inventée et ce n'est pas de sa faute.

43.4 Conséquences

C'est maintenant que les choses vraiment graves commencent (cela soit dit sans vouloir te faire peur). Affin d'un peu simplifier les choses, nous n'allons étudier que les mouvements en une dimension, c'est à dire sur une droite.

43.4.1 Ligne d'univers

Un événement a une coordonnée (t, x). Si je pose un objet juste à mes pieds (disons en $x = 0$), ses coordonnées seront à tout moment $(t, 0)$. Il est bon de voir cette coordonnée comme l'équation

paramétrique d'une droite horizontale dans le plan des coordonnées t et x. Plus généralement quand un mobile effectue un mouvement $x(t)$, on appelle la **ligne d'univers** du mobile la ligne (pas forcément droite) $(t, x(t))$. Dans le premier exemple, on avait $x(t) = 0$ pour tout t.

Le cas d'un mobile se déplaçant à vitesse constante v donne comme ligne d'univers la droite[3] $(t, x_0 + vt)$, et un objet qui se déplace selon un MRUA a comme ligne d'univers

$$\big(t, x_0 + v_0 t + \frac{at^2}{2}\big).$$

43.4.2 Transformations de Lorentz

Reprenons les amours scientifiques de Bob et Alice, mais cette-fois, analysons celles-ci en tenant compte du fait que la vitesse de la lumière soit invariante. Maintenant, si Bob voit se passer quelque chose au temps t à l'endroit x, on va dire qu'Alice voit cette chose au temps t' à la position x', et on va chercher (t', x') en fonction de (t, x).

Posé en termes mathématiques, le problème s'énonce ainsi : trouver les fonctions f et g telles que les formules

$$t' = f(t, x) \tag{43.11a}$$
$$x' = g(t, x) \tag{43.11b}$$

donnent les coordonnées vues par Alice pour un événement vu par Bob à l'instant t au point x. Une première étape importante est franchie par la proposition suivante[4].

Proposition 43.1.
Les fonctions f et g contenues dans les transformations (43.11) sont nécessairement linéaires (affines), c'est à dire qu'elles doivent s'écrire sous la forme

$$t' = \alpha t + \beta x + p$$
$$x' = \gamma t + \delta x + q$$

pour certaines fonctions α, β, γ, δ, p et q de la vitesse d'Alice relativement à Bob.

Démonstration. Pendant qu'Alice court et que Bob la regarde, Ève tente de lancer une pierre sur Alice (Ève est jalouse). Bob et Alice regardent deux événements. Le premier est la pierre qui quitte la main de Ève, et le second est la pierre qui percute le sol. Pour Bob, le jet s'est passée au temps t_0 au point x_0, et la pierre touche le sol un petit peu plus tard, au temps $t_0 + \Delta t$ et un peu plus loin, au point $x_0 + \Delta x$. Bob écrit donc ceci sur sa feuille de papier :

$$E_1 = (t_0, x_0)$$
$$E_2 = (t_0 + \Delta t, x_0 + \Delta x),$$

tandis qu'Alice, en observant les mêmes deux événements, aura noté

$$E_1' = \big(f(t_0, x_0), g(t_0, x_0)\big)$$
$$E_2' = \big(f(t_0 + \Delta t, x_0 + \Delta x), g(t_0 + \Delta t, x_0 + \Delta x)\big).$$

Bob et Alice se demandent combien de temps la pierre est restée en l'air et quelle distance elle a parcourue. Par le principe général d'homogénéité, les deux seules quantités pertinentes (qui ont un sens physique) pour Bob sont $(t_0 + \Delta t) - t_0$ et $(x_0 + \Delta x) - x_0$, c'est à dire Δt et Δx. En effet, si Bob avait choisi de s'asseoir autre part et si Alice avait commencé à courir un peu plus tard, ça n'aurait rien changé à la longueur du jet de Ève.

D'une façon ou d'une autre, il doit exister une façon de déduire les mesures de Alice en connaissant celles de Bob ; je ne connais pas avec quelles formules, mais ces formules ne peuvent contenir que Δt, Δx et v parce que ce sont les seules quantités qui définissent tous les événements.

3. bon exercice de révision de ton cours de math de vérifier que c'est une droite.
4. dont je te suggère fortement de ne pas lire la preuve si tu ne veux pas que ton cerveau éclate.

Cela dit, Alice va caractériser le mouvement de la pierre avec la différence des coordonnée entre le jet et la chute sur le sol mesurées par elle-même. En d'autres termes, pour Alice ce qui compte c'est la différence entre E_1' et E_2', soit

$$\big(f(t_0+\Delta t,x_0+\Delta x),g(t_0+\Delta t,x_0+\Delta x)\big)-\big(f(t_0,x_0),g(t_0,x_0)\big). \tag{43.12}$$

Mais nous venons de signaler que ce qu'Alice mesurait devait pouvoir être exprimé en termes de Δt et Δx. Nous concluons que la différence (43.12) ne dépend en fait pas de x et t mais seulement de Δt et Δt.

Prenons maintenant une notation plus compacte et notons $X=(t,x)$, $\Delta X=(\Delta t,\Delta x)$ puis $F=(f,g)$. Avec ça, l'expression (43.12) se note $F(X+\Delta X)-F(X)$. Comme mentionné, cette expression ne dépend que de Δx. En particulier, elle ne dépend pas de X.

Maintenant tu vas comprendre pourquoi on apprend les dérivées dans ton cours de math. Comme $F(X+\Delta X)-F(X)$ ne dépend pas de X, le rapport $\big(F(X+\Delta X)-F(X)\big)/\Delta X$ non plus. La limite de ce rapport quand ΔX tend vers zéro non plus :

$$\lim_{\Delta X\to 0}\frac{F(X+\Delta X)-F(X)}{\Delta X} \tag{43.13}$$

ne dépend pas de X. Tu reconnais là la dérivée de F au point X. En d'autres termes, $F'(X)$ est constante, elle ne dépend pas de X. Disons donc que $F'(X)=a$. Tu connais beaucoup de fonctions dont la dérivée est constante ? Non ? En effet, il n'y en a pas beaucoup. Les fonctions qui vérifient $F'(X)=a$ signifie sont toutes de la forme

$$F(X)=aX+b.$$

À ce niveau, il convient de re-déballer les notations compactes : si $a=\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}$ et $b=(p,q)$ on trouve

$$f(t,x)=\alpha t+\beta x+p \tag{43.14a}$$
$$g(t,x)=\gamma t+\delta x+q, \tag{43.14b}$$

comme annoncé.

□

Nous savons que lorsque $(t,x)=(0,0)$, alors $(t',x')=(0,0)$. En effet, Bob et Alice ont lancés leurs chronos en même temps au moment où ils étaient au même endroit. En mettant $(t,x)=(0,0)$ dans les équations (43.14), on trouve $(t',x')=(p,q)$, et donc $p=q=0$. Ça fait une chose de réglée ; on se retrouve avec

$$\begin{cases} t'=\alpha t+\beta x & \text{(43.15a)}\\ x'=\gamma t+\delta x. & \text{(43.15b)}\end{cases}$$

Quelles sont les contraintes à vérifier pour que ces transformations décrivent correctement la physique que l'on cherche à écrire ?

(1) Il faut que les transformations décrivent correctement que Alice avance à une vitesse v par rapport à Bob,

(2) dans le même ordre d'idée, il faut que l'on trouve que Bob avance à la vitesse $-v$ par rapport à Alice,

(3) il faut que si Alice et Bob observent un rayon lumineux, ce rayon aille à la vitesse c par rapport à Alice et à la *même* vitesse c par rapport à Bob,

(4) enfin, il faut avoir le principe de relativité, c'est à dire que comme les équations (43.15) disent ce que Alice voit en fonction de ce que Bob voit, on demande que les équations qui disent ce que Bob voit en fonction de ce que Alice voit soient les mêmes. En d'autres termes, il faut que les transformations et les transformations inverses soient les mêmes au changement près du signe de v.

Étudions une à une ce que chacune de ses contraintes impose. Rappelons que (t,x) et (t',x') sont les coordonnées que Bob et Alice mettent sur le même événement. Par exemple sur l'événement qui consiste à ce que Ève, par jalousie envers Bob, jette une peau de banane sous les pieds d'Alice. Cet événement a lieu à un certain moment, à un certain endroit. C'est ce moment et cet endroit qui sont notés (t,x) et (t',x').

(1) Les coordonnées (t,x) et (t',x') peuvent décrire n'importe quoi. Regardons les coordonnées de Alice qui cours. Pour Alice, cela correspond à $(t',x') = (t',0)$ parce que si x' désigne la position de Alice par rapport à Alice, alors x' est toujours nul. Pour Bob par contre, Alice ne reste pas en place, mais se déplace à une vitesse v. C'est à dire que si (t,x) sont les coordonnées de Alice pour Bob, alors $x/t = v$. Écrivons les équations (43.15) en tenant compte de tout ça : avec $x' = 0$, la seconde équation donne

$$0 = \gamma t + \delta x, \tag{43.16}$$

d'où on déduit que $x/t = -\gamma/\delta$. En imposant que cela soit v, on trouve $\gamma = -v\delta$, et on ré-écrit les transformations en tenant compte de ça :

$$\begin{cases} t' = \alpha t + \beta x & \text{(43.17a)} \\ x' = -v\delta t + \delta x. & \text{(43.17b)} \end{cases}$$

Nous voilà débarrassé d'un paramètre.

(2) Maintenant, on regarde ce qu'il se passe quand (t,x) et (t',x') décrivent les positions de Bob. On a $(t,x) = (t,0)$ parce que selon Bob, Bob est au repos. Les équations deviennent :

$$t' = \alpha t \qquad\qquad x' = -v\delta t. \tag{43.18}$$

La vitesse de Bob par rapport à Alice est $-v$, donc on exige que $x'/t' = -v$, c'est à dire que

$$\frac{-v\delta t}{\alpha t} = -v,$$

ce qui implique que $\delta = \alpha$. On avance encore un peu. Écrivons à nouveau les lois de transformation en en tenant compte :

$$\begin{cases} t' = \alpha t + \beta x & \text{(43.19a)} \\ x' = -v\alpha t + \alpha x. & \text{(43.19b)} \end{cases}$$

(3) Si maintenant Bob et Alice regardent un même rayon de lumière (comme c'est romanesque !), alors (t,x) et (t',x') expriment les coordonnées d'un rayon lumineux expriment les coordonnées d'un rayon lumineux. Le fait que Bob regarde un rayon lumineux fait que $x = ct$, et donc que les coordonnées du rayon lumineux, observé par Alice sont :

$$t' = \alpha t + \beta ct \qquad\qquad x' = -\alpha vt + \alpha ct. \tag{43.20}$$

L'invariance de la vitesse de la lumière exige que Alice mesure une vitesse c pour le rayon de lumière, c'est à dire $x' = ct'$. On exige donc que

$$-\alpha vt + \alpha ct = c\alpha t + \beta c^2 t,$$

ce qui implique que

$$\beta = -\frac{\alpha v}{c^2}.$$

Une fois de plus, l'avant-dernière, on ré-écrit les lois de transformations en tenant compte de ce fait ; mais cette fois, on fait l'effort d'écrire aussi les transformations inverses :

$$t' = \alpha t - \frac{\alpha v}{c^2}x \qquad\qquad t = \frac{1}{\Delta}\big(\alpha t' + \frac{\alpha v}{c^2}x'\big) \tag{43.21}$$

$$x' = -\alpha vt + \alpha x \qquad\qquad x = \frac{1}{\Delta}(\alpha vt' + \alpha x') \tag{43.22}$$

où $\Delta = \alpha^2 - \frac{\alpha^2 v^2}{c^2}$ que tu noteras au passage être toujours positif, et nul uniquement quand $v = c$.

(4) Maintenant il reste à imposer le principe de relativité. Les transformations (43.21) montrent comment Alice voit le monde (c'est à dire (t', x')) en fonction de la façon dont Bob voit le monde (c'est à dire (t, x)). On se demande donc quelle seraient, pour Bob, les coordonnées (t, x) d'un point vu en (t', x') par Alice. Cela signifie que l'on impose que les deux systèmes (43.21) soient en réalité les mêmes, à un changement de signe près.
Attention : il *a priori* faux de dire qu'en changeant le signe de v dans $\alpha v/c^2$, j'obtiens $-\alpha v/c^2$ parce que α est une fonction de v. En réalité, il faut noter $\alpha(v)v/c^2$ et donc le changement de signe de v donne $-\alpha(-v)v/c^2$. Ceci étant clair, on peut un petit peu calculer.
Commençons par égaliser le coefficient de x dans t' à celui de x' dans t, en changeant le signe de v :

$$\frac{\alpha(-v)v}{c^2} = \frac{\alpha(v)v}{c^2},$$

et donc $\alpha(v) = \alpha(-v)$. Ça c'est une bonne nouvelle. Égalisons maintenant le coefficient de t dans t' à celui de t' dans t en changeant le signe de v :

$$\alpha(-v) = \frac{\alpha(v)}{\Delta(v)} = \frac{\alpha(v)}{\alpha(v)^2\left(1-\frac{v^2}{c^2}\right)}.$$

Comme $\alpha(-v) = \alpha(v)$, on en déduit que

$$\alpha(v) = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}. \tag{43.23}$$

Maintenant qu'on a tout, on peut écrire les transformations de Lorentz. On met donc l'expression (43.23) dans les lois de transformations (43.21) :

$$\begin{aligned} t' &= \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\left(t - \frac{v}{c^2}x\right) & t &= \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\left(t' + \frac{v}{c^2}x'\right) \\ x' &= \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}(x - vt) & x &= \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}(vt' + x'). \end{aligned} \tag{43.24}$$

Tu remarqueras que $\Delta = 1$; si tu ne sais pas ce qu'est le déterminant d'une application linéaire, ça n'a pas d'importance. Mais si tu sais ce qu'est le déterminant d'une application linéaire, alors ce $\Delta = 1$ est crucial !

Affin d'avoir des équations un peu plus courtes, à partir de maintenant nous allons noter

$$\gamma(v) = \sqrt{1-\frac{v^2}{c^2}}.$$

43.4.3 Conditions d'existence

Comme tu vois une racine carrée et un dénominateur dans ces formules, tu dois te demander quelles sont les conditions d'existence. Étant donné que $v < c$, on a $v^2/c^2 \leqslant 1$ et en particulier, $v^2/c^2 = 1$ si et seulement si $v = c$.

Ce qui se trouve dans la racine carrée ne pose donc jamais de problèmes parce que ce n'est jamais négatif.

Le dénominateur est par contre plus problématique : quand $v = c$ il n'y a plus rien qui fonctionne. Quelle est la physique de ce problème ? Pour le comprendre, il faut se souvenir ce que représente v. Nous avons dit que v est la vitesse à laquelle Alice court. Ce que la condition d'existence nous enseigne, c'est que personne ne peut courir à la vitesse de la lumière. C'est une vitesse que l'on ne peut pas atteindre.

Dit en termes plus savants, on ne peut pas choisir un repère qui se déplace à la vitesse de la lumière.

La question qui se pose alors est « ah bon, on ne peut pas atteindre la vitesse de la lumière ! Et la lumière, comment elle fait ? ». Bonne question, merci de l'avoir posée. Hélas la réponse sort du cadre de ce cours.

Loi numéro 2.
Aucun objet ne peut atteindre la vitesse de la lumière.

Loi numéro 3.
Tu ne dois pas te demander pourquoi la lumière elle-même se déplace à la vitesse de la lumière malgré la loi numéro 2.

43.4.4 La notion d'intervalle

Un **événement** est quelque chose qui se passe à un endroit à un certain moment. C'est donc caractérisé par le moment et le lieu. Comme on travaille à une dimension, c'est un couple de réels (t, x).

Regardons un rayon de lumière. Un événement est le fait d'allumer une lampe de poche, et un autre est le fait que le lumière arrive sur l'objet qu'on éclaire. Appelons-les (t_1, x_1) et (t_2, x_2). Comme d'habitude, on note $\Delta t = t_2 - t_1$ et $\Delta x = x_2 - x_1$. Comme le rayon de lumière va à la vitesse c, on a $c = \Delta x/\Delta t$, ou encore

$$c^2\Delta t^2 - \Delta x^2 = 0.$$

Pour cette raison, on va dire que l'**intervalle** entre deux événements (t_1, x_1) et (t_2, x_2) vaut en général

$$s^2 = c^2(t_2 - t_1)^2 - (x_2 - x_1)^2. \tag{43.25}$$

Par invariance de la vitesse de la lumière, si un intervalle est nul pour un observateur, il sera nul pour tous les observateurs.

43.4.4.1 En mécanique newtonienne

Affin de voir un peu mieux l'enjeu de l'invariance de l'intervalle, regardons un exemple chiffré. Si par exemple je me déplace de 10 m en 5 s, mon intervalle mesuré par une personne extérieure est

$$c^2\Delta t^2 - \Delta x^2 = (300.000.000)^2 \cdot (5)^2 - (10)^2 = 2,25 \cdot 10^{18}\,\text{m}.$$

Si je fais le calcul pour moi, j'ai que $\Delta x' = 0$ parce que je ne me déplace pas, et $\Delta t' = 5$ parce que je me suis déplacé en 5 secondes. Le truc est que à côté de $(300.000.000)^2$, l'intervalle spatial Δx ne pèse pas grand chose. Ça ne change presque rien qu'il soit de 5 mètres ou de zéro. Ça ne change pas grand chose, mais ça change quand même ! Entre moi qui calcule ou une personne extérieure, l'intervalle change de 100 sur un nombre de la grandeur de 200.000.000.000.000.000.0000 !

Reprenons plus clairement le raisonnement. D'après la mécanique classique, l'intervalle mesuré par deux personnes est différent, mais très peu différent. Inutile de dire que du temps de Newton, on n'avait pas les moyens techniques de mesurer si cet intervalle est effectivement différent ou bien s'il est en réalité égal. C'est un peu comme si on te mettait un spot dans les yeux et qu'on te demandait si c'est un spot de 1000 W ou de 1001 W. Bonne chance pour le dire !

43.4.4.2 En mécanique relativiste

Maintenant qu'on a des moyens techniques nettement plus poussés que Newton, on a pu mesurer que l'intervalle est égal. L'intervalle est un invariant. Cela n'est pas un nouveau principe physique parce qu'il découle des transformations de Lorentz.

43.4.5 Le cône de lumière d'un point

Il est intéressant de dessiner dans le plan (t, x) l'ensemble des points atteints par le rayon lumineux. Le point (t, x) est atteint si $c^2t^2 - x^2 = 0$, ou encore si $x = \pm ct$. Cela forme deux droites dans le plan tracé par les coordonnées t et x. Ces deux droites forment ce qu'on appelle le **cône de lumière** du point $(0, 0)$.

43.4.6 Contraction des longueurs

Bob prend un morceau de bois qu'il mesure de longueur l et le dépose devant lui. À l'instant t (de Bob), les deux extrémités sont aux coordonnées $e_1 = (t, 0)$ et $e_2 = (t, l)$.

Afin de savoir quelle est la longueur de ce même morceau de bois pour Alice, il faut qu'elle mesure les deux extrémités en même temps (pour elle), et qu'elle fasse la différence. Comme Bob et Alice déclenchent leurs chronomètres en même temps, le plus simple est de faire la mesure à cet instant.

Pour Bob, c'est clair : les coordonnées des deux extrémités sont $e_1 = (0, 0)$ et $e_2 = (0, l)$. La longueur du bois est l. Pour savoir quelle est la longueur mesurée par Alice, on utilise les transformations de Lorentz qui donnent les coordonnées e'_1 et e'_2 relatives à Alice. On trouve $e'_1 = (0, 0)$ et

$$e'_2 = \left(\frac{-vl/c^2}{\gamma(v)}, \frac{l}{\gamma(v)} \right). \tag{43.26}$$

En d'autres termes, on a $x_1 = 0$ et $x_2 = l/\gamma(v)$, ce qui fait que la longueur observée par Alice est $l' = x_2 - x_1 = l/\gamma(v)$.

Eh bien ce résultat est faux. Si tu vois pourquoi sans lire la suite, tu es très fort.

Pour mesurer la longueur d'un objet, il faut mesurer la position des deux bouts *en même temps* puis faire la différence entre les deux. Effectivement, e_1 et e_2 sont en même temps pour Bob, et donc Bob peut mesurer la longueur de son bout de bois en faisant la différence $x_2 - x_1$. Mais comme le montre les coordonnées (43.26), les événements e'_1 et e'_2 ne se passent pas en même temps pour Alice ! Eh oui : $t'_1 = 0$ et $t'_2 = -vl/c^2\gamma(v)$; c'est pas la même chose.

Il faut donc trouver un événement qui pour Alice correspond à l'extrémité du bout de bois au temps $t' = 0$. Comme l'événement général qui correspond au bout du bois pour Bob est (t, l), l'événement général est pour Alice

$$t' = \frac{t - \frac{v}{c^2}l}{\gamma(v)} \qquad\qquad x' = \frac{l - vt}{\gamma(v)}. \tag{43.27}$$

Affin d'avoir $t' = 0$, il faut $t = vl/c^2$. En mettant cette valeur de t dans x', on trouve

$$x' = \frac{l - v\left(\frac{vl}{c^2}\right))}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{l\left(1 - \frac{v^2}{c^2}\right))}{\sqrt{1 - \frac{v^2}{c^2}}} = l\gamma(v).$$

Et là, c'est la bonne formule. Si un objet a une longueur l dans le référentiel où il est au repos, il aura une longueur

$$l' = l\sqrt{1 - \frac{v^2}{c^2}} \tag{43.28}$$

dans un référentiel qui se déplace à la vitesse v par rapport à l'objet.

43.4.7 Dilatation des intervalles de temps

Encore un petit effort et promit, je te donne une application concrète que tu connais des bizarreries de la relativité. Mais en attendant, regarde bien ta montre, tu ne va pas en croire tes yeux !

Reprenons Bob et Alice. On se rappelle que Bob et Alice avaient déclenchés leurs chronomètres en même temps quand Alice était passée devant Bob. Un peu plus tard, Alice regarde sa montre qui indique un temps t. Et elle se demande si Bob a aussi à ce moment une montre qui indique un temps t.

Ce serait dingue que non hein ! ? ! En effet, si je synchronise ma montre avec quelqu'un et que je pars faire un tour, ma montre ne sera pas tout d'un coup désynchronisée. Oui, mais Alice, elle cours presque à la vitesse de la lumière ... et à ces vitesses-là, on a déjà vu des choses incroyables. Calculons donc pour en avoir le cœur net.

Le fait qu'Alice regarde sa montre est un événement qui se passe pour Alice aux coordonnées $(t', 0)$ (le zéro c'est parce que par rapport à elle-même, Alice est toujours au repos). À quelles coordonnées (t, x) pour Bob correspond cet événement ?

L'équation de t en fonction de t' et x' dans les transformations de Lorentz (43.24) prise avec $x' = 0$ donnent

$$t = \frac{t'}{\gamma(v)}.$$

Et si, juste pour le plaisir, on faisait l'inverse ? Bob regarde sa montre, il voir un temps t et sa coordonnée spatiale est $x = 0$. À quel temps d'Alice cela correspond ? Mettons $x = 0$ dans la transformation de Lorentz de t' en fonction de t et x. Ce qu'on obtient c'est

$$t' = \frac{t}{\gamma(v)}.$$

N'est-ce pas génial ? C'est la même ! Évidemment, ça ne pouvait pas être autre chose : le principe de relativité demande qu'on ne puisse pas faire la différence entre Alice qui cours vers la droite avec Bob assis et Alice assise avec Bob qui cours vers la gauche. C'est exactement pour ça que dans une gare, quand le train d'à côté démarre, il t'arrive de croire que c'est ton train qui démarre : tu ne peux pas faire la différence, c'est un principe physique.

43.4.8 Invariance de l'intervalle

Dans deux secondes, je vais te montrer comment une utilisation intelligente des exponentielles permet de trouver un résultat très fort en relativité. Quoi ? Les exponentielles, les mêmes qu'au cours de math ? Eh oui : la même exponentielle que celle qu'on t'a introduit avec des populations de bactéries qui se multiplient, cette même exponentielle qui la la miraculeuse propriété d'être égale à sa propre dérivée.

Mais n'anticipons pas.

Nous avons déjà signalé que si la quantité $\Delta s^2 = c^2\Delta t^2 - \Delta x^2$ était nulle pour un observateur, alors elle était nulle pour tous les observateurs. Supposons deux événements A et B observés par Alice et Bob. Bob les note aux coordonnées (t_a, x_a), et (t_b, x_b) tandis qu'Alice les note en (t'_a, x'_a) et (t'_b, x'_b).

L'intervalle entre les deux événements mesuré par Bob sera

$$s^2 = c^2(t_b - t_a)^2 - (x_b - x_a)^2,$$

tandis que ce même intervalle mesuré par Alice sera

$$s'^2 = c^2(t'_b - t'_a)^2 - (x'_b - x'_a)^2.$$

On peut bien entendu remplacer dans la première équation les t_a, t_b, x_a et x_b par leurs valeurs en termes de t'_a, t'_b, x'_a et x'_b données par les transformations de Lorentz. Tu paries que les trois quart des termes dans le calcul se simplifient et qu'il restera exactement s'^2 ? Je te dis que oui, et je te conseille de me croire sur parole, sinon tu vas devoir lire le calcul suivant :

$$\begin{aligned} s^2 = c^2(t_b - t_a)^2 - (x_b - x_a)^2 = c^2 &\left(\frac{1}{\gamma(v)}(t'_b + \frac{v}{c^2}x'_b) - \frac{1}{\gamma(v)}(t'_a + \frac{v}{c^2}x'_a)\right)^2 \\ &- \left(\frac{1}{\gamma(v)}(vt'_b + x'_b) - \frac{1}{\gamma(v)}(vt'_a + x'_a)\right)^2. \end{aligned}$$

Jusqu'ici, on n'a fait que remplacer les choses par leurs valeurs données par les transformations de Lorentz. Maintenant on regroupe à l'intérieur de chaque parenthèse les termes de façon à faire

apparaître $\Delta x'$ et $\Delta t'$:

$$\begin{aligned} s^2 &= \frac{c^2}{\gamma(v)^2}\left((t'_b - t'_a) + \frac{v}{c^2}(x'_b - x'_a)\right)^2 \\ &\quad - \frac{1}{\gamma(v)^2}\big((x'_b - x'_a) + v(t'_b - t'_a)\big)^2 \\ &= \frac{c^2}{\gamma(v)^2}\left((\Delta t')^2 + 2\frac{v}{c^2}\Delta t'\Delta x' + \frac{v^2}{c^4}(\Delta x')^2\right) \\ &\quad - \frac{1}{\gamma(v)^2}\Big((\Delta x')^2 + 2v\Delta x'\Delta t' + v^2(\Delta t')^2\Big). \end{aligned}$$

Là, on a utilisé le produit remarquable $(a+b)^2 = a^2+2ab+b^2$, et on a systématiquement renommé tous les intervalles avec la notation Δ pour être plus compact. Maintenant, on va regrouper tous les termes contentant $(\Delta t')^2$ ensemble, tous ceux qui contiennent $\Delta t'\Delta x'$ ensemble et ceux qui contiennent $(\Delta x')^2$ ensemble. Autre manière de le dire, on met les Δ en évidence comme on peut. On trouve ceci :

$$\begin{aligned} (\Delta t')^2\left(\frac{c^2}{\gamma(v)^2} - \frac{v^2}{\gamma(v)^2}\right) + \Delta t'\Delta x'\left(\frac{2vc^2}{\gamma(v)^2c^2} - \frac{2v}{\gamma(v)^2}\right) \\ + (\Delta x')^2\left(\frac{c^2v^2}{c^4\gamma(v)^2} - \frac{1}{\gamma(v)^2}\right). \end{aligned}$$

À partir de là, je te laisse vérifier (en utilisant le fait que $\gamma(v)^2 = 1 - v^2/c^2$) que les coefficients se simplifient beaucoup et valent finalement respectivement c^2, 0 et -1 comme il se doit. Avec tout ça, nous avons montré le résultat très important suivant :

> L'intervalle entre deux événements est invariant sous les changements de repères d'inertie, c'est à dire que la valeur mesurée par n'importe qui qui se déplace en MRU sera toujours la même.

Pourquoi cela est tellement important ? À cause de Pythagore et d'une petite démonstration à coups d'exponentielles[5].

43.4.8.1 Rappel de trigonométrie hyperbolique

Les fonctions de trigonométrie hyperboliques sont :

$$\cosh(x) = \frac{e^x + e^{-x}}{2} \qquad \sinh = \frac{e^x - e^{-x}}{2}. \tag{43.29}$$

Elles ont pas mal de propriétés en commun avec les sinus cosinus et normaux. D'abord, leurs dérivées sont faciles à calculer :

$$\begin{aligned} \cosh'(x) &= \sinh(x) \\ \sinh'(x) &= \cosh(x) \end{aligned}$$

où tu noteras qu'il n'y a pas de signe moins qui apparaît, contrairement au cas de la trigonométrie normale. Une autre propriété qui ressemble fort à une propriété de la trigonométrie est :

Proposition 43.2.
Pour tout $x \in \mathbb{R}$,

$$\cosh^2(x) - \sinh^2(x) = 1 \tag{43.30}$$

avec un signe moins comme différence avec la trigonométrie.

5. oui oui tout ton cours de math va finir par y passer.

Démonstration. La preuve revient simplement à calculer en utilisant le produit remarquable de $(a+b)^2$. D'abord, on a :

$$\cosh^2(x) = \frac{1}{4}\big(e^x + e^{-x}\big)^2 = \frac{1}{4}\big(e^{2x} + 2e^x e^{-x} + e^{-2x}\big) = \frac{1}{4}(e^{2x} + 2 + e^{-2x})$$

où l'on a utilisé le fait que $(e^x)^2 = e^{2x}$ et que $e^x e^{-x} = 1$. Il te reste à faire la même chose pour $\sinh^2(x)$, la réponse est :

$$\sinh^2(x) = \frac{1}{4}\big(e^{2x} - 2 + e^{-2x}\big).$$

En faisant la différence entre les deux, il reste 1. □

Une propriété qui est par contre très différente entre la trigonométrie plane et la trigonométrie hyperbolique, c'est la périodicité. Les fonctions usuelles cos et sin sont périodiques. Pas les fonctions hyperboliques.

Proposition 43.3.
La fonction $\sinh\colon \mathbb{R} \to \mathbb{R}$ *est bijective.*

Démonstration. Il faut démontrer que sinus hyperbolique est injective et surjective. Calculons d'abord les limites. Comme tu sais que $\lim_{x\to\infty} e^x = \infty$ et $\lim_{x\to-\infty} e^x = 0$, tu vois facilement que

$$\lim_{x\to-\infty} \sinh(x) = -\infty \qquad \lim_{x\to\infty} \sinh(x) = \infty. \tag{43.31}$$

Par ailleurs, la fonction sinus hyperbolique est continue et respecte donc le théorème de la valeur intermédiaire[6] 14.41. Soit $y \in \mathbb{R}$. Voyons s'il existe un $x \in \mathbb{R}$ tel que $\sinh(x) = y$. Les deux limites indiquent qu'il existe $x_1 \in \mathbb{R}$ tel que $\sinh(x_1) < y$ et $x_2 \in \mathbb{R}$ tel que $\sinh(x_2) > y$. Le théorème de la valeur intermédiaire conclu qu'il existe un x entre x_1 et x_2 tel que $\sinh(x) = y$. Cela prouve la surjectivité.

Pour l'injectivité, on va utiliser le théorème de Rolle et une petite preuve par l'absurde. Supposons que $\sinh(x_1) = \sinh(x_2)$ avec $x_1 \neq x_2$. Dans ce cas, le théorème de Rolle nous dit qu'il existe un x entre x_1 et x_2 tel que $\sinh'(x) = 0$. La dérivée de sinus hyperbolique étant cosinus hyperbolique, il faut se demander il existe un x tel que $\cosh(x) = 0$. Étant donné que $e^x > 0$ pour tout x, en fait le cosinus hyperbolique ne s'annule jamais. □

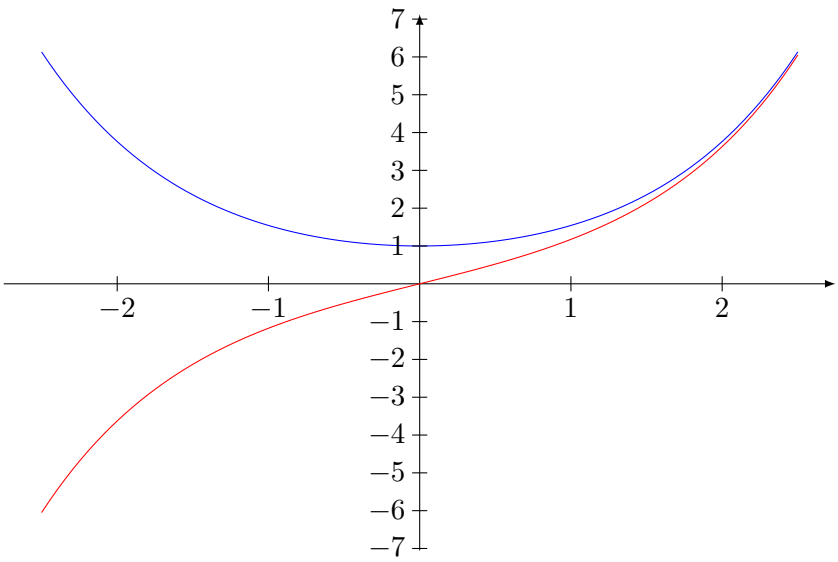


FIGURE 43.1 – En rouge, la fonction $x \mapsto \sinh(x)$ et en bleu, la fonction $x \mapsto \cosh(x)$.

6. Je t'avais dit que tout tons cours de math allait y passer hein.

Un très bon exercice serait de faire un étude complète des fonctions cosinus et sinus hyperbolique. Leur graphes sont donnés à la figure 43.1

Un corollaire de la surjectivité de sinh sur $\mathbb{R}$ est que si je prends n'importe quel deux nombres dont la différence des carrés vaut 1, alors ces carrés sont représentables avec des fonctions hyperboliques :

$$\forall x, y \in \mathbb{R} \text{ tels que } x^2 - y^2 = 1, \exists \xi \in \mathbb{R} \text{ tel que } x^2 = \cosh(\xi) \text{ et } y^2 = \sinh(\xi).$$

La **tangente hyperbolique** est définie par

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)}. \tag{43.32}$$

Un bon exercice est de prouver les deux relations suivantes :

$$\sinh(x) = \frac{\tanh(x)}{\sqrt{1 - \tanh^2(x)}} \qquad \cosh(x) = \frac{1}{\sqrt{1 - \tanh^2(x)}}. \tag{43.33}$$

43.4.8.2 Les transformations de Lorentz (bis)

Nous avons prouvé qu'en relativité, l'intervalle est un invariant. Pour cela, nous avons utilisé les transformations de Lorentz démontrées à partir de l'hypothèse d'invariance de la vitesse de la lumière. Eh bien, oublions un instant que la vitesse de la lumière soit invariante, et posons à la place comme hypothèse que l'intervalle soit invariant. C'est à dire que si Bob mesure un événement aux coordonnées (t, x) et Alice en (t', x'), alors $c^2t^2 - x^2 = c^2(t')^2 - (x')^2$.

Théorème 43.4.
Les transformations de Lorentz sont les seules qui laissent l'intervalle invariant.

Démonstration. Toute la partie comme quoi les transformations doivent êtres linéaires reste parce que cette partie ne demandait pas l'invariance de la vitesse de la lumière.

Nous cherchons donc les transformations entre Alice et Bob sous la forme

$$\begin{aligned} t' &= \alpha t + \beta x \\ x' &= \gamma t + \delta x \end{aligned}$$

telles que $c^2(t')^2 - (x')^2 = c^2t^2 - x^2$. Lorsque Alice passe devant Bob, ils déclenchent tout deux leurs chronomètre et leurs axes. C'est à dire que si à ce moment un événement se trouve à droite pour Alice, il est aussi à droite pour Bob. On doit donc avoir, quand $t = t' = 0$, que $x > 0$ implique $x' > 0$. Cela donne la contrainte que $\delta > 0$. D'autre part, comme leurs chronomètres vont dans le même sens (ils choisissent tout les deux de *compter* le temps et non *décompter*), on a $\alpha > 0$.

En développant l'expression de $(s')^2$ en termes de t et x, on trouve la condition d'invariance de l'intervalle sous la forme :

$$c^2(\alpha^2t^2 + 2\alpha\beta tx + \beta^2x^2) - (\gamma^2t^2 + 2\gamma\delta tx + \delta^2x^2) = c^2t^2 - x^2, \tag{43.34}$$

qui doit être valable pour tout t et pour tout x. En $t = 0$ on trouve la condition

$$\delta^2 - c^2\beta^2 = 1. \tag{43.35}$$

Cela implique qu'il existe un $\xi \in \mathbb{R}$ tel que $\delta^2 = \cosh^2(\xi)$ et $c^2\beta^2 = \sinh(\xi)$. La première équation donne $\delta = \cosh(\xi)$ (il faut rejeter $\delta = -\cosh(\xi)$ parce qu'on a demandé que $\delta > 0$). Pour la seconde, on trouve $c\beta = \sinh(\xi)$ où l'on peut oublier la possibilité $c\beta = -\sinh(\xi)$ parce que cela revient juste à renommer $\xi \to -\xi$ (la fonction sinus hyperbolique est impaire). Bref, il existe un ξ tel que

$$\begin{aligned} \delta &= \cosh(\xi) \\ \beta &= \frac{\sinh(\xi)}{c} \end{aligned} \tag{43.36}$$

En mettant maintenant $x = 0$ dans la condition (43.34), on trouve la condition

$$\alpha^2 - \frac{\gamma^2}{c^2} = 1.$$

Pour les mêmes raisons qu'avant, il existe un $\eta \in \mathbb{R}$ tel que

$$\begin{aligned} \alpha &= \cosh(\eta) \\ \gamma &= c\sinh(\eta). \end{aligned} \tag{43.37}$$

Rien qu'en regardant deux cas très particuliers, on a déjà bien avancé, non ? Remettons maintenant les valeurs (43.36) et (43.37) dans la condition (43.34). En utilisant l'identité $\cosh^2(x) - \sinh^2(x) = 1$, et en séparant les termes en t^2, x^2 et tx pour satisfaire la condition, il faut

$$\cosh(\eta)\sinh(\xi) = \sinh(\eta)\cosh(\xi) \tag{43.38}$$

parce que les termes en t^2 et x^2 donnent exactement $c^2t^2 - x^2$ et qu'il faut que le terme en tx s'annule. Mettons la condition (43.38) au carré, et substituons $\cosh^2(\eta) = 1 + \sinh^2(\eta)$ et $\cosh^2(\xi) = 1 + \sinh^2(\xi)$, il reste

$$\sinh^2\xi = \sinh^2\eta,$$

ce qui signifie $\sinh\xi = \pm\sinh\eta$, ou encore $\xi = \pm\eta$. On voit que $\xi = -\eta$ ne fonctionne pas dans (43.38), donc on reste avec $\xi = \eta$ et les transformations prennent la forme

$$\begin{aligned} t' &= \cosh(\xi)t + \frac{\sinh(\xi)}{c}x \\ x' &= c\sinh(\xi)t + \cosh(\xi)x. \end{aligned} \tag{43.39}$$

Ce que nous avons prouvé, c'est qu'il existe un $\xi \in \mathbb{R}$ tel que les transformations entre Alice et Bob aient cette forme. Il faut trouver ce que vaut ξ en fonction de la vitesse v à laquelle Alice court.

Pour ce faire, étudions le mouvement d'Alice. Bob la voit aux coordonnées (t, vt), ce qui correspond à

$$x' = c\sinh(\xi)t + \cosh(\xi)vt$$

pour Alice. Mais ces coordonnées sont celles de Alice elle-même, donc $x' = 0$, ce qui donne[7] $vt = -c\sinh(\xi)t/\cosh(\xi)$, ou encore

$$\tanh(\xi) = -\frac{v}{c} \tag{43.40}$$

En utilisant les relations (43.33), on trouve

$$\cosh(\xi) = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad \sinh(\xi) = \frac{-v/c}{\sqrt{1 - \frac{v^2}{c^2}}}. \tag{43.41}$$

En remettant ces valeurs dans les transformations (43.39), on trouve

$$t' = \frac{t - \frac{v}{c^2}x}{\gamma(v)} \tag{43.42}$$

$$x' = \frac{x - vt}{\gamma(v)}, \tag{43.43}$$

exactement les transformations de Lorentz !

□

Ce résultat est important pour une raison assez simple : maintenant, la théorie de la relativité est indépendante de toute considérations sur la lumière. En effet, ce que nous venons de prouver, c'est que s'il existe une vitesse c telle que $c^2t^2 - x^2 = c^2(t')^2 - (x')^2$, alors (t, x) et (t', x') sont liés par les transformations de Lorentz.

7. Conditions d'existence : $\cosh(\xi) \neq 0$; heureusement, nous avons vu que le cosinus hyperbolique ne s'annule jamais.

43.4.9 Vitesse limite

Affin de nous passer de l'hypothèse d'invariance de la vitesse de la lumière, nous avons prouvé que l'hypothèse d'invariance de l'intervalle était suffisante. Mais il faut avouer que cette hypothèse n'est pas très intuitive. Nous allons montrer maintenant que l'existence d'une vitesse limite est une troisième hypothèse qui peut être utilisée comme alternative aux deux premières.

43.5 Applications

Une première application sympa est le logiciel [8] *lightspeed*. Si tu es sous Ubuntu-Linux, installe juste le paquet nommé lightspeed, et régales-toi ! Tu verras c'est marrant. Si tu utilise des fenêtres, je laisse faire l'adage « Windows c'est facile ».

43.5.1 Le GPS

Pour qu'un système GPS puisse te localiser, en gros, il t'envoie un signal, tu lui réponds et il mesure le temps qu'il a fallu à la lumière pour faire l'aller-retour. Déjà, tu remarques que cela n'est possible que grâce au fait que la vitesse de la lumière soit finie. Sinon, le GPS ne fonctionnerait pas. Mais il y a mieux.

Comme pour te localiser il faut plusieurs satellites en plus de ton appareil, il faut que les horloges internes de tout ce petit monde soient bien synchronisées, sinon pour mesurer des intervalles de temps et calculer des distances, c'est mal parti. Eh mais un satellite, ça bouge assez vite (surtout que les mesures doivent être très précises), et en plus ça ne fait même pas un MRU, vu que ça tourne en rond. Comme tu vois tout le travail qu'il a fallu faire pour trouver les transformations de Lorentz d'un MRU, tu t'imagines le travail pour un mouvement circulaire ! Eh bien ce travail a été fait, et le résultat est que si on en tient pas compte, les contractions temporelles liées à la relativité sont suffisamment grandes pour complètement dérégler le GPS.

43.5.2 Les ondes électromagnétiques

Tu te souviens qu'au début du chapitre, nous avons dit que le problème qui a amené la relativité était la propagation du champ électrique. Maintenant que nous avons déjà vu une partie des conséquences du problèmes, il est temps de se rendre compte que les champs électriques et magnétiques sont les objets les plus soumis aux bizarreries relativistes du monde : elles se propagent à la vitesse de la lumière. Regarde un coup autour de toi ; tout ce qui est champ électromagnétique a besoin de la relativité pour être bien compris : GSM, lumière, four à micro-onde, radio, wifi, fibre optique, ...

Si un jour un ingénieur te dit qu'il n'y a pas besoin de connaître la relativité pour inventer la radio (c'est vrai : la radio a été inventée avant la relativité), ni pour construire une fibre optique, dis lui en pensant à moi qu'il utilise tout le temps les équations de Maxwell [9], et que ces équations sont relativistes.

Bref, soit convaincu que tu vis dans un monde relativiste et que les transformations de Lorentz te suivent à chacun de tes pas.

43.6 Mécanique relativiste

Cela est bien beau, mais la dilatation du temps, et les contractions de longueurs doivent bien avoir des répercussions sur la cinématique et la dynamique des objets. Est-ce que le théorème de l'énergie cinétique est encore valable ? est-ce que les lois de Newton tiennent encore la route ?

8. jeu de mot sur « application » ! ah ah !
9. C'est sous ce nom là qu'on nome l'ensemble des équations de l'électromagnétisme comme la loi de l'induction.

43.6.1 Des problèmes, toujours des problèmes

Attardons-nous un peu pour faire quelques commentaires sur cette citation du chevalier pégase dans les chevaliers du zodiaque :

> Ses coups vont à la vitesse de la lumière et pourtant je les vois distinctement arriver.

Est-ce possible ? Nous avons vu qu'il y avait des dénominateurs qui s'annulent quand des objets se déplacent plus vite que la lumière ; or pour voir venir un rayon de lumière qui vient vers soi, il faudrait que le rayon émette de la lumière devant elle. Ça semble un peu mal parti pour respecter les lois de la relativité, non ?

Cela pose en tout cas une question qu'il faudra résoudre. On *entend* venir une ambulance parce qu'elle émet du son qui avance plus vite qu'elle. Pas de problèmes avec ça. Mais quid de la *voir* venir ?

On peut voir venir un tram parce qu'il émet de la lumière ; cette lumière allant plus vite que le tram, elle arrive à nos yeux avant le tram lui-même. Cela est très bien. Mettons que le tram avance à 50 km/h ; pour le conducteur, la lumière de son phare avant avance devant lui à la vitesse c. Par conséquent pour un observateur au sol, cette même lumière devrait avancer à la vitesse $c+50$. Encore une fois, on a un problème d'invariance de la vitesse de la lumière ; mais comme c'est de la lumière, on est habitué à ce que des trucs bizarres arrivent. On ne sera pas étonné que $c+50=c$ d'une manière ou d'une autre[10]. Pire. Si un vaisseau spatial avance à la vitesse 200000 km/s et qu'il envoie en reconnaissance un vaisseau devant lui à la vitesse de 150000 km/s, le vaisseau de reconnaissance ira à la vitesse 150000 km/s par rapport au vaisseau principal. Et par rapport au sol, il ira à la vitesse $150000+200000=350000$ km/s, ce qui est impossible. Il faudra trouver quelque chose pour que ça se passe bien.

Un autre problème maintenant.

Prenons une masse m que l'on soumet à une force constante F. Par la loi de Newton, $a=F/m$ est constante et la vitesse après un temps t vaut $v=Ft/m$. Pas de bol, ça devient plus grand que la vitesse de la lumière à partir du temps $t=cm/F$. Ça est un problème hein ? Il faut trouver un truc pour qu'avec une force constante, l'accélération diminue.

43.6.2 Loi d'addition des vitesses

Si Bob observe un objet se déplacer à la vitesse V, alors Alice devrait l'observer bouger à la vitesse $V-v$. Tout comme si une vache voit passer un train à 90 km/h, alors le vélo qui avance à 25 km/h le voit passer à 65 km/h.

Maintenant, tu es habitué à ce que rien ne se passe comme d'habitude, donc tu te doutes bien qu'en réalité la bonne formule ne va pas être $V-v$.

Bob observe l'objet aux coordonnées (t,Vt), ce qui fait pour Alice :

$$\left(\frac{t-\frac{v}{c^2}Vt}{\gamma(v)},\frac{Vt-vt}{\gamma(v)}\right).$$

En divisant le x' d'Alice par le t' d'Alice, on trouve la vitesse mesurée par Alice :

$$V'=\frac{(V-v)t}{\gamma(v)}\frac{\gamma(v)}{t\left(1-\frac{vV}{c^2}\right)}=\frac{V-v}{1-\frac{vV}{c^2}}.$$

La loi de transformation des vitesses relativiste est donc

$$V'=\frac{V-v}{1-\frac{vV}{c^2}}. \tag{43.44}$$

Qu'en est-il de notre $c+50=c$? Disons que Bob lance un bisou à Alice pendant qu'elle arrive vers lui. Le bisou arrive à la vitesse de la lumière (càd $V=c$) tandis que Alice s'approche de Bob

10. et je ne te cache pas que c'est ce qui va arriver.

à la vitesse 50 m/s (càd $v = 50$). Donc la vitesse à laquelle Alice devrait voir arriver le bisou est bien $c + 50$. En utilisant la formule d'addition relativiste des vitesses (43.44), nous trouvons

$$V' = \frac{c - 50}{1 - \frac{50c}{c^2}} = \frac{c - 50}{1 - \frac{50}{c}} = \frac{c(c - 50)}{c - 50} = c.$$

Donc effectivement en relativité quand on additionne des vitesses il faut penser à la règle du « $c + 50 = c$ ».

43.6.3 L'action d'une force

L'équation fondamentale de la mécanique classique est

$$F = ma.$$

Or tu n'es pas sans savoir que l'accélération est la dérivée seconde de la position par rapport au temps. Nous noterions donc $F = mx''(t)$. Le problème est évidemment que si F est constante, on trouve $v = Ft/m$ qui dépasse toujours la vitesse c quand t est assez grand. Il faudra donc modifier la loi $F = ma$. Pour cela, posons-nous des questions sur la dérivée $x'(t)$. On dérive par rapport au temps ; oui mais nous avons vu que le temps n'est pas le même pour tout le monde. Introduisons donc la notation

$$v = \frac{dx}{dt} \tag{43.45}$$

qui ne signifie rien d'autre que nous dérivons x par rapport à t et non par rapport au temps t' de quelqu'un d'autre. Dans le cadre de la relativité, ce que signifie l'équation (43.45) est que v est la dérivée de x par rapport à t. Dans le cas où x et t sont les coordonnées de la position d'Alice mesurées par Bob, cela signifie qu'on dérive la position *mesurée par Bob* par rapport au temps *mesuré par Bob*.

Ce que dit la relativité est que cette quantité v ne peut pas varier proportionnellement à la force sous peine de dépasser la vitesse de la lumière. La subtilité est de modifier la loi de Newton en disant que la quantité qui varie sous l'action d'une force n'est plus $dx/dt = v$, mais

$$\frac{dx}{dt'} = \frac{v}{\sqrt{1 - \frac{v^2}{c^2}}},$$

c'est à dire la dérivée de la position *mesurée par Bob* par rapport au temps *mesuré par Alice* ! La loi de Newton $v = Ft/m$ devient donc

$$\frac{v}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{Ft}{m}. \tag{43.46}$$

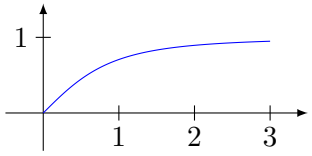
Est-ce que cela résous le problème ? Pour le savoir, regardons la vitesse acquise par le mobile de masse m soumit à la force F pendant un temps t Il faut résoudre l'équation (43.46) par rapport à v et voir si cela reste bien toujours inférieur à c. On commence par mettre la racine à droite et à élever toute l'équation au carré :

$$\begin{aligned} v^2 &= \frac{F^2t^2}{m^2}\left(1 - \frac{v^2}{c^2}\right) \\ v^2\left(1 + \frac{F^2t^2}{c^2m^2}\right) &= \frac{F^2t^2}{m^2} \\ v &= \frac{\sqrt{F^2t^2/m^2}}{\sqrt{1 + \frac{F^2t^2}{c^2m^2}}}, \end{aligned}$$

et donc finalement

$$v(t) = \frac{Ft}{m\sqrt{1 + \frac{F^2t^2}{c^2m^2}}}. \tag{43.47}$$

Tu dois remarquer que si F et t ne sont pas trop grands, l'expression F^2t^2/c^2m^2 est minuscule parce que c est énorme. Si on fait l'approximation $F^2t^2/c^2m^2 = 0$ dans cette expression, on retrouve $v = Ft/m$. Cela montre qu'à moins de faire des expérience avec de très grandes forces pendant énormément de temps, on ne peut pas voir la différence entre la mécanique de Newton et la mécanique relativiste.



Sur le graphe suivant, la vitesse en fonction du temps lorsqu'une particule de masse $m = 1$ est soumise à une force constante. Pour les besoins du graphique, nous avons mit à 1 la vitesse c. Tu vois que quand la vitesse n'est pas très grande, le graphiques est presque celui d'une droite ; et à partir d'un certain moment, la courbe s'infléchit pour tendre vers 1 sans l'atteindre.

Remarque que si on maintient une accélération constante égale à celle de la gravité terrestre pendant deux heures, on arrive déjà sur la Lune, à une vitesse de 75 km/s, c'est à dire encore rien par rapport à la vitesse de la lumière ! Cela pour te dire que la formule (43.47) a l'air d'être très différente de la formule classique $v = Ft/m$, mais en réalité tant qu'on n'atteint pas des forces énormes, elle ressemble très fort.

Vérifions maintenant que la formule (43.47) n'est pas en contradiction avec l'impossibilité de dépasser la vitesse de la lumière. Pour cela, regardons ce qu'il se passe si on applique une force constante F sur un objet de masse m pendant un temps très long. C'est à dire : calculons la limite

$$\lim_{t\to\infty} v(t).$$

Tu vois tout de suite qu'on est sur un cas $\frac{\infty}{\infty}$, ce qui t'oblige à utiliser la règle de l'Hospital. On peut cependant un peu réfléchir et deviner la réponse sans passer par des math trop compliquées.

En effet, quand t est vraiment énorme, l'expression $\frac{F^2t^2}{m^2c^2}$ devient très grande, et le 1 qui se trouve à côté ne vaut plus grand chose, on peut le négliger.

$$\begin{aligned}\lim_{t\to\infty} \frac{Ft}{m\sqrt{1+\frac{F^2t^2}{c^2m^2}}} &= \lim_{t\to\infty} \frac{Ft}{m\sqrt{\frac{F^2t^2}{m^2c^2}}} \\ &= \lim_{v\to\infty} \frac{Ft}{m\frac{Ft}{cm}} \\ &= c.\end{aligned} \tag{43.48}$$

Tout est bien : on arrive au maximum à la vitesse de la lumière, mais il faut un temps infini pour y parvenir. Conclusion : il n'est pas possible d'accélérer un objet jusqu'à atteindre la vitesse de la lumière.

43.6.4 Équivalence entre la masse et l'énergie

Le moment est venu de montrer ce que signifie la fameuse formule $E = mc^2$.

43.7 Principe de correspondance

Nous ne sommes pas parvenu à démontrer la formule (43.46) de la mécanique relativiste qui montre comme un objet accélère sous l'effet d'une force constante. Nous avons juste montré qu'il fallait modifier la loi $v = Ft/m$ et nous avons prit la première modification qui nous soit tombée sous la main, à savoir qu'il faut dériver la position par rapport au temps de l'objet qu'on observe plutôt que par rapport au temps de l'observateur.

En fait, il est possible de prouver rigoureusement [11] la formule

$$\frac{Ft}{m} = \frac{\alpha v}{\sqrt{1-\frac{v^2}{c^2}}}.$$

Mais il n'y a pas moyen de trouver la valeur de la constante α. Tout ce qu'il y a moyen de trouver avec l'hypothèse de l'invariance de la vitesse de la lumière est l'existence d'une constante telle que cette formule soit vraie.

Affin de fixer la constante α, il faut faire intervenir un principe physique supplémentaire, le **principe de correspondance**

Loi numéro 4.
Lorsque la vitesse d'une particule est faible, les équations doivent être en première approximation les mêmes que celles de la mécanique classique.

Que signifie *en première approximation* ? Tu sais qu'une fonction $x \mapsto f(x)$ peut être approximée (pour des petits x) par la formule

$$f(x) \simeq f(0) + xf'(0).$$

Nous voudrions donc que Ft/m soit en première approximation égal à v. Nous devons étudier la fonction

$$f(v) = \frac{\alpha v}{\sqrt{1-\frac{v^2}{c^2}}}.$$

Voir ce que vaut cette fonction en première approximation lorsque v est petit est un exercice de dérivation. En utilisant la règle de dérivation des fractions, on trouve que

$$f'(v) = \frac{\alpha}{\sqrt{1-\frac{v^2}{c^2}}} + \frac{\alpha v^2}{c^2\left(1-\frac{v^2}{c^2}\right)^{3/2}},$$

et donc que $f'(0) = \alpha$. Bien entendu, $f(0) = 0$. En première approximation, nous trouvons donc

$$f(v) \simeq \alpha v \tag{43.49}$$

qui doit être égal à la quantité non relativiste v. Nous en déduisons qu'il faut fixer $\alpha = 1$, et on tombe sur la formule relativiste proposée plus haut

$$\frac{Ft}{m} = \frac{v}{\sqrt{1-\frac{v^2}{c^2}}}$$

L'utilisation cruciale du principe de correspondance a une répercussion énorme sur notre vision de la physique. En effet, la relativité d'Einstein ne parvient pas à *remplacer* la mécanique de Newton. On a besoin d'invoquer la mécanique de Newton pour fixer la théorie. On peut écrire l'axiome suivant :

$$\lim_{v\to 0} \text{Einstein} = \text{Newton}. \tag{43.50}$$

Cela n'est pas une propriété de la théorie d'Einstein, mais un de ses axiomes !

La relativité ne fait donc pas table rase des principes physiques de la mécanique newtonienne : elle les complète et les contient.

11. Mais il n'existe pas de démonstrations simples à ma connaissance.

Chapitre 44

Exemples avec Sage

Ce chapitre est un foure-tout de choses que l'on peut faire avec Sage.

44.0.1 Graphiques

Pour afficher le graphe d'une fonction, vous pouvez faire

```
+--------------------------------------------------------------------+
| SageMath version 8.1, Release Date: 2017-12-07                     |
| Type "notebook()" for the browser-based notebook interface.        |
| Type "help()" for help.                                            |
+--------------------------------------------------------------------+
sage: plot(cos(x),0,5)
Launched png viewer for Graphics object consisting of 1 graphics primitive
sage: f(x)=sin(x)
sage: f.plot(-pi,pi)
Launched png viewer for Graphics object consisting of 1 graphics primitive
```

Un programme externe se lance automatiquement pour afficher le graphique que vous avez demandé.

Il se peut qu'aucun programme ne se lance et vous ayez, au lieu de `Launched png viewer for Graphics object ...` uniquement `Created graphics object ...`. Disons pour faire court que Sage a produit un `png` et qu'il ne sait pas quel programme externe utiliser pour l'afficher.

La solution est à l'adresse http://doc.sagemath.org/html/en/reference/misc/sage/misc/viewer.html

44.0.2 Autres

Dans le but d'automatiser certaines tâches, j'ai écrit ce module, nomé `outilsINGE.sage`, dans le cadre d'un cours de première année donné à des ingénieurs. Certaines des fonctions définies ici sont utilisée dans les exemples qui suivent.

```
1  # -*- coding: utf8 -*-
2  from sage.all import *
3
4  """
5  This module provides _pragmatic_ tools for solving exercise for
6  a first year in general mathematics.
7  """
8
9  # TODO : trouver une bonne traduction pour "point de selle."
10
```

```
def automatedVar(symbol,n):
  """ If symbol = "x" and n=4, return the string 'x1,x2,x3,x4'
     """
  s = ",".join([ symbol+str(i) for i in range(1,n+1)])
  return s

class SolveLinearSystem(object):
  """
  Solve Ax=v and print it in a nice way

  Example :

  A=matrix([ [1,-2,3,-2,0],[3,-7,-2,4,0],[4,3,5,2,0] ])
  v=vector((0,0,0,0,0))
  print SolveLinearSystem(A,v)
  """
  def __init__(self,A,v):
    self.matrix = A
    self.vector = v
    self.nvars = A.ncols()
    s = automatedVar("x",self.nvars)
    self.xx=var(s)
  def equations(self):
    """Return the equations corresponding to the
                self.matrix and self.vector as a list of equations
                   """
    X=matrix( [self.xx[i] for i in range(0,self.nvars) ]).
       transpose()
    eqs=[]
    for i in range(0,self.matrix.nrows()):
      exp = (self.matrix*X)[i][0]==self.vector[i]
      eqs.append(exp)
    return eqs
  def solutions(self):
    return solve(self.equations(),self.xx)
  def latex(self):
    """Return the LaTeX's code of the system."""
    a=[]
    a.append(r"""

        \item
          $
          \left\{
          \begin{array}{ll}
    """)
    for eq in self.equations():
      a.append("          "+str(eq).replace("x","x_").replace("*"
         ,"").replace("==","=")+"\\\\ \n")
    a.append(r"""
          \end{array}
          \right.
          $
          """)
```

```
60      return "".join(a)
61    def __str__(self):
62      a = []
63      a.append("The given matrix corresponds to the system")
64      for eq in self.equations():
65        a.append(str(eq))
66      a.append("And the solutions are")
67      a.append(str(self.solutions()))
68      return "\n".join(a)
69
70  def QuadraticMap(A,v):
71    """
72    Return the result of the quadratic form associated
73          with A applied on the vector v, that is the number
74    A_ij v^iv^j
75    using the summation convention.
76    """
77    n = A.nrows()
78    if not A.is_symmetric():
79      print "Warning : Given matrix is not symmetric"
80    if not A.is_square():
81      raise TypeError,"Error : The matrix A is not square"
82    if not v.degree()==n :
83      raise TypeError,"The size do not agree"
84    return sum([ A[i,j]*v[i]*v[j] for i in range(n) for j in range(↩
       n)  ]).simplify_full()
85
86  class SymmetricMatrix(object):
87    """
88    Provide informations about the matrix A assuming it is symmetric↩
         .
89    """
90    def __init__(self,A):
91      if not A.is_square():
92        print "Error : A symmetric matrix must be square"
93        raise TypeError
94      self.matrix = A
95      self.degree = A.nrows()
96      self.matrix.set_immutable()
97    def primary_principal_submatrix(self,n):
98      """
99      Return the primary principal submatrix of order n, that is the↩
          matrix obtained
100                 by removing the n last lines and columns from self↩
                       .
101     """
102     taille=self.degree-n
103     v=[]
104     for i in range(0,taille):
105       v.append(self.matrix[i][0:taille])
106     return matrix(v)
107   def principal_minors(self):
108     """
```

```
109     Return the list of principal minors. The principal minor of order k is
110                 the determinant of the primary principal matrix of order k.
111     """
112     a=[]
113     for i in range(self.degree):
114       a.append(self.primary_principal_submatrix(i).determinant())
115     return a
116   def genre_list(self):
117     """
118     Return the genius of the matrix as a list of booleans in the order
119     positive defined, negative defined;
120                 semidefinite positive, semidefinite negative, indefinite.
121
122     """
123     defpos = True
124     defneg = True
125     semidefpos = True
126     semidefneg = True
127     indefinie=True
128     mineurs = self.principal_minors()
129     for i in range(len(mineurs)):
130       m = mineurs[i]
131       if m == 0:
132         defneg=False
133         defpos=False
134       if m < 0:
135         defpos=False
136         semidefpos=False
137         if i%2==0:
138           defneg=False
139       if m > 0:
140         semidefneg=False
141         if i%2==1:
142           defneg=False
143     if 0 not in mineurs:
144       semidefneg=False
145       semidefpos=False
146     if (defpos==True)or(defneg==True)or(semidefpos==True)or(semidefneg==True):indefinie=False
147     return [defpos,defneg,semidefpos,semidefneg,indefinie]
148   def __str__(self):
149     return str(self.matrix)
150
151 class QuadraticForm(SymmetricMatrix):
152   """
153   From a symmetric matrix A, provide informations concerning the associated quadratic form.
154   """
155   def __init__(self,A):
```

```
156     SymmetricMatrix.__init__(self,A)
157     if not A.is_symmetric():
158       print "Warning : matrix is not symmetric"
159   def evaluate(self,v):
160     """
161     Return the value of the quadratic form on the vector v.
162     """
163     return QuadraticMap(self.matrix,v)
164   def diagonalizing_martrix(self):
165     """
166     Return the matrix B such that B^tAB is diagonal.
167     """
168     # The transposition is because, in the matrix B, the ↩
          eigenvectors have
169     # to be read as column while Sage's matrix constructor takes ↩
          rows.
170     return matrix(self.orthonormal_basis()).transpose()
171   def new_variables(self):
172     """
173     Give the change of variables needed to put the quadratic form ↩
          under its normal form
174     X=BY
175     where X are the "old" variables
176     """
177     variables = var(automatedVar("y",self.degree))
178     Y = vector(variables)
179     return self.diagonalizing_martrix()*Y
180   def eigenmatrix_left(self):
181     return self.matrix.eigenmatrix_left()
182   def eigenvectors(self):
183     """
184     Return a list of eigenvectors of the matrix.
185
186     As the matrix is symmetric, that list has to be a basis.
187     """
188     D,P = self.eigenmatrix_left()
189     return [P[i] for i in range(P.nrows())]
190   def eigenvalues(self):
191     """
192     Return a list of eigenvalues of the matrix in the same order ↩
          as the list of eigenvectors given in
193       self.eigenvectors()
194     """
195     D,P = self.eigenmatrix_left()
196     return [ D[i,i] for i in range(D.nrows()) ]
197   def orthonormal_basis(self):
198     """
199     Return a basis of eigenvectors normalised to 1 as a list.
200     """
201     M,mu = matrix(self.eigenvectors()).gram_schmidt()
202     return [ v/v.norm() for v in M ]
203   def verification(self):
204     """
```

```
205     return the value of the quadratic form on the vector ↩
           new_variables()
206     """
207     return self.evaluate(self.new_variables())
208
209   def __str__(self):
210     a = []
211     a.append("Hi guy; I'm the quadratic form associated with the ↩
           matix")
212     a.append(str(self.matrix))
213     a.append("My eigenvalues and eigenvectors are : ")
214     veps = self.eigenvectors()
215     vaps = self.eigenvalues()
216     for i in range(len(veps)):
217       a.append("%s -> %s"%(str(vaps[i]),str(veps[i])))
218     a.append("I've the following orthonormal basis of eigenvectors↩
           :")
219     for v in self.orthonormal_basis():
220       a.append(str(v))
221     a.append("A matrix B such that B^tAB is diagonal is ")
222     a.append(str(self.diagonalizing_martrix()))
223     a.append("I'm quite pretty in the following variables ...")
224     for i in range(self.degree):
225       a.append("x%s = %s"%(str(i+1),str(self.new_variables()[i]))↩
            )
226     a.append("Look at me when I wear my cool variables")
227     a.append(str(self.verification()))
228     return "\n".join(a)
229
230 class Extrema(object):
231   """
232   From a function f, provides the informations for the study of ↩
         the extrema :
233   partial derivative
234   critical points
235   Hessian matrix at the critical points
236   Genius of the Hessian and conclusion as local min/max
237
238   Dear student : remember that this class does not furnish any ↩
         informations
239         concerning *global* extrema. The latter have to be found
240     among the critical points OR on the border of the domain.
241   """
242   def __init__(self,f):
243     var('x,y')
244     self.fun = f
245     self.gx=self.fun.diff(x).full_simplify()
246     self.gy=self.fun.diff(y).full_simplify()
247     self.gxx=self.gx.diff(x).simplify_full()
248     self.gxy=self.gx.diff(y).full_simplify()
249     self.gyy=self.gy.diff(y).full_simplify()
250     self.cp = solve( [self.gx(x,y)==0,self.gy(x,y)==0],[x,y] )
251   def critical_points(self):
```

```
252     """
253     Return the critical points as a list of tuples (x,y)
254     """
255     a = []
256     for pt in self.cp :
257       try :
258         px = SR(pt[0].rhs())
259         py = SR(pt[1].rhs())
260         a.append((px,py))
261       except TypeError :
262         a.append("  I'm not able to solve these equations.")
263     return a
264   def hessienne(self,a,b):
265     return matrix(SR,2,2,[self.gxx(a,b),self.gxy(a,b),self.gxy(a,↩
          b),self.gyy(a,b)])
266   def __str__(self):
267     a = []
268     a.append ("The function :")
269     a.append(str(self.fun))
270     a.append ("Derivative x and y :")
271     a.append(str(self.gx))
272     a.append(str(self.gy))
273     a.append ("Hessian matrix :")
274     a.append(str(self.hessienne(x,y)))
275     a.append ("Critical points :")
276     for pt in self.critical_points() :
277       a.append(str(pt))
278     for pt in self.critical_points():
279       try :
280         px = pt[0]
281         py = pt[1]
282         a.append("At (%s,%s), the Hessian is"%(str(px),str(py)))
283         try :
284           Hess = SymmetricMatrix(self.hessienne(px,py))
285           for l in Hess.matrix:
286             a.append("  "+str(l))
287           a.append("  Primary principal minors are %s"%str(Hess.↩
                principal_minors()))
288           l = Hess.genre_list()
289           if l[0]==True:
290             a.append("  Hessian positive defined")
291             a.append("  local minimum")
292           if l[1]==True:
293             a.append("  Hessian negative defined")
294             a.append("  local maximum")
295           if l[2]==True:
296             a.append("  Hessian positive semidéfinite")
297             a.append("  I don't conclude")
298           if l[3]==True:
299             a.append("  Hessian negative semidefinite")
300             a.append("  I don't conclude")
301           if l[4]==True:
302             a.append("  Undefinite Hessian")
```

```
303             a.append("  «selle» point")
304         except RuntimeError,data :
305           a.append("  "+str(data))
306       except TypeError :
307         a.append("  I'm not able to solve these equations.")
308     return "\n".join(a)
```

tex/sage/outilsINGE.sage

Exemple 44.1
Calculer la limite

$$\lim_{x\to\infty} \frac{\sin(x)\cos(x)}{x} \tag{44.1}$$

```
var('x')
f(x)=sin(x)*cos(x)/x
limit(f(x),x=oo)
```

La première ligne déclare que la lettre `x` désignera une variable. Pour la troisième ligne, notez que l'infini est écrit par deux petits « o ». △

Exemple 44.2
Quelques limites et graphes avec Sage.

(1) $\lim_{x\to 0} \frac{\sin(\alpha x)}{\sin(\beta x)}$.
Pour effectuer cet exercice avec Sage, il faut taper les lignes suivantes :

```
sage: var('x,a,b')
(x, a, b)
sage: f(x)=sin(a*x)/sin(b*x)
sage: limit( f(x),x=0  )
a/b
```

Noter qu'il faut déclarer les variables `x`, `a` et `b`.

(2) $\lim_{x\to\pm\infty} \frac{\sqrt{x^2+1}-x}{x-2}$

```
sage: f(x)=(sqrt(x**2+1))/(x-2)
sage: limit(f(x),x=oo)
1
sage: limit(f(x),x=-oo)
-1
```

Noter la commande pour la racine carré : `sqrt`. Étant donné que cette fonction diverge en $x = 2$, si nous voulons la tracer, il faut procéder en deux fois :

```
sage: plot(f,(-100,1.9))
Launched png viewer for Graphics object consisting of 1 graphics primitive
sage: plot(f,(2.1,100))
Launched png viewer for Graphics object consisting of 1 graphics primitive
```

La première ligne trace de -100 à 1.9 et la seconde de 2.1 à 100. Ces graphiques vous permettent déjà de voir les limites. Attention : ils ne sont pas des *preuves* ! Mais ils sont de sérieux indices qui peuvent vous inspirer dans vos calculs.

△

Exemple 44.3

Calculer les dérivées partielles $\partial_x f$, $\partial_y f$, $\partial^2_x f$, $\partial^2_{xy} f$, $\partial^2_{yx} f$ et $\partial^2_y f$ des fonctions suivantes.

(1) $2x^3 + 3x^2y - 2y^2$
(2) $\ln(xy^2)$
(3) $\tan(x/y)$
(4) $\frac{xy^2}{x+y}$

Le script Sage suivant (`exoDV002.sage`) résout l'exercice :

```
# -*- coding: utf8 -*-

def LesCalculs(f):
  print "Pour la fonction %s"%str(f)
  print "d_x",f.diff(x).simplify_full()
  print "d_y",f.diff(y).simplify_full()
  print "d^2_x",f.diff(x).diff(x).simplify_full()
  print "d_xd_y",f.diff(x).diff(y).simplify_full()
  print "d_yd_x",f.diff(y).diff(x).simplify_full()
  print "d^2_y",f.diff(y).diff(y).simplify_full()
  print ""

def exercise_DV002():
  var('x,y')
  fa(x,y)=2*x**3+3*x**2*y-2*y**2
  fb(x,y)=ln(x*y**2)
  fc(x,y)=tan(x/y)
  fd(x,y)=x*y**2/(x+y)
  LesCalculs(fa)
  LesCalculs(fb)
  LesCalculs(fc)
  LesCalculs(fd)
```

tex/sage/exoDV002.sage

La sortie est :

```
Pour la fonction (x, y) |--> 2*x^3 + 3*x^2*y - 2*y^2
d_x (x, y) |--> 6*x^2 + 6*x*y
d_y (x, y) |--> 3*x^2 - 4*y
d^2_x (x, y) |--> 12*x + 6*y
d_xd_y (x, y) |--> 6*x
d_yd_x (x, y) |--> 6*x
d^2_y (x, y) |--> -4

Pour la fonction (x, y) |--> log(x*y^2)
d_x (x, y) |--> 1/x
d_y (x, y) |--> 2/y
d^2_x (x, y) |--> -1/x^2
d_xd_y (x, y) |--> 0
d_yd_x (x, y) |--> 0
d^2_y (x, y) |--> -2/y^2
```

```
Pour la fonction (x, y) |--> tan(x/y)
d_x (x, y) |--> 1/(y*cos(x/y)^2)
d_y (x, y) |--> -x/(y^2*cos(x/y)^2)
d^2_x (x, y) |--> 2*sin(x/y)/(y^2*cos(x/y)^3)
d_xd_y (x, y) |--> -(2*x*sin(x/y) + y*cos(x/y))/(y^3*cos(x/y)^3)
d_yd_x (x, y) |--> -(2*x*sin(x/y) + y*cos(x/y))/(y^3*cos(x/y)^3)
d^2_y (x, y) |--> 2*(x^2*sin(x/y) + x*y*cos(x/y))/(y^4*cos(x/y)^3)

Pour la fonction (x, y) |--> x*y^2/(x + y)
d_x (x, y) |--> y^3/(x^2 + 2*x*y + y^2)
d_y (x, y) |--> (2*x^2*y + x*y^2)/(x^2 + 2*x*y + y^2)
d^2_x (x, y) |--> -2*y^3/(x^3 + 3*x^2*y + 3*x*y^2 + y^3)
d_xd_y (x, y) |--> (3*x*y^2 + y^3)/(x^3 + 3*x^2*y + 3*x*y^2 + y^3)
d_yd_x (x, y) |--> (3*x*y^2 + y^3)/(x^3 + 3*x^2*y + 3*x*y^2 + y^3)
d^2_y (x, y) |--> 2*x^3/(x^3 + 3*x^2*y + 3*x*y^2 + y^3)
```

△

Exemple 44.4

Résoudre les systèmes suivants.

(1) $\begin{cases} x_1 - 2x_2 + 3x_3 - 2x_4 = 0 \\ 3x_1 - 7x_2 - 2x_3 + 4x_4 = 0 \\ 4x_1 + 3x_2 + 5x_3 + 2x_4 = 0 \end{cases}$

(2) $\begin{cases} 2x_1 + x_2 - 2x_3 + 3x_4 = 0 \\ 3x_1 + 2x_2 - x_3 + 3x_4 = 4 \\ 3x_1 + 3x_2 + 3x_3 - 3x_4 = 9 \end{cases}$

(3) $\begin{cases} x_1 + 2x_2 - 3x_3 = 0 \\ 2x_1 + 5x_2 + 2x_3 = 0 \\ 3x_1 - x_2 - 4x_3 = 0 \end{cases}$

(4) $\begin{cases} x_1 + 2x_2 - x_3 = 0 \\ 2x_1 + 5x_2 + 2x_3 = 0 \\ x_1 + 4x_2 + 7x_3 = 0 \\ x_1 + 3x_2 + 3x_3 = 0 \end{cases}$

(5) $\begin{cases} x_1 + x_2 + x_3 + x_4 = 0 \\ x_1 + x_2 + x_3 - x_4 = 4 \\ x_1 + x_2 - x_3 + x_4 = -4 \\ x_1 - x_2 + x_3 + x_4 = 2 \end{cases}$

(6) $\begin{cases} x_1 + 3x_2 + 3x_3 = 1 \\ x_1 + 3x_2 + 4x_3 = 0 \\ x_1 + 4x_2 + 3x_3 = 3 \end{cases}$

(7) $\begin{cases} x_1 - 3x_2 + 2x_3 = -6 \\ -3x_1 + 3x_2 - x_3 = 17 \\ 2x_1 - x_2 = 3 \end{cases}$

(8) $\begin{cases} x_1 - 2x_2 + 3x_3 - 2x_4 = 0 \\ 3x_1 - 7x_2 - 2x_3 + 4x_4 = 0 \\ 4x_1 + 3x_2 + 5x_3 + 2x_4 = 0 \end{cases}$

(9) $\begin{cases} 2x_1 + x_2 - 2x_3 + 3x_4 = 0 \\ 3x_1 + 2x_2 - x_3 + 3x_4 = 4 \\ 3x_1 + 3x_2 + 3x_3 - 3x_4 = 9 \end{cases}$

(10) $\begin{cases} x_1 + 2x_2 - 3x_3 = 0 \\ 2x_1 + 5x_2 + 2x_3 = 0 \\ 3x_1 - x_2 - 4x_3 = 0 \end{cases}$

(11) $\begin{cases} x_1 + 2x_2 - x_3 = 0 \\ 2x_1 + 5x_2 + 2x_3 = 0 \\ x_1 + 4x_2 + 7x_3 = 0 \\ x_1 + 3x_2 + 3x_3 = 0 \end{cases}$

(12) $\begin{cases} x_1 + x_2 + x_3 + x_4 = 0 \\ x_1 + x_2 + x_3 - x_4 = 4 \\ x_1 + x_2 - x_3 + x_4 = -4 \\ x_1 - x_2 + x_3 + x_4 = 2 \end{cases}$

(13) $\begin{cases} x_1 + 3x_2 + 3x_3 = 1 \\ x_1 + 3x_2 + 4x_3 = 0 \\ x_1 + 4x_2 + 3x_3 = 3 \end{cases}$

(14) $\begin{cases} x_1 - 3x_2 + 2x_3 = -6 \\ -3x_1 + 3x_2 - x_3 = 17 \\ 2x_1 - x_2 = 3 \end{cases}$

Nous résolvons les systèmes en utilisant Sage avec le script suivant.

```
1 # -*- coding: utf8 -*-
2 """
```

```
"""
Ce script Sage résout un certain nombre
de systèmes d'équations linéaires du cours INGE1121
"""

import outilsINGE

def exercise_1_1_bcdefhi():
  # Exercice 1.1.b (INGE1121)
  A=matrix([ [1,-2,3,-2,0],[3,-7,-2,4,0],[4,3,5,2,0] ])
  v=vector((0,0,0,0,0))
  print outilsINGE.SolveLinearSystem(A,v)
  # Exercice 1.1.c (INGE1121)
  A=matrix([ [2,1,-2,3],[3,2,-1,3],[3,3,3,-3] ])
  v=vector((0,4,9))
  print outilsINGE.SolveLinearSystem(A,v)
  # Exercice 1.1.d (INGE1121)
  A=matrix([ [1,2,-3],[2,5,2],[3,-1,-4] ])
  v=vector((0,0,0))
  print outilsINGE.SolveLinearSystem(A,v)
  # Exercice 1.1.e (INGE1121)
  A=matrix([ [1,2,-1],[2,5,2],[1,4,7],[1,3,3] ])
  v=vector((0,0,0,0))
  print outilsINGE.SolveLinearSystem(A,v)
  # Exercice 1.1.f (INGE1121)
  A=matrix([ [1,1,1,1],[1,1,1,-1],[1,1,-1,1],[1,-1,1,1] ])
  v=vector((0,4,-4,2))
  print outilsINGE.SolveLinearSystem(A,v)
  # Exercice 1.1.h (INGE1121)
  A=matrix([ [1,3,3],[1,3,4],[1,4,3] ])
  v=vector((1,0,3))
  print outilsINGE.SolveLinearSystem(A,v)
  # Exercice 1.1.i (INGE1121)
  A=matrix([ [1,-3,2],[-3,3,-1],[2,-1,0] ])
  v=vector((-6,17,3))
  print outilsINGE.SolveLinearSystem(A,v)
```

tex/sage/exo11.sage

Le résultat est le suivant :

```
The given matrix corresponds to the system
x1 - 2*x2 + 3*x3 - 2*x4 == 0
3*x1 - 7*x2 - 2*x3 + 4*x4 == 0
4*x1 + 3*x2 + 5*x3 + 2*x4 == 0
And the solutions are
[
[x1 == -23/16*r19, x2 == -5/16*r19, x3 == 15/16*r19, x4 == r19, x5 == r18]
]
The given matrix corresponds to the system
2*x1 + x2 - 2*x3 + 3*x4 == 0
3*x1 + 2*x2 - x3 + 3*x4 == 4
3*x1 + 3*x2 + 3*x3 - 3*x4 == 9
And the solutions are
[
[x1 == 3*r20 - 7, x2 == -4*r20 + 11, x3 == r20, x4 == 1]
```

```
]
The given matrix corresponds to the system
x1 + 2*x2 - 3*x3 == 0
2*x1 + 5*x2 + 2*x3 == 0
3*x1 - x2 - 4*x3 == 0
And the solutions are
[
[x1 == 0, x2 == 0, x3 == 0]
]
The given matrix corresponds to the system
x1 + 2*x2 - x3 == 0
2*x1 + 5*x2 + 2*x3 == 0
x1 + 4*x2 + 7*x3 == 0
x1 + 3*x2 + 3*x3 == 0
And the solutions are
[
[x1 == 9*r21, x2 == -4*r21, x3 == r21]
]
The given matrix corresponds to the system
x1 + x2 + x3 + x4 == 0
x1 + x2 + x3 - x4 == 4
x1 + x2 - x3 + x4 == -4
x1 - x2 + x3 + x4 == 2
And the solutions are
[
[x1 == 1, x2 == -1, x3 == 2, x4 == -2]
]
The given matrix corresponds to the system
x1 + 3*x2 + 3*x3 == 1
x1 + 3*x2 + 4*x3 == 0
x1 + 4*x2 + 3*x3 == 3
And the solutions are
[
[x1 == -2, x2 == 2, x3 == -1]
]
The given matrix corresponds to the system
x1 - 3*x2 + 2*x3 == -6
-3*x1 + 3*x2 - x3 == 17
2*x1 - x2 == 3
And the solutions are
[
[x1 == 37, x2 == 71, x3 == 85]
]
```

△

Exemple 44.5

Pour chacun des systèmes suivants $A \cdot X = B$,

(1) Résoudre le système par échelonnement,

(2) Calculer A^{-1},

(3) Vérifier votre réponse en calculant $A^{-1}B$. Qu'êtes-vous censé obtenir ?

Les énoncés sont

(1)

$$A = \begin{pmatrix} 2 & 1 & -2 \\ 3 & 2 & 2 \\ 5 & 4 & 3 \end{pmatrix}, \quad B = \begin{pmatrix} 10 \\ 1 \\ 4 \end{pmatrix} \tag{44.2}$$

Nous utilisons Sage pour fournir la réponse. Le code suivant résout le système et donne l'inverse de la matrice :

```
# -*- coding: utf8 -*-

import outilsINGE

def exercise_1_3():
  A=matrix([[2,1,-2],[3,2,2],[5,4,3]])
  v=vector((10,1,4))
  print outilsINGE.SolveLinearSystem(A,v)
  print "Matrice inverse :"
  print A.inverse()
```

tex/sage/exo13.sage

La sortie est ici :

```
The given matrix corresponds to the system
2*x1 + x2 - 2*x3 == 10
3*x1 + 2*x2 + 2*x3 == 1
5*x1 + 4*x2 + 3*x3 == 4
And the solutions are
[
[x1 == 1, x2 == 2, x3 == -3]
]
Matrice inverse :
[  2/7  11/7  -6/7]
[ -1/7 -16/7  10/7]
[ -2/7   3/7  -1/7]
```

△

Exemple 44.6

Sachant que $(-1, 0, 1, 0)$ est un vecteur propre de la matrice

$$A = \begin{pmatrix} 2 & 1 & -1 & 1 \\ 1 & 0 & 1 & 1 \\ -1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \tag{44.3}$$

(1) Diagonaliser A au moyen d'une matrice orthogonale

(2) Écrire la forme quadratique X^tAX sous forme d'une somme pondérée de carrés.

Calculons Av afin de savoir la valeur propre associée au vecteur donné :

$$\begin{pmatrix} 2 & 1 & -1 & 1 \\ 1 & 0 & 1 & 1 \\ -1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ 3 \\ 0 \end{pmatrix}. \tag{44.4}$$

La valeur propre est donc 3. Nous savons donc que $(\lambda - 3)$ pourra être factorisé dans le polynôme caractéristique.

Pour le reste de l'exercice c'est standard et c'est résolu de la façon suivante :

```
# -*- coding: utf8 -*-

import outilsINGE

def exercise_6_5():
  A=matrix(QQ,4,4,[2,1,-1,1,1,0,1,1,-1,1,2,1,1,1,1,0])
  x=outilsINGE.QuadraticForm(A)
  print x
```

tex/sage/exo65.sage

qui retourne

```
Hi guy; I'm the quadratic form associated with the matix
[ 2  1 -1  1]
[ 1  0  1  1]
[-1  1  2  1]
[ 1  1  1  0]
My eigenvalues and eigenvectors are :
3 -> (1, 0, -1, 0)
3 -> (0, 1, 2, 1)
-1 -> (1, 0, 1, -2)
-1 -> (0, 1, 0, -1)
I've the following orthonormal basis of eigenvectors :
(1/2*sqrt(2), 0, -1/2*sqrt(2), 0)
(1/2, 1/2, 1/2, 1/2)
(1/6*sqrt(6), 0, 1/6*sqrt(6), -1/3*sqrt(6))
(-1/2*sqrt(1/3), 3/2*sqrt(1/3), -1/2*sqrt(1/3), -1/2*sqrt(1/3))
A matrix B such that B^tAB is diagonal is
[   1/2*sqrt(2)            1/2    1/6*sqrt(6) -1/2*sqrt(1/3)]
[             0            1/2             0  3/2*sqrt(1/3)]
[  -1/2*sqrt(2)            1/2    1/6*sqrt(6) -1/2*sqrt(1/3)]
[             0            1/2   -1/3*sqrt(6) -1/2*sqrt(1/3)]
I'm quite pretty in the following variables ...
x1 = -1/2*sqrt(1/3)*y4 + 1/2*sqrt(2)*y1 + 1/6*sqrt(6)*y3 + 1/2*y2
x2 = 3/2*sqrt(1/3)*y4 + 1/2*y2
x3 = -1/2*sqrt(1/3)*y4 - 1/2*sqrt(2)*y1 + 1/6*sqrt(6)*y3 + 1/2*y2
x4 = -1/2*sqrt(1/3)*y4 - 1/3*sqrt(6)*y3 + 1/2*y2
Look at me when I wear my cool variables
3*y1^2 + 3*y2^2 - y3^2 - y4^2
```

△

Exemple 44.7

Rechercher les extrema des fonctions suivantes

(1) $f(x,y) = 2 - \sqrt{x^2+y^2}$

(2) $f(x,y) = x^3 + 3xy^2 - 15x - 12y$

(3) $f(x,y) = \frac{x^3}{3} + \frac{4y^3}{3} - x^2 - 3x - 4y - 3$

Les corrigés sont créés par le script Sage `exo101.sage`

```
# -*- coding: utf8 -*-

import outilsINGE

def exercise_10_1_A():
   var('x,y')
   f(x,y)=2-sqrt(x**2+y**2)
   print outilsINGE.Extrema(f)
def exercise_10_1_B():
   var('x,y')
   f(x,y)=x**3+3*x*y**2-15*x-12*y
   print outilsINGE.Extrema(f)
def exercise_10_1_C():
   var('x,y')
   f(x,y)=x**3/3+4*y**3/3-x**2-3*x-4*y-3
   print outilsINGE.Extrema(f)
```

Des réponses :

(1)
```
The function :
(x, y) |--> -sqrt(x^2 + y^2) + 2
Derivative x and y :
(x, y) |--> -x/sqrt(x^2 + y^2)
(x, y) |--> -y/sqrt(x^2 + y^2)
Hessian matrix :
[-sqrt(x^2 + y^2)*y^2/(x^4 + 2*x^2*y^2 + y^4)                          x*y/(x^2 + y^2)^
[                       x*y/(x^2 + y^2)^(3/2) -sqrt(x^2 + y^2)*x^2/(x^4 + 2*x^2*y^2 +
Critical points :
(0, 0)
At (0,0), the Hessian is
   power::eval(): division by zero
```

Ici nous voyons que Sage a du mal à calculer la matrice hessienne en $(0,0)$. En effet, nous tombons sur une division par zéro. Pour résoudre l'exercice, il faut se rendre compte que la fonction $(x,y) \mapsto \sqrt{x^2+y^2}$ est toujours positive et est nulle seulement au point $(0,0)$. Donc f est toujours plus petite ou égale à deux tandis que $f(0,0) = 2$. Le point est donc un maximum global.

(2)
```
The function :
(x, y) |--> x^3 + 3*x*y^2 - 15*x - 12*y
Derivative x and y :
(x, y) |--> 3*x^2 + 3*y^2 - 15
(x, y) |--> 6*x*y - 12
Hessian matrix :
[6*x 6*y]
[6*y 6*x]
Critical points :
(2, 1)
(1, 2)
(-1, -2)
(-2, -1)
At (2,1), the Hessian is
   (12, 6)
```

```
   (6, 12)
   Primary principal minors are [108, 12]
   Hessian positive defined
   local minimum
At (1,2), the Hessian is
   (6, 12)
   (12, 6)
   Primary principal minors are [-108, 6]
   Undefinite Hessian
   «selle» point
At (-1,-2), the Hessian is
   (-6, -12)
   (-12, -6)
   Primary principal minors are [-108, -6]
   Undefinite Hessian
   «selle» point
At (-2,-1), the Hessian is
   (-12, -6)
   (-6, -12)
   Primary principal minors are [108, -12]
   Hessian negative defined
   local maximum
```

(3)
```
The function :
(x, y) |--> 1/3*x^3 - x^2 + 4/3*y^3 - 3*x - 4*y - 3
Derivative x and y :
(x, y) |--> x^2 - 2*x - 3
(x, y) |--> 4*y^2 - 4
Hessian matrix :
[2*x - 2       0]
[      0     8*y]
Critical points :
(3, 1)
(-1, 1)
(3, -1)
(-1, -1)
At (3,1), the Hessian is
   (4, 0)
   (0, 8)
   Primary principal minors are [32, 4]
   Hessian positive defined
   local minimum
At (-1,1), the Hessian is
   (-4, 0)
   (0, 8)
   Primary principal minors are [-32, -4]
   Undefinite Hessian
   «selle» point
At (3,-1), the Hessian is
   (4, 0)
   (0, -8)
   Primary principal minors are [-32, 4]
   Undefinite Hessian
   «selle» point
```

```
At (-1,-1), the Hessian is
   (-4, 0)
   (0, -8)
   Primary principal minors are [32, -4]
   Hessian negative defined
   local maximum
```

△

Exemple 44.8

Déterminer les valeurs extrêmes et les points de selle des fonctions suivantes.

(1) $f(x,y) = x^2 + 4x + y^2 - 2y$.
(2) $f(x,y) = e^{x^2+xy}$.
(3) $f(x,y) = e^x \sin(y)$.

Certains corrigés de cet exercice ont étés réalisés par Sage. Le script utilisé est `exo103.sage`

```
# -*- coding: utf8 -*-

import outilsINGE

def exercise_10_3_A():
  var('x,y')
  f(x,y)=x**2+4*x+y**2-2*y
  print outilsINGE.Extrema(f)

def exercise_10_3_H():
  var('x,y')
  f(x,y)=exp(x**2+x*y)
  print outilsINGE.Extrema(f)

def exercise_10_3_Q():
  var('x,y')
  f(x,y)=exp(x)*sin(y)
  print outilsINGE.Extrema(f)
```

tex/sage/exo103.sage

Des réponses :

(1)
```
The function :
(x, y) |--> x^2 + y^2 + 4*x - 2*y
Derivative x and y :
(x, y) |--> 2*x + 4
(x, y) |--> 2*y - 2
Hessian matrix :
[2 0]
[0 2]
Critical points :
(-2, 1)
At (-2,1), the Hessian is
   (2, 0)
   (0, 2)
```

```
   Primary principal minors are [4, 2]
   Hessian positive defined
   local minimum
```

(2)
```
The function :
(x, y) |--> e^(x^2 + x*y)
Derivative x and y :
(x, y) |--> (2*x*e^(x^2) + y*e^(x^2))*e^(x*y)
(x, y) |--> x*e^(x^2 + x*y)
Hessian matrix :
[(4*x*y*e^(x^2) + y^2*e^(x^2) + 2*(2*x^2 + 1)*e^(x^2))*e^(x*y)                    (x*y*e^
[                  (x*y*e^(x^2) + (2*x^2 + 1)*e^(x^2))*e^(x*y)
Critical points :
(0, 0)
At (0,0), the Hessian is
   (2, 1)
   (1, 0)
   Primary principal minors are [-1, 2]
   Undefinite Hessian
   «selle» point
```

(3)
```
The function :
(x, y) |--> e^x*sin(y)
Derivative x and y :
(x, y) |--> e^x*sin(y)
(x, y) |--> e^x*cos(y)
Hessian matrix :
[ e^x*sin(y)  e^x*cos(y)]
[ e^x*cos(y) -e^x*sin(y)]
Critical points :
   I'm not able to solve these equations.
   I'm not able to solve these equations.
At (   ,I), the Hessian is
   I'm not able to solve these equations.
At (   ,I), the Hessian is
   I'm not able to solve these equations.
```

Ici, Sage n'est pas capable de résoudre les équations qui annulent le jacobien. Les équations à résoudre sont pourtant faciles :

$$\begin{cases} e^x \cos(y) = 0 & \text{(44.5a)} \\ e^x \sin(y) = 0 & \text{(44.5b)} \end{cases}$$

Étant donné que l'exponentielle ne s'annule jamais, il faudrait avoir en même temps $\cos(y) = 0$ et $\sin(y) = 0$, ce qui est impossible. La fonction n'a donc aucun extrema local.

△

Exemple 44.9

Considérons la fonction

$$f(x, y) = xy^2 e^{-(x^2+y^2)/4}. \tag{44.6}$$

(1) Montrer qu'il y a une infinité de points critiques.

(2) Déterminer leur nature.

Voici la fonction Sage qui fournit les informations :

```
# -*- coding: utf8 -*-

import outilsINGE

def exercise_10_4():
  var('x,y')
  f(x,y)=x*y**2*exp(-(x**2+y**2)/4)
  print outilsINGE.Extrema(f)
```

tex/sage/exo104.sage

La sortie est

```
The function :
(x, y) |--> x*y^2*e^(-1/4*x^2 - 1/4*y^2)
Derivative x and y :
(x, y) |--> -1/2*(x^2 - 2)*y^2*e^(-1/4*x^2 - 1/4*y^2)
(x, y) |--> -1/2*(x*y^3 - 4*x*y)*e^(-1/4*x^2 - 1/4*y^2)
Hessian matrix :
[                1/4*(x^3 - 6*x)*y^2*e^(-1/4*x^2 - 1/4*y^2) 1/4*((x^2 - 2)*y^3 - 4*(x^2 - 2
[1/4*((x^2 - 2)*y^3 - 4*(x^2 - 2)*y)*e^(-1/4*x^2 - 1/4*y^2)        1/4*(x*y^4 - 10*x*y^2 + 
Critical points :
(r17, 0)
(-sqrt(2), -2)
(sqrt(2), -2)
(-sqrt(2), 2)
(sqrt(2), 2)
At (r17,0), the Hessian is
   (0, 0)
   (0, 2*r17*e^(-1/4*r17^2))
   Primary principal minors are [0, 0]
   Hessian positive semidéfinite
   I don't conclude
   Hessian negative semidefinite
   I don't conclude
At (-sqrt(2),-2), the Hessian is
   (4*sqrt(2)*e^(-3/2), 0)
   (0, 4*sqrt(2)*e^(-3/2))
   Primary principal minors are [32*e^(-3), 4*sqrt(2)*e^(-3/2)]
   Hessian positive defined
   local minimum
At (sqrt(2),-2), the Hessian is
   (-4*sqrt(2)*e^(-3/2), 0)
   (0, -4*sqrt(2)*e^(-3/2))
   Primary principal minors are [32*e^(-3), -4*sqrt(2)*e^(-3/2)]
   Hessian negative defined
   local maximum
At (-sqrt(2),2), the Hessian is
   (4*sqrt(2)*e^(-3/2), 0)
   (0, 4*sqrt(2)*e^(-3/2))
   Primary principal minors are [32*e^(-3), 4*sqrt(2)*e^(-3/2)]
   Hessian positive defined
```

```
   local minimum
At (sqrt(2),2), the Hessian is
   (-4*sqrt(2)*e^(-3/2), 0)
   (0, -4*sqrt(2)*e^(-3/2))
   Primary principal minors are [32*e^(-3), -4*sqrt(2)*e^(-3/2)]
   Hessian negative defined
   local maximum
```

Notez la présence de `r1` comme paramètres dans les solutions. Tous les points avec $y = 0$ sont des points critiques. Cependant, Sage [1] ne parvient pas à conclure la nature de ces points $(x, 0)$.

Notons que le nombre $f(x, y)$ a toujours le signe de x parce que y^2 et l'exponentielle sont positives. Toujours ? En tout cas lorsque $x \neq 0$. Prenons un point $(a, 0)$ avec $a > 0$. Dans un voisinage de ce point, nous avons $f(x, y) > 0$ parce que si $a > 0$, alors $x > 0$ dans un voisinage de a. Le point $(a, 0)$ est un minimum local parce que $0 = f(a, 0) \leqslant f(x, y)$ pour tout (x, y) dans un voisinage de $(a, 0)$.

De la même façon, les points $(a, 0)$ avec $a < 0$ sont des maxima locaux parce que dans un voisinage, la fonction est négative.

Le point $(0, 0)$ n'est ni maximum ni minimum local. C'est un point de selle.

△

Exemple 44.10

Dériver les fonctions suivantes.

(1) $\sin\big(\ln(x)\big)$

(2) $\dfrac{\sin x}{x}$;

(3) e^{x^2}

(4) $\cos(x)^{\sin(x)}$

Le programme suivant par Sage résout l'exercice :

```
#! /usr/bin/sage -python
# -*- coding: utf8 -*-

from sage.all import *

var('x')
f=sin(ln(x))
print f.diff(x)
f=sin(x)/x
print f.diff(x)
f=exp(x**2)
print f.diff(x)
f=cos(x)**(sin(x))
print f.diff(x)
```

tex/sage/corrDerive_0002.sage

Le résultat est :

```
cos(log(x))/x
cos(x)/x - sin(x)/x^2
```

1. ou, plus précisément, le programme que j'ai écrit avec Sage.

```
2*x*e^(x^2)
(log(cos(x))*cos(x) - sin(x)^2/cos(x))*cos(x)^sin(x)
```

△

Exemple 44.11

Donner une approximation de ln(1.0001).

```
----------------------------------------------------------------------
| Sage Version 4.5.3, Release Date: 2010-09-04                       |
| Type notebook() for the GUI, and license() for information.        |
----------------------------------------------------------------------
sage: numerical_approx(ln(1.0001))
0.0000999950003332973
```

△

Chapitre 45

Épilogue : la constante de Weiner

Nous voici à la fin du Frido. Nous avons étudié beaucoup de math, et beaucoup reste à voir. En guise de conclusion, je voudrais vous parler de la constante de Weiner, introduite dans [435]. Il s'agit d'une constante, qui comme π ou e intervient dans à peu près tous les domaines de la mathématique.

Comme toujours, il existe énormément de définitions équivalentes différentes ; nous choisissons celle-ci, motivée par le lemme 28.58.

Définition 45.1.
La ***constante de Weiner*** *est l'unique entier p tel que l'espace L^p soit un espace de Hilbert.*

Cette constante intervient de façon centrale dans de nombreux résutats dans tous les domaines ; nous en citons quelques-uns.

(1) La moyenne de tout couple de réels peut être calculée en divisant leur somme par la constante de Weiner [1].

(2) La constante de Weiner donne l'indice du groupe alterné dans le groupe symétrique pour tous les ordres, théorème 6.30.

(3) La constante de Weiner donne une borne inférieure optimale pour l'ensemble des nombres premiers.

(4) L'unique point fixe non trivial de la fonction factorielle est la constante de Weiner.

(5) La fameuse droite critique de la conjecture de Riemann est donnée par l'inverse de la constante de Weiner.

(6) Tout automorphisme d'anneau a un polynôme minimal dont le degré est donné par la constante de Weiner.

(7) Pour l'anecdote, la constante de Weiner donne le rapport τ/π.

(8) L'égalité $ab = 0$ dans un anneau n'implique pas spécialement $a = 0$ ou $b = 0$ lorsque la caractéristique de l'anneau est égale à la constante de Weiner, et seulement dans ce cas.

D'aucuns pourraient objecter que tout cela n'est que fantaisie et trivialité. Il n'en est rien. La preuve que la constante de Weiner est centrale en mathématique est précisément qu'elle avait déjà un nom et un symbole réservé bien avant le début de l'histoire des mathématiques.

Le fait est que toutes les mathématiques que vous connaissez se basent sur les nombres entiers ; cela n'est pas du tout une trivialité.

1. C'est historiquement la première propriété énoncée de la constante de Weiner ; elle suggère également une notion de constante de Weiner généralisée pour moyenner un nombre arbitraire de nombres. La construction des nombres de Weiner généralisés est en projet dans la section 2.2.

Chapitre 46

GNU Free Documentation License

Version 1.3, 3 November 2008
Copyright © 2000, 2001, 2002, 2007, 2008 Free Software Foundation, Inc.

http://fsf.org/

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom : to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation : a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals ; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "**Document**", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "**you**". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "**Modified Version**" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "**Secondary Section**" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical

connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "**Invariant Sections**" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "**Cover Texts**" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "**Transparent**" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "**Opaque**".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "**Title Page**" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "**publisher**" means any person or entity that distributes copies of the Document to the public.

A section "**Entitled XYZ**" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "**Acknowledgements**", "**Dedications**", "**Endorsements**", or "**History**".) To "**Preserve the Title**" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties : any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts : Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version :

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
D. Preserve all the copyright notices of the Document.
E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.
I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another ; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original

documents, forming one section Entitled “History”; likewise combine any sections Entitled “Acknowledgements”, and any sections Entitled “Dedications”. You must delete all sections Entitled “Endorsements”.

COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an “aggregate” if the copyright resulting from the compilation is not used to limit the legal rights of the compilation’s users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document’s Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled “Acknowledgements”, “Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/ .

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM : How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page :

> Copyright © YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation ; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the “with . . . Texts.” line with this :

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Liste des notations

$N \lhd G$ Le sous-groupe N est normal dans G, page 131

Algèbre

$[\mathbb{L} : \mathbb{K}]$ degré d'une extension de corps, page 288

$\mathcal{L}(E, F)$ Ensemble des applications linéaires de E dans F, page 204

$\mathbb{K}(A)$ corps contenant $\mathbb{K}$ et A, page 296

$\mathbb{K}[A]$ anneau contenant $\mathbb{K}$ et A, page 296

$\mathbb{N}_0$ les naturels non nuls : $\mathbb{N}_0 = \mathbb{N}\backslash\{0\}$, page 103

$\mathcal{M}_{n\times m}$ l'ensemble des matrices $n \times m$, page 204

∇f gradient de la fonction f, page 681

$\mathtt{proj}_V$ projection de $V \times W$ sur V, page 507

$\mathrm{Span}(A)$ l'ensemble des combinaisons linéaires finies d'éléments de A, page 197

$C^1(U, \mathbb{R}^n)$ Les applications une fois continument dérivables, page 694

$df_a(u)$ Application de la différentielle de f sur le vecteur u, page 666

$f^{(n)}$ La n-ième dérivée de la fonction f, page 706

$o(x)$ fonction tendant rapidement vers zéro, page 707

(p) idéal engendré par p, page 167

$\mathbb{F}_p$ lorsque p est premier, page 271

$\mathbb{F}_{p^n}$ corps fini à p^n éléments, page 1168

$\mathrm{Frac}(\mathbb{A})$ Le corps des fractions de l'anneau $\mathbb{A}$, page 111

$\mathrm{Fun}(X, Y)$ les applications de X vers Y, page 105

$\mathrm{S}(E)$ Les opérateurs autoadjoints de E, page 414

U_n Le groupe des racines n^{e}de l'unité., page 1151

$\mathrm{res}(P, Q)$ résultat des polynômes P et Q, page 405

$\sqrt{A}$ racine d'une matrice hermitienne positive, page 747

$\theta_\alpha(P)$ la multiplicité de α par rapport à P, page 194

$A[X]$ tous les polynômes de degré fini à coefficients dans A, page 190

$A_n[X]$ les polynômes à coefficients dans A et de degré inférieur à n, page 191

$C(P)$ matrice compagnon, page 470

$D \mid P$ D divise P, page 192

$E_\lambda(u)$ Espace propre de u, page 426

$mat_{\mathcal{B}}(q)$ matrice de q dans la base $\mathcal{B}$, page 1034

$U(A)$ ensemble des inversibles, page 107

Ensembles de matrices

$S^+(n,\mathbb{R})$ matrices symétriques définies positives, page 452

$S^{++}(n,\mathbb{R})$ matrices symétriques strictement définies positives, page 452

$\mathrm{Aut}(E)$ automorphisme de l'espace vectoriel E, page 205

$\mathrm{End}(E)$ les endomorphismes de E, page 205

$L(E,F)$ applications linéaires bornées (continues), page 502

$\mathrm{O}(n,\mathbb{R})$ le groupe des matrices orthogonales, page 415

$\Omega(E)$ formes quadratiques non dégénérées, page 1037

$Q(E)$ formes quadratiques réelles sur E, page 233

$Q^+(E)$ formes quadratiques positives, page 1037

$Q^{++}(E)$ formes quadratiques strictement définies positives, page 1037

$S_n^{p,q}(\mathbb{R})$ matrices symétriques réelles de signature (p,q), page 1038

$\beta(s)$ Vecteur unitaire de la binormale, page 1321

$\gamma \sim g$ Équivalence d'arcs paramétrés, page 1312

$\nu(s)$ Vecteur unitaire de la normale principale, page 1321

$c(s)$ rayon de courbure, page 1321

$t(s)$ Torsion, page 1322

Géométrie

$(x_0 : \ldots : x_n)$ coordonnées homogènes dans un espace projectif, page 1372

$\mathcal{C}\mathrm{onv}(A)$ enveloppe convexe, page 564

$\mathbb{C}[G]$ combinaisons d'éléments de G à coefficients dans $\mathbb{C}$, page 947

$\mathrm{PGL}(E)$ groupe projectif, page 1360

B^o orthogonal dans le dual, page 238

$P(E)$ l'espace projectif de E, page 1353

$P_1(\mathbb{C})$ sphère de Riemann, page 1375

Chaînes de Markov

$\pi(x)$ lié au temps de retour, page 1941

Probabilités et statistique

$\sigma(X)$ La tribu engendrée par la variable aléatoire X, page 1815

$a \wedge b$ $\min(a,b)$, page 1959

K_X matrice de covariance d'un vecteur gaussien, page 1865

$m(\mathcal{A})$ Ensemble des fonctions $\mathcal{A}$-mesurables, page 784

Théorie des groupes

$(G/H)_g$ classes à gauche, page 154

$[a]_p$ ensemble des $a + kp$, page 146

$[G, G]$ groupe dérivé, page 134

$[g, h]$ commutateur dans un groupe, page 134

$\mathrm{Aff}(\mathbb{R}^n)$ Le groupe des applications affines bijectives de $\mathbb{R}^n$., page 573

gr groupe engendré, page 131

$\hat{G}$ groupe des caractères de G, page 943

σ_x réflexion par rapport à x, page 1055

A_n groupe alterné, page 257

$D(G)$ groupe dérivé, page 134

D_n groupe diédral, page 1102

G^{ab} groupe abélianisé de G, page 134

$N \times_\phi H$ produit semi-direct, page 162

S_n le groupe symétrique, page 158

Topologie et théorie des ensembles

$\mathrm{Adh}(A)$ adhérence de A, page 324

$\complement A$ Le complémentaire de l'ensemble A, page 101

$\mathrm{Diam}(A)$ Diamètre de la partie A, page 618

$\mathrm{Int}(A)$ intérieur de A, page 324

∂A La frontière de l'ensemble A, page 423

$A \Delta B$ différence symétrique, page 102

A^c complémentaire de A, page 101

Analyse

$\mathrm{Isom}(X)$ Le groupe des isométries de X, page 358

$\mu \ll \nu$ La mesure μ est absolument continue par rapport à la mesure ν., page 820

$C^\infty\big(\mathbb{R}, \mathscr{S}'(\mathbb{R}^d)\big)$ Fonctions à valeurs dans les distributions., page 1673

$C^\infty(I, \mathscr{D}'(\mathbb{R}^d))$ fonctions à valeurs dans les distributions, page 1602

$(S, \hat{\mathcal{F}}, \hat{\mu})$ complété de l'espace mesuré $(S, \hat{\mathcal{F}}, \hat{\mu})$, page 777

$\mathcal{L}(E, F)$ Les applications linéaires de E vers F, page 533

$\mathcal{L}^{(n)}(V, W)$ L'espace des applications n-linéaires $V^n \to W$, page 698

$\arg(z)$ La valeur principale de l'argument de $z \in \mathbb{C}$, page 1531

L Les applications linéaires continues de E vers F, page 533

$\mathbb{D}_b$ l'ensemble de écritures décimales en base b, page 519

$\mathbb{R}$ l'ensemble des réels, page 120

$\mathbb{R}^+$ les réels positifs ou nuls, page 122

exp exponentielle, page 906

$\mathscr{H}'$ dual, page 1425

$\liminf a_n$ limite inférieure, page 717

$\limsup a_n$ limite supérieure, page 717

$\mathcal{L}^p$ espace de Lebesgue, sans les classes, page 1448

$\mu \perp \nu$ mesures perpendiculaires, page 820

μ^* La mesure extérieure associée à la mesure μ, page 772

∂_z,$\partial_{\bar{z}}$ dérivées partielles d'une fonction complexe, page 1499

$\mathtt{proj}_K(x)$ projection orthogonale de x sur y, page 1422

$\sigma(\mathcal{A})$ tribu engendrée par $\mathcal{D}$, page 760

$\mathscr{D}(\Omega)$ Les fonctions C^∞ à support compact sur Ω, page 1588

$\mathscr{S}'(\mathbb{R}^d)$ espace des distributions tempérées, page 1594

$A^2(\Omega)$ espace de Bergman, page 1543

$A^\perp$ orthogonal d'une partie., page 1423

$C_c(I)$ fonctions continues à support compact dans I, page 1463

$f \sim g$ fonctions ayant des limites équivalentes, page 654

$H^1(\Omega)$ espace de Sobolev sur Ω, page 1616

$H^1(I)$ espace de Sobolev, page 1611

$H^m(M)$ espace de Sobolev, page 1617

$L^1_{loc}(I)$ fonctions intégrables sur les compacts de I, page 1611

L^p espace de Lebesgue avec les classes, page 1449

$M_i\varphi$ La fonction $x \mapsto x_i\varphi(x)$, page 1494

$S_n f$ somme partielle de série de Fourier, page 1474

Bibliographie

[1] L'auteur de ces lignes. Invention personnelle de l'auteur de ces lignes, lire avec prudence et merci de me dire si c'est correct.

[2] A. Casamayou, N Cohen, G. Connan, T. Dumont, L. Fousse, F. Maltey, M. Meulien, M. Mezzarobba, C. Pernet, M. N. Thiéry, and Zimmermann P. Calcul mathématique avec Sage. 2013. URL http://sagebook.gforge.inria.fr/.

[3] cdrcprds. Question sur les corps de rupture. 2017. URL https://github.com/LaurentClaessens/mazhe/issues/62.

[4] Arnaud Girand. Développements pour l'agrégation externe de mathématiques. 18 août 2012. URL http://perso.univ-rennes1.fr/arnaud.girand/pdf/dvp_agreg/dvp.pdf.

[5] Rémi Peyre and Frédéric Simon. Travaux dirigés de probabilités. 2009. URL http://www.normalesup.org/~rpeyre/pro/enseignement/td09pbas-e.pdf.

[6] Sylvie Benzoni. Distributions tempérées. 13 novembre 2013. URL http://math.univ-lyon1.fr/~benzoni/Agreg/Distributions.pdf.

[7] Wikipédia. Axiome du choix — wikipédia, l'encyclopédie libre, 2015. URL http://fr.wikipedia.org/w/index.php?title=Axiome_du_choix&oldid=116151439. [En ligne ; Page disponible le 9-septembre-2015].

[8] O. Teytaud, C. Antonini, J.-B. Bardet, J.-F. Quint, M. De Crisenoy, P. Borgnat, J. Bérard, E. Lebeau, E. Souche, and A. Chateau. Les mathématiques pour l'Agrégation. 21 mai 2002. http://les.mathematiques.free.fr/pdf/agreg.zip.

[9] Pedro Tamaroff. Prove that if A is an infinite set then Ax2 is equipotent to A. 2014. URL https://math.stackexchange.com/questions/1041731/prove-that-if-a-is-an-infinite-set-then-a-times-2-is-equipotent-to-a.

[10] Daniel Daigle. Cardinalité. URL http://mysite.science.uottawa.ca/asavag2/mat2762/cardinal.pdf.

[11] Christine Laurent-Thiébaut. Aximatique des nombres. URL http://ljk.imag.fr/membres/Bernard.Ycart/mel/ax/ax.pdf.

[12] Ilie Grigorescu. Lecture 4 - finite and infinite sets. URL http://www.math.miami.edu/~igrigore/teaching/mth533/lecture4_mth433.pdf.

[13] Patrice Tauvel. *Cours d'algèbre*. Dunod, 1999. ISBN 2-10-005490-3.

[14] Wikipédia. Morphisme d'anneaux — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Morphisme_d%27anneaux&oldid=129892066. [En ligne ; Page disponible le 23-septembre-2016].

[15] Wikipédia. Corps des fractions — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Corps_des_fractions&oldid=130674526. [En ligne ; Page disponible le 14-octobre-2016].

[16] Frédéric Paulin. Topologie, analyse et calcul différentiel. 2008-2009. http://www.math-question-center.com/publications-pdf/cours_d'analyse+topologie_et_calcul_differntiel_Frederic_Paulin.pdf.

[17] Wikipédia. Nombre positif — wikipédia, l'encyclopédie libre, 2018. URL http://fr.wikipedia.org/w/index.php?title=Nombre_positif&oldid=146043202. [En ligne ; Page disponible le 3-mars-2018].

[18] Peter Hendrikus Kropholler. Groups rings and fields. 2004-2005. http://www.maths.gla.ac.uk/~phk/ulearning.pdf.

[19] Anonyme. Théorie des groupes. 2008. Lien original mort : http://theoriedesgroupes.perso.sfr.fr/cours/theoriePDF.pdf autre lien, qui fonctionne pour l'instant : http://ekladata.com/CKoLuTDRT9S2sx5BPX9DkCDAcdo.pdf et https://web.archive.org/web/20150701233527/http://theoriedesgroupes.perso.sfr.fr/cours/theoriePDF.pdf.

[20] Wikiversité. Arithmétique/divisibilité et congruences dans z — wikiversité,, 2017. URL https://fr.wikiversity.org/w/index.php?title=Arithm%C3%A9tique/Divisibilit%C3%A9_et_congruences_dans_Z&oldid=650617. [En ligne ; accédé le 25-mai-2017].

[21] Mongi Amorri. Théorème de bezout, théorème de gauss. http://lux.lyceefrancais-brasilia.net/documents/cours/ts/bezout_gauss.pdf.

[22] Gilles Costantini. pgcd et ppcm dans $\mathbb{Z}$, théorème de Bezout, applications. . http://gilles.costantini.pagesperso-orange.fr/agreg_fichiers/bezout.pdf.

[23] Wikipédia. Lemme d'euclide — wikipédia, l'encyclopédie libre, 2015. URL http://fr.wikipedia.org/w/index.php?title=Lemme_d%27Euclide&oldid=112046247. [En ligne ; Page disponible le 28-juin-2015].

[24] Wikipédia. Théorème fondamental de l'arithmétique — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_fondamental_de_l%27arithm%C3%A9tique&oldid=110036012. [En ligne ; Page disponible le 26-juin-2015].

[25] Georges Skandalis. Algèbre générale et algèbre linéaire. 6 mars 2015. URL http://www.math.univ-paris-diderot.fr/_media/formations/prepa/agreginterne/polycopiealgebre.pdf.

[26] Jean Luc W. groupe-quotient et création de nouveaux sous-groupes. 2007. URL https://fr.wikipedia.org/wiki/Discussion:Groupe_quotient.

[27] Wikipédia. Théorème de cauchy (groupes) — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Cauchy_(groupes)&oldid=140167231. [En ligne ; Page disponible le 17-septembre-2017].

[28] Alexei Pantchichkine. Algèbre2. 2004-2005. http://www-fourier.ujf-grenoble.fr/~panchish/05ensl.pdf.

[29] François Combes. *Algèbre et géométrie*. Bréal, 2003. ISBN 2-84291-202-0.

[30] Niels Borne. Fiche numéro 2 : morphismes, sous-groupe distingué, quotient. 2009-2012. http://math.univ-lille1.fr/~borne/Enseignement/TD2_M308.pdf.

[31] Pierre Lissy. Théorème de Jordan-Hölder. 6 avril 2010. http://www.ljll.math.upmc.fr/~lissy/Agreg/developpements/JordanHolder.pdf.

[32] Muriel Galley. Groupes résolubles. 24 mai 2012. URL http://matthieu.gendulphe.com/Galley.pdf.

[33] Fabrice Castel. Groupes finis. 2009-2012. http://agreg-maths.univ-rennes1.fr/documentation/docs/Groupes_finis.pdf.

[34] Wikipédia. Groupe symétrique — wikipédia, l'encyclopédie libre, 2018. URL http://fr.wikipedia.org/w/index.php?title=Groupe_sym%C3%A9trique&oldid=146656747. [En ligne ; Page disponible le 21-mars-2018].

[35] Jean Delcourt. Groupes-permutations. URL http://delcourt.u-cergy.fr/StrucAlg/chap_5.pdf.

[36] Wikipédia. Permutation — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Permutation&oldid=89805226. [En ligne ; Page disponible le 24-juin-2013].

[37] cdrcprds. Questions (9), (15) et (21) d'algèbre. 2017. URL https://github.com/LaurentClaessens/mazhe/issues/60.

[38] Wikipédia. Formule du binôme de newton — wikipédia, l'encyclopédie libre, 2018. URL http://fr.wikipedia.org/w/index.php?title=Formule_du_bin%C3%B4me_de_Newton&oldid=144161875. [En ligne ; Page disponible le 4-janvier-2018].

[39] Wikipédia. Caractéristique d'un anneau — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Caract%C3%A9ristique_d%27un_anneau&oldid=136316021. [En ligne ; Page disponible le 22-mai-2017].

[40] Wikipédia. Module sur un anneau — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Module_sur_un_anneau&oldid=141582982. [En ligne ; Page disponible le 16-octobre-2017].

[41] Wikipédia. Algèbre sur un corps — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Alg%C3%A8bre_sur_un_corps&oldid=98992040. [En ligne ; Page disponible le 6-mars-2014].

[42] Anonyme. sans titre. http://www.ac-grenoble.fr/champo/IMG/arithmetique.pdf.

[43] Wikipédia. Primalité dans un anneau — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Primalit%C3%A9_dans_un_anneau&oldid=127919489. [En ligne ; Page disponible le 30-mai-2017].

[44] Wikipédia. Idéal premier — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Id%C3%A9al_premier&oldid=125964642. [En ligne ; Page disponible le 7-mai-2016].

[45] Danny-Jack Mercier. Anneaux factoriels. 5 octobre 2003. http://megamaths.perso.neuf.fr/cours/ari/cann0001.pdf.

[46] Wikipédia. Anneau euclidien — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Anneau_euclidien&oldid=126145344. [En ligne ; Page disponible le 1-juin-2017].

[47] Wikiversité. Anneau (mathématiques)/exercices/exercices — wikiversité,, 2017. URL https://fr.wikiversity.org/w/index.php?title=Anneau_(math%C3%A9matiques)/Exercices/Exercices&oldid=676468. [En ligne ; accédé le 26-octobre-2017].

[48] Sébastien Pellerin. Développements d'algèbre pour l'oral de l'agrégation. . http://pellerin.xyz/doc/agreg/algebre.pdf.

[49] Wikipédia. Triplet pythagoricien — wikipédia, l'encyclopédie libre, 2015. URL http://fr.wikipedia.org/w/index.php?title=Triplet_pythagoricien&oldid=114628048. [En ligne ; Page disponible le 24-juin-2015].

[50] Wikipédia. Primalité dans un anneau — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Primalit%C3%A9_dans_un_anneau&oldid=127919489. [En ligne ; Page disponible le 1-juin-2017].

[51] Wikipédia. Lemme de gauss (polynômes) — wikipédia, l'encyclopédie libre, 2015. URL http://fr.wikipedia.org/w/index.php?title=Lemme_de_Gauss_(polyn%C3%B4mes)&oldid=112056131. [En ligne ; Page disponible le 1-juin-2017].

[52] Wikipédia. Espace vectoriel — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Espace_vectoriel&oldid=143370323. [En ligne ; Page disponible le 10-décembre-2017].

[53] Espaces vectoriels de dimension infinie. . URL http://www.les-mathematiques.net/b/e/e/node8.php.

[54] Wikipedia contributors. Tensor product — Wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Tensor_product&oldid=842778935. [Online ; accessed 18-June-2018].

[55] Marie-Pierre Lebaux. Rappels sur les applications linéaires. URL https://perso.univ-rennes1.fr/marie-pierre.lebaud/agint/ecrit/algebre-lineaire/applications-lineaires/V-appli-lin.pdf.

[56] Daniel Ferrand. Étendre le corps. Juillet 2007. URL http://webusers.imj-prg.fr/~daniel.ferrand/ExtCorps.pdf.

[57] G. Donald Allen. Matrices and linear algebra. 2004. URL http://www.math.tamu.edu/~dallen/m640_03c/lectures/chapter2.pdf.

[58] Wikipédia. Mineur (algèbre linéaire) — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Mineur_(alg%C3%A8bre_lin%C3%A9aire)&oldid=133960288. [En ligne ; Page disponible le 25-janvier-2017].

[59] Wikipédia. Application bilinéaire — wikipédia, l'encyclopédie libre, 2018. URL http://fr.wikipedia.org/w/index.php?title=Application_bilin%C3%A9aire&oldid=144184234. [En ligne ; Page disponible le 5-janvier-2018].

[60] Wikipédia. Forme quadratique — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Forme_quadratique&oldid=103591878. [En ligne ; Page disponible le 7-mai-2014].

[61] Marc Sage. Dualité en dimension finie. 24 octobre 2005. http://www.normalesup.org/~sage/Enseignement/Cours/DualDimFinie.pdf.

[62] Wikipédia. Théorème de cauchy (groupes) — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Cauchy_(groupes)&oldid=122684298. [En ligne ; Page disponible le 1-août-2017].

[63] Wikiversité. Groupe (mathématiques)/exercices/premiers résultats sur les groupes simples — wikiversité,, 2016. URL https://fr.wikiversity.org/w/index.php?title=Groupe_(math%C3%A9matiques)/Exercices/Premiers_r%C3%A9sultats_sur_les_groupes_simples&oldid=597195. [En ligne ; accédé le 31-mai-2017].

[64] Wikiversité. Groupe (mathématiques)/automorphismes d'un groupe cyclique — wikiversité,, 2016. URL https://fr.wikiversity.org/w/index.php?title=Groupe_(math%C3%A9matiques)/Automorphismes_d%27un_groupe_cyclique&oldid=595608. [En ligne ; accédé le 3-juin-2017].

[65] Structure des groupes d'ordre pq. URL http://agregmaths.free.fr/doc/docs_nicolas/developpementAlgebre/grouped'ordrepq.pdf.

[66] Olivier Debarre. Td4 : produit semi-direct. 2015-2016. URL http://www.math.ens.fr/~debarre/TDC4.pdf.

[67] Christophe Mourougane. Théorie des groupes et géométrie. 2009-2010. http://perso.univ-rennes1.fr/christophe.mourougane/enseignements/2009-10/THGG/poly.groupes.pdf.

[68] Michel Emsalem and Pierre Dèbes. Théorème de sylow. http://exo7.emath.fr/ficpdf/fic00023.pdf.

[69] N.G.J. Pagnon. Le groupe symétrique et le groupe alterné. URL http://amatheux.com/IMG/pdf/groupe_symetrique.pdf.

[70] Wikipédia. Corps ordonné — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Corps_ordonn%C3%A9&oldid=125185412. [En ligne ; Page disponible le 22-mai-2017].

[71] Automorphisme de R et continuité. 2008. URL http://forums.futura-sciences.com/mathematiques-superieur/261451-automorphisme-de-r-continuite.html.

[72] Sylvain Duchet. Congrugence dans $\mathbb{Z}$, anneaux $\mathbb{Z}/n\mathbb{Z}$, applications. http://epsilon.2000.free.fr/Csup/congruences.pdf.

[73] Wikipédia. Polynôme irréductible — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Polyn%C3%B4me_irr%C3%A9ductible&oldid=132607764. [En ligne ; Page disponible le 12-décembre-2016].

[74] Patrick Polo. Extension de corps, caractéristique, corps de rupture, corps de décomposition, clôtures algébriques. 2007-2008. URL https://webusers.imj-prg.fr/~patrick.polo/M1Galois/ATG07chVII.pdf.

[75] C. Antonini, JF. Quint, P. Borgnat, J. Bérard, E. Lebeau, E. Souche, A. Chateau, and O. Teytaud. Le théorème de wedderburn. . URL http://www.les-mathematiques.net/d/a/w/node5.php.

[76] cdrcprds. Lemme 5.43 sur les pgcd de polynômes. 2017. URL https://github.com/LaurentClaessens/mazhe/issues/52#issuecomment-333251728.

[77] Wikipédia. Extension de corps — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Extension_de_corps&oldid=133373097. [En ligne ; Page disponible le 5-janvier-2017].

[78] Wikipédia. Extension de corps — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Extension_de_corps&oldid=133373097. [En ligne ; Page disponible le 5-janvier-2017].

[79] David Harari. Algèbre 1 - notions de théorie des corps. URL https://www.math.u-psud.fr/~harari/exposes/corps.pdf.

[80] Pierre Bernard. Entiers algébriques et racines de l'unité. 7 janvier 2012. URL http://allken-bernard.org/pierre/weblog/?p=2061.

[81] Mortajine Abdellatif. Les extensions de corps. 2012-2013. URL http://www.iecn.u-nancy.fr/~mortajin/chap3-L3-S5.pdf.

[82] Wikipédia. Corps de décomposition — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Corps_de_d%C3%A9composition&oldid=139881008. [En ligne ; Page disponible le 19-août-2017].

[83] Patrick Polo. Extensions normales, séparables, galoisiennes. corps fini. 2007-2008. http://www.math.jussieu.fr/~polo/M1/ATG07chVIII.pdf.

[84] Olivier Dodane. Le théorème de zéros de Hilbert. 8 octobre 2007. http://www.math.ens.fr/~debarre/nullstellensatz.pdf.

[85] Patrick Polo. Polynômes symétriques et résolutions d'équations. 13 décembre 2004. http://www.math.jussieu.fr/~polo/M1/ATGch9.pdf.

[86] Vincent Beck, Jérôme Malick, and Gabriel Peyré. Sur c tout est connexe ! URL http://objagr.gforge.inria.fr/documents/files/connexite-polynome.pdf.

[87] El Hage. Équation générale de degré n. http://les.mathematiques.free.fr/pdf/gal10.pdf.

[88] Florian Morel Chevillet. Résolubilité par radicaux des équations algébriques. 31 mai 2012. URL http://matthieu.gendulphe.com/MorelChevillet.pdf.

[89] Wikipédia. Théorème fondamental de l'algèbre — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_fondamental_de_l%27alg%C3%A8bre&oldid=139327371. [En ligne ; Page disponible le 20-août-2017].

[90] Wikipédia. Espace séparé — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Espace_s%C3%A9par%C3%A9&oldid=97013072. [En ligne ; Page disponible le 27-septembre-2013].

[91] C. Antonini, JF. Quint, P. Borgnat, J. Bérard, E. Lebeau, E. Souche, A. Chateau, and O. Teytaud. Généralités. . URL http://www.les-mathematiques.net/a/a/b/node19.php.

[92] Emmanuel Vieillard Baron. Sous-espaces compacts. 2001. URL http://www.les-mathematiques.net/a/t/c/node5.php.

[93] C. Antonini, JF. Quint, P. Borgnat, J. Bérard, E. Lebeau, E. Souche, A. Chateau, and O. Teytaud. Espaces métriques compacts. . http://www.les-mathematiques.net/a/a/b/node22.php.

[94] Éric Brunelle. Norme matricielle. http://www.dms.umontreal.ca/~math1600/6Supplement/Normematricielle-1.pdf.

[95] Wikipédia. Théorème de borel-lebesgue — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Borel-Lebesgue&oldid=96560796. [En ligne ; Page disponible le 11-mars-2014].

[96] Wikipédia. Formule de stirling — wikipédia, l'encyclopédie libre, Mai 2013. URL http://fr.wikipedia.org/w/index.php?title=Formule_de_Stirling&oldid=93224272. [En ligne ; Page disponible le 20-mai-2013].

[97] Wikipédia. Théorème des suites adjacentes — wikipédia, l'encyclopédie libre, 2018. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_des_suites_adjacentes&oldid=148318442. [En ligne ; Page disponible le 10-mai-2018].

[98] Mathieu Mansuy. Autour des séries alternées. URL http://www.mathieu-mansuy.fr/pdf/ECS2-complément1.pdf.

[99] Claessens Laurent. Définition de la limite. 2018. URL https://fr.wikipedia.org/wiki/Discussion:Espace_topologique#D%C3%A9finition_de_la_limite.

[100] Wikipédia. Connexité (mathématiques) — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Connexit%C3%A9_(math%C3%A9matiques)&oldid=123744973. [En ligne ; Page disponible le 26-avril-2016].

[101] C. Antonini, JF. Quint, P. Borgnat, J. Bérard, E. Lebeau, E. Souche, A. Chateau, and O. Teytaud. Connexité. . URL http://www.les-mathematiques.net/a/a/b/node23.php.

[102] Huiqiang Jiang. Functional analysis. URL http://www.math.pitt.edu/~hqjiang/2303/functional.pdf.

[103] Nicolas Bourbaki. General topology 2 chapters 5 - 10. URL https://books.google.be/books?id=bQwhdmL6IjUC. Oui je sais c'est une honte de citer Bourbaki en anglais en pointant vers une version partielle disponible sur Googlebooks. Mais c'est surtout une honte que ce livre ne soit pas disponible gratuitement en version électronique.

[104] Guillaume Carlier. Analyse fonctionnelle. 2009-20010. URL https://www.ceremade.dauphine.fr/~carlier/poly2010.pdf.

[105] Raz Kupferman. Topological vector spaces. URL http://www.ma.huji.ac.il/~razk/iWeb/My_Site/Teaching_files/TVS.pdf.

[106] Ron Freiwald. Chapter iv completeness and compactness. URL http://www.math.wustl.edu/~freiwald/ch4.pdf.

[107] Wikipédia. Suite de cauchy — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Suite_de_Cauchy&oldid=98046451. [En ligne ; Page disponible le 11-mars-2014].

[108] Wikipédia. Continuité — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Continuit%C3%A9&oldid=96548980. [En ligne ; Page disponible le 17-septembre-2013].

[109] Wikipédia. Théorème des fermés emboités — Wikipédia, l'encyclopédie libre, 2012. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_des_ferm%C3%A9s_emboit%C3%A9s&oldid=85987870. [En ligne ; Page disponible le 1-avril-2013].

[110] C. Antonini, JF. Quint, P. Borgnat, J. Bérard, E. Lebeau, E. Souche, A. Chateau, and O. Teytaud. Le théorème de tykhonov. . URL http://www.les-mathematiques.net/a/a/b/node20.php.

[111] Wikipédia. Équicontinuité — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=%C3%89quicontinuit%C3%A9&oldid=137910699. [En ligne ; Page disponible le 4-juin-2017].

[112] Zied Ammari. Analyse fonctionnelle : Pré-requis. http://perso.univ-rennes1.fr/zied.ammari/other/pdf/chapitre1.pdf.

[113] Jean Saint Raymond. Semi-normes. URL http://www.math.jussieu.fr/~raymond/preprints/seminormes.pdf.

[114] Stéphane Mischler. Semi-norme et introduction aux evtlcs. Février 2007. URL https://www.ceremade.dauphine.fr/~mischler/Enseignements/AFAENS/Chap1evtlcs.pdf.

[115] Wikipédia. Semi-norme — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Semi-norme&oldid=89981857. [En ligne ; Page disponible le 22-septembre-2013].

[116] Wikipédia. Théorème de baire — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Baire&oldid=95077621. [En ligne ; Page disponible le 4-août-2013].

[117] Wikipédia. Hermitien — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Hermitien&oldid=124476255. [En ligne ; Page disponible le 18-mars-2016].

[118] Daniel Li. Espaces de hilbert. . http://www.editions-ellipses.fr/product_info.php?products_id=11183 http://www.editions-ellipses.fr/product_info.php?products_id=9387.

[119] Frank Jones. Chapter 7 : cross product. 2004.

[120] Emmanuel Vieillard Baron. Déterminant d'une matrice, d'une application linéaire. 2001. http://www.les-mathematiques.net/b/e/d/node5.php.

[121] Robert Rolland. Le théorème de Müntz-szász. . http://megamaths.perso.neuf.fr/rr/fichexo_201.pdf.

[122] Wikipédia. Résultant — wikipédia, l'encyclopédie libre, Mai 2013. URL http://fr.wikipedia.org/w/index.php?title=R%C3%A9sultant&oldid=90109864. [En ligne ; Page disponible le 24-mai-2013].

[123] Jean-Etienne Rombaldi. Polynômes d'endomorphismes en dimension finie. applications. . http://www-fourier.ujf-grenoble.fr/~rombaldi/AgregInterne/Oral1/110.pdf.

[124] Matthieu Romagny. Endomorphismes cycliques. 2012. URL https://perso.univ-rennes1.fr/matthieu.romagny/agreg/dvt/endom_cycliques.pdf.

[125] Callus. URL http://math.stackexchange.com/questions/1829332/prove-that-if-v-is-finite-dimensional-then-v-is-even-dimensional.

[126] La taverne de l'Irlandais. De la réduction des endomorphismes. 2001. URL http://tanopah.jo.free.fr/epilogues/reduction.pdf.

[127] Wikipédia. Trigonalisation — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Trigonalisation&oldid=109085753. [En ligne ; Page disponible le 22-décembre-2014].

[128] Henry C. King. Unitary diagonalization of matrices. http://www-users.math.umd.edu/~hck/Normal.pdf.

[129] David Delaunay. Nilpotence. 10 juillet 2014. URL http://mp.cpgedupuydelome.fr/pdf/Réductiondesendomorphismes-Nilpotence.pdf.

[130] G. Donald Allen. Lectures on linear algebras. 2004. http://www.math.tamu.edu/~dallen/m640_03c/readings.htm.

[131] Wikipedia. Spectral theorem — wikipedia, the free encyclopedia, 2013. URL http://en.wikipedia.org/w/index.php?title=Spectral_theorem&oldid=575488135. [Online ; accessed 21-October-2013].

[132] theorem for normal triangular matrices. URL http://planetmath.org/theoremfornormaltriangularmatrices.

[133] Michel Granget. Normes matricielles, conditionnement. URL http://math.univ-angers.fr/~granger/ananum/Chapitre_II.pdf.

[134] Maxime Pouvreau. Pseudo-réduction simultanée. 2013-2014. URL http://minerve.bretagne.ens-cachan.fr/images/Pseudored.pdf.

[135] Robert Rollan. Produit tensorile d'espaces vectoriels. URL http://robert.rolland.acrypta.com/telechargements/algebre/tensor.pdf.

[136] Marc Sage. Algèbre multilinéaire. . URL http://www.normalesup.org/~sage/Cours/ProdTens.pdf.

[137] Ministère de l'éducation nationale. Rapport de jury de concours – agrégation de mathématiques, concours externe. 2011. http://agreg.org/Rapports/rapport2011provisoire.pdf.

[138] Hervé Carrieu, Maurice Fadel, Etienne Fieux, Patrice Lassère, and Frédéric Rodriguez. Autour des matrices de Frobenius ou compagnon. 2007. http://www.math.univ-toulouse.fr/~lassere/pdf/vfcomp.pdf.

[139] Grégory Vial. Autour du théorème des invariants de similitude. Octobre 2005. http://w3.bretagne.ens-cachan.fr/math/people/gregory.vial/files/cplts/ivs.pdf.

[140] Arnaud Moncet. Invariants de similitude. http://blogperso.univ-rennes1.fr/arnaud.moncet/public/IVS.pdf.

[141] Commutant d'un endomorphisme. . URL http://minerve.bretagne.ens-cachan.fr/images/Commutant.pdf.

[142] Bernard Alken. Dans un commentaire de « dimension du commutant d'une matrice ». 2009. URL http://www.mathoman.com/index.php/1538-dimension-du-commutant-d-une-matrice.

[143] Wikipédia. Théorème de burnside (problème de 1902) — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Burnside_(probl%C3%A8me_de_1902)&oldid=98510166. [En ligne ; Page disponible le 2-novembre-2015].

[144] Pierre Monmarché. Développements. 18 mai 2011. URL http://laurent.claessens-donadello.eu/pdf/1-total.pdf.

[145] Christian Squarcini. Dégénérescence. 2005. URL http://christian-squarcini.pagesperso-orange.fr/AgregInterne/Algebrelineaire/2_2.pdf.

[146] William F. Trench. Introduction to real analysis. 2010. http://ramanujan.math.trinity.edu/wtrench/texts/TRENCH_REAL_ANALYSIS.PDF.

[147] Wikipédia. Norme d'opérateur — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Norme_d%27op%C3%A9rateur&oldid=125963366. [En ligne ; Page disponible le 26-avril-2017].

[148] Wikipedia. Spectral radius — wikipedia, the free encyclopedia, 2017. URL https://en.wikipedia.org/w/index.php?title=Spectral_radius&oldid=758072091. [Online ; accessed 26-April-2017].

[149] Pierre-Emmanuel Jabin. Analyse numérique, correction du TD 8. 2008-2009. URL http://math.unice.fr/~jabin/CTD3-8.pdf.

[150] Christian Squarcini. Applications linéaires. . URL http://christian-squarcini.pagesperso-orange.fr/AgregInterne/Topo/4_3.pdf.

[151] Espaces métriques et espaces normés. URL http://www.les-mathematiques.net/a/a/b/node4.php.

[152] Gilles Costantini. Espaces vectoriels de dimension infinie, normes usuelles, équivalence de normes. . URL http://gilles.costantini.pagesperso-orange.fr/agreg_fichiers/evn.pdf.

[153] Gallouët !Thierry. Norme et conditionnement d'une matrice. URL https://www.i2m.univ-amu.fr/~gallouet/licence.d/anum.d/anum-tg2.pdf.

[154] David Wilkins. Normed vector spaces and functional analysis. 1997-2001. URL http://www.maths.tcd.ie/~dwilkins/Courses/212/212PtIII.pdf.

[155] Gabriel Nagy. Operator theory in Hilbert spaces. URL https://www.math.ksu.edu/~nagy/real-an/2-07-op-th.pdf.

[156] H. Jerome Keisler. Elementaty calculus, an infinitesimal approach. 2010. http://www.math.wisc.edu/~keisler/keislercalc-810.pdf.

[157] Wikipédia. Suite arithmético-géométrique — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Suite_arithm%C3%A9tico-g%C3%A9om%C3%A9trique&oldid=88168841. [En ligne ; Page disponible le 16-juin-2013].

[158] Wikipedia contributors. Direct sum of modules — Wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Direct_sum_of_modules&oldid=835636392. [Online ; accessed 22-June-2018].

[159] Wikipedia contributors. Tensor product — Wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Tensor_product&oldid=846586202. [Online ; accessed 21-June-2018].

[160] Arjeh Cohen. Tensor product. 2004. URL https://www.win.tue.nl/~amc/ow/lba/lba3.pdf.

[161] Jerome E. Marsden and Tudor S. Ratiu. Manifolds, tensor analysis and applications. 2002. URL http://www.springer.com/gp/book/9780387967905.

[162] F. Laudenbach. *Calcul différentiel et intégral*. les Éd. de l'École polytechnique, 2000. ISBN 9782730207249. URL http://books.google.fr/books?id=Ws9A7ZoRJNEC.

[163] Marie-Claure David, Frédéric Haglund, and Daniel Perrin. Géométrie affine. 8 décembre 2003. http://webens.math.u-psud.fr/~mcld/GAEL/Gael/bar.pdf.

[164] Marie-Claude David, Frédéric Haglund, and Daniel Perrin. Géometrie affine. Document de travail pour la préparation au CAPES. Deuxième partit : barycentre. 8 décembre 2003. URL http://omphale.math.u-psud.fr/~mcld/GAEL/Gael/bar.pdf.

[165] Marie-Claude David, Frédéric Haglund, and Daniel Perrin. Géometrie affine. Document de travail pour la préparation au CAPES. Troisième partie : convexité. 8 décembre 2003. URL http://omphale.math.u-psud.fr/~mcld/GAEL/Gael/conv.pdf.

[166] Jean-Marc Decauwert. Géométrie affine. 8 novembre 2011. URL http://ljk.imag.fr/membres/Bernard.Ycart/mel/ga/ga.pdf.

[167] Wikiversité. Barycentre/théorème de l'associativité du barycentre — wikiversité,, 2016. URL https://fr.wikiversity.org/w/index.php?title=Barycentre/Th%C3%A9or%C3%A8me_de_l%27associativit%C3%A9_du_barycentre&oldid=582092. [En ligne ; accédé le 16-janvier-2017].

[168] Adrien Fontaine. Développement : théorème de Sarkiowski. 8 octobre 2013. URL http://perso.eleves.bretagne.ens-cachan.fr/~afontain/dvptthmdesarkowski.pdf.

[169] Wikipédia. Opérations sur les dérivées — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Op%C3%A9rations_sur_les_d%C3%A9riv%C3%A9es&oldid=99099598. [En ligne ; Page disponible le 10-avril-2016].

[170] URL https://www.math.hmc.edu/calculus/tutorials/quotient_rule/proof.pdf.

[171] Wikiversité. Fonctions d'une variable réelle/dérivabilité — wikiversité,, 2018. URL https://fr.wikiversity.org/w/index.php?title=Fonctions_d%27une_variable_r%C3%A9elle/D%C3%A9rivabilit%C3%A9&oldid=718842. [En ligne ; accédé le 24-mai-2018].

[172] Wikiversité. Fonction dérivée/dérivée d'un quotient — wikiversité,, 2017. URL https://fr.wikiversity.org/w/index.php?title=Fonction_d%C3%A9riv%C3%A9e/D%C3%A9riv%C3%A9e_d%27un_quotient&oldid=675129. [En ligne ; accédé le 24-mai-2018].

[173] Wikipédia. Opérations sur les dérivées — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Op%C3%A9rations_sur_les_d%C3%A9riv%C3%A9es&oldid=99099598. [En ligne ; Page disponible le 28-juillet-2014].

[174] Wikipédia. Théorème de rolle — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Rolle&oldid=132483699. [En ligne ; Page disponible le 28-décembre-2016].

[175] Guillaume Connan. Dérivation. URL http://gconnan.free.fr/les%20pdf/Deriv.pdf.

[176] Keith Konrad. The remainder in Taylor series. URL http://www.math.uconn.edu/~kconrad/blurbs/analysis/TaylorRemainder.pdf.

[177] Wikipédia. Règle de l'hôpital — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=R%C3%A8gle_de_L%27H%C3%B4pital&oldid=131491717. [En ligne ; Page disponible le 29-décembre-2016].

[178] Livio Flaminio. Éléments de géométrie différentielle. 29 octobre 2009. URL http://www-gat.univ-lille1.fr/~flaminio/M403/2009-2010/cours_geo_top.pdf.

[179] G.B. Folland. Taylor's formula. URL https://sites.math.washington.edu/~folland/Math425/taylor.pdf.

[180] Wikipedia. Développement de taylor, 2016. URL http://fr.wikipedia.org/wiki/Développement_de_Taylor.

[181] Wikiversité. Calcul différentiel/théorèmes utiles — wikiversité,, 2017. URL https://fr.wikiversity.org/w/index.php?title=Calcul_diff%C3%A9rentiel/Th%C3%A9or%C3%A8mes_utiles&oldid=657239. [En ligne ; accédé le 10-juillet-2017].

[182] Wikipédia. Convergence uniforme — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Convergence_uniforme&oldid=95493463. [En ligne ; Page disponible le 15-octobre-2013].

[183] Suites et séries de fonctions. . URL http://blog.psi945.fr/public/maths-psi/cours-psi-suites-series-fonctions.pdf.

[184] F. Poupaud. Analyse fonctionnelle pour la licence. URL http://math.unice.fr/~rascle/pdffiles/coursanapp/ana-fonc.pdf.

[185] Rached Mneimé and Frédéric Testard. *Groupes de Lie classiques*. Hermann, 1986. ISBN 2-7056-6040-2.

[186] Yoann Gelineau. Générateurs du groupe linéaire. . URL http://math.univ-lyon1.fr/~gelineau/devagreg/Generateurs_Groupe_Lineaire.pdf.

[187] Sandrine Caruso. Générateurs de $\mathrm{GL}_n(K)$ et $\mathrm{SL}_n(K)$. URL http://sandrine.toonywood.org/pageperso/agreg/geneSL.pdf.

[188] Wikipédia. Théorème de cayley-hamilton — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Cayley-Hamilton&oldid=104470625. [En ligne ; Page disponible le 10-juin-2014].

[189] David Monniaux. 9 juin 2014. URL http://david.monniaux.free.fr/dotclear/index.php/post/2014/06/09/Mémoire-à-effaçage-rapide. David Monniaux esquisse la preuve dans un commentaire à cette note.

[190] Pierre Lissy. Décomposition polaire. 30 avril 2010. URL http://www.ljll.math.upmc.fr/~lissy/Agreg/developpements/DecPol.pdf.

[191] Wikipédia. Décomposition polaire — wikipédia, l'encyclopédie libre, avril 2013. URL http://fr.wikipedia.org/w/index.php?title=D%C3%A9composition_polaire&oldid=90240040. [En ligne ; Page disponible le 24-avril-2013].

[192] Jean-Etienne Rombaldi. Endomorphismes symétriques d'un espace vectoriel. applications. . http://www-fourier.ujf-grenoble.fr/~rombaldi/AgregInterne/Oral1/120.pdf.

[193] Brice Loustau. Développements d'algèbre pour l'agrégation. 2008. URL http://brice.loustau.free.fr/deval.pdf.

[194] Michel Coste. Sous-groupes compacts du groupe linéaire. 29 septembre 2006. URL http://agreg-maths.univ-rennes1.fr/documentation/docs/ssgrpecompact.pdf.

[195] Daniel Li. Notions fondamentales de la théorie des probabilités. . http://labomathlens.free.fr/Liens/ProbaM1/PROBA01.pdf.

[196] Daniel Saada. Tribu de borel et tribu de baire d'un espace topologique. 2009. URL http://www.daniel-saada.eu/fichiers/16-Tribus_de_Baire.pdf.

[197] Daniel Li. Construction de la mesure de lebesgue. 28 javier 2008. http://www.editions-ellipses.fr/product_info.php?products_id=11183 http://www.editions-ellipses.fr/product_info.php?products_id=9387.

[198] Wikipédia. Mesure (mathématiques) — wikipédia, l'encyclopédie libre, 2018. URL http://fr.wikipedia.org/w/index.php?title=Mesure_(math%C3%A9matiques)&oldid=145203435. [En ligne ; Page disponible le 5-février-2018].

[199] ProofWiki. Outer measure of limit of increasing sequence of sets — proofwiki,, 2012. URL http://www.proofwiki.org/w/index.php?title=Outer_Measure_of_Limit_of_Increasing_Sequence_of_Sets&oldid=85952. [Online ; accessed 2-February-2014].

[200] C. Antonini, JF. Quint, P. Borgnat, J. Bérard, E. Lebeau, E. Souche, A. Chateau, and O. Teytaud. Quelques résultats d'unicité. . URL http://www.les-mathematiques.net/a/i/d/node1.php#T42.

[201] C. Antonini, JF. Quint, P. Borgnat, J. Bérard, E. Lebeau, E. Souche, A. Chateau, and O. Teytaud. Ensembles négligeables et complétion de tribus. . URL http://www.les-mathematiques.net/a/i/c/node1.php.

[202] Wikipédia. Complétion d'une mesure — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Compl%C3%A9tion_d%27une_mesure&oldid=89819650. [En ligne ; Page disponible le 27-mars-2014].

[203] Daniel Li. Tribus et mesures. 24 mars 2011. http://www.editions-ellipses.fr/product_info.php?products_id=11183 http://www.editions-ellipses.fr/product_info.php?products_id=9387.

[204] Gerald Teschl. Topics in real and functional analysis. 2013. URL http://www.mat.univie.ac.at/~gerald/ftp/book-fa/fa.pdf.

[205] Emmanuel Vieillard Baron. Les fonctions mesurables. 2001. URL http://www.les-mathematiques.net/a/i/b/node1.php.

[206] Wikipédia. Fonction étagée — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Fonction_%C3%A9tag%C3%A9e&oldid=104836104. [En ligne ; Page disponible le 6-septembre-2015].

[207] B. Mauray. Intégration jusqu'au théorème de Lebesgue. 2006. URL https://webusers.imj-prg.fr/~bernard.maurey/agreg/Textes/Lebesgue.pdf.

[208] Daniel Choï. Mathématiques pour la mécanique. 2008. http://www.meca.unicaen.fr/Enseignement/Document/CoursChoi/mathmeca.pdf.

[209] Gilles Dubois. URL http://dubois.gilles.pagesperso-orange.fr/analyse_reelle/intlimites.html.

[210] Daniel Li. Construction de l'intégrale de lebesgue. 10 février 2011. http://www.editions-ellipses.fr/product_info.php?products_id=11183 http://www.editions-ellipses.fr/product_info.php?products_id=9387.

[211] Yao Jianfeng. Statistique et logiciels. 2010. http://perso.univ-rennes1.fr/jian-feng.yao/pedago/statlog/.

[212] Daniel Li. Théorème de radon-nikodým et applications. . http://www.editions-ellipses.fr/product_info.php?products_id=11183 http://www.editions-ellipses.fr/product_info.php?products_id=9387.

[213] Bruce Driver. Analysis tools with applications. April 10, 2003. URL http://www.math.ucsd.edu/~bdriver/240-01-02/Lecture_Notes/anal.pdf.

[214] Daniel Li. Intégration sur un espace produit. 29 mars 2011. http://www.editions-ellipses.fr/product_info.php?products_id=11183 http://www.editions-ellipses.fr/product_info.php?products_id=9387.

[215] B. Mauray. Préparation à l'agrégation, analyse. 2011. http://www.math.jussieu.fr/~maurey/agreg/index.html.

[216] Thierry Gallouët and Raphaèle Herbin. Mesure, intégration, probabilités. 2012. http://www.cmi.univ-mrs.fr/~gallouet/licence.d/mes-int-pro.pdf.

[217] Wikipédia. Ensemble de vitali — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Ensemble_de_Vitali&oldid=100925109. [En ligne ; Page disponible le 9-septembre-2015].

[218] John Hunter. Measure theory. 2011. URL https://www.math.ucdavis.edu/~hunter/measure_theory/measure_notes.pdf.

[219] Bruno Demange. Théorème de changement de variables. 2012. URL https://www-fourier.ujf-grenoble.fr/~demange/integration/2012/integration-chap5.pdf.

[220] Daniel Li. Changement de variables dans les intégrales sur un ouvert de $\mathbb{R}^n$. 8 juin 2010. URL http://li.perso.math.cnrs.fr/textes/Integration/change-v.pdf.

[221] Wikipédia. Intégrale impropre — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Int%C3%A9grale_impropre&oldid=103902526. [En ligne ; Page disponible le 5-août-2014].

[222] Giuseppe De Marco. *Analisi due*. 1999.

[223] romain Boillaud. Séries de fonctions. 2010-2011. URL http://www.rblld.fr/cours/chap6.pdf.

[224] Gérard Eguether. Ex-séries entières. URL http://iecl.univ-lorraine.fr/~Gerard.Eguether/zARTICLE/EX.pdf.

[225] Georges Skandalis. Analyse, résumés et exercices. 6 mars 2015. URL https://www.math.univ-paris-diderot.fr/_media/formations/prepa/agreginterne/polycopieanalyse.pdf.

[226] Rahul Krishna. Approximations by taylor polynomials. URL http://math.columbia.edu/~krishna/08092011.pdf.

[227] Vogel Thomas. Analytic functions. 2012. URL http://www.math.tamu.edu/~tvogel/410/sect74a.pdf.

[228] ENS Cachan. Théorème de stabilité de Lyapunov. URL http://minerve.bretagne.ens-cachan.fr/images/Lyapunov.pdf.

[229] Rached Mneimé. *Réduction des endomorphismes*. Calvage et Mounet, 2006. ISBN 2-916352-01-5.

[230] Matthieu Romagny. Représentations linéaires des groupes finis. 2010-2011. http://www.math.jussieu.fr/~romagny/agreg/theme/rlgf_paysage.pdf.

[231] Michael Gechele. Théorème de stone-weierstrass. http://michael.gechele.perso.neuf.fr/Agregation/Theoreme_Stone_Weierstrass.pdf.

[232] Primitives et intégrales. URL http://math.univ-lyon1.fr/capes/IMG/pdf/new.primitive.pdf.

[233] Sébastien Pellerin. Développements d'analyse pour l'oral de l'agrégation. . http://pellerin.xyz/doc/agreg/analyse.pdf.

[234] Kevin Quirin. Théorème de Müntz. http://kevin.quirin.free.fr/Trucs/agreg/Dvt_Muntz.pdf.

[235] Pierre Lissy. Déterminant de cauchy et application au théorème de Müntz. 5 mai 2010. http://www.ljll.math.upmc.fr/~lissy/Agreg/developpements/CauchyMuntz.pdf.

[236] Benjamin Dadoun. Théorème de Müntz. URL http://benjamin.dadoun.free.fr/muntz.pdf.

[237] Wikipédia. Lemme de hadamard — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Lemme_de_Hadamard&oldid=90659396. [En ligne ; Page disponible le 14-juillet-2014].

[238] Jean-Pierre Demailly. Analyse numérique et équations différentielles. 1991.

[239] Clémence Minazzo and Kelsey Rider. Théorèmes du point fixe et applications aux équations différentielles. 2006-2007. http://math.unice.fr/~eaubry/Enseignement/M1/memoire.pdf.

[240] Ivan Nourdin. *Leçons d'analyse, probabilités algèbre et géométrie*. Dunod, 2001. ISBN 2-10-005668-9.

[241] Sandrine Caruso. Théorème de Cauchy-Lipschitz. 2009. http://boumbo.toonywood.org/sandrine/pageperso/agreg/cauchy-lipschitz.pdf Voir aussi la page toute pleine de développement http://boumbo.toonywood.org/sandrine/pageperso/agreg.html.

[242] Franck Boyer. Équations différentielles ordinaires. 2012. URL http://www.latp.univ-mrs.fr/~fboyer/Enseignement/Agreg/cours_EDO_Agreg_FBoyer.pdf.

[243] Théo Pierron. Théorème de Cauchy-Lipschitz global. 5 janvier 2014. URL http://perso.eleves.ens-rennes.fr/~tpier758/agreg/dvpt/maths/cauchy_lipschitz.pdf.

[244] Thierry Audibert. Le problème de Cauchy, résultats fondamentaux. URL http://www.univenligne.fr/Documents/PolycopiesAgreg/EDO1.pdf.

[245] Thomas Budzinski. Théorème de Cauchy-Lipschitz. 2014. URL https://www.eleves.ens.fr/home/budzinsk/polys/Nonolympique/2014_cauchylipschitz.pdf.

[246] Wikipédia. Théorème d'inversion locale — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_d%27inversion_locale&oldid=100975433. [En ligne ; Page disponible le 19-février-2014].

[247] Antoine Chambert-Loir. Introduction aux groupes et algèbres de Lie. 2004-2005. URL http://www.math.u-psud.fr/~chambert/enseignement/2004-05/lie/lie.pdf.

[248] E.P. van den Ban. Lie groups. 2003. http://www.math.uu.nl/people/ban/lecnot.html.

[249] Joachim Stubbe. URL http://mathaa.epfl.ch/prob/enseignement/analyse2/series/AnalysisE.pdf.

[250] Laurent Guillopé. Optimisation sous contrainte. 2015-2016. URL https://www.math.sciences.univ-nantes.fr/~guillope/l3-osc/osc.pdf.

[251] Bruno Galerne. Optimisation, algorithmique. 2016-2017. URL http://www.math-info.univ-paris5.fr/~bgalerne/m1_opti_algo/poly_opti_algo.pdf.

[252] Wikipédia. Fonction convexe — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Fonction_convexe&oldid=94429550. [En ligne ; Page disponible le 8-août-2013].

[253] Wikipédia. Fonction convexe — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Fonction_convexe&oldid=98971686. [En ligne ; Page disponible le 8-février-2014].

[254] Wikiversité. Fonctions convexes/définition et premières propriétés — wikiversité,, 2013. URL http://fr.wikiversity.org/w/index.php?title=Fonctions_convexes/D%C3%A9finition_et_premi%C3%A8res_propri%C3%A9t%C3%A9s&oldid=381775. [En ligne ; accédé le 9-février-2014].

[255] Wikiversité. Fonctions convexes/fonctions convexes dérivables — wikiversité,, 2013. URL http://fr.wikiversity.org/w/index.php?title=Fonctions_convexes/Fonctions_convexes_d%C3%A9rivables&oldid=379779. [En ligne ; accédé le 9-février-2014].

[256] Stephen Boyd and Lieven Vandenberghe. Convex optimization. 2004. URL http://www.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf.

[257] Document 33 : fonctions convexes. . URL http://math.univ-lyon1.fr/capes/IMG/pdf/new.convexe.pdf.

[258] Wikipédia. Inégalité arithmético-géométrique — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=In%C3%A9galit%C3%A9_arithm%C3%A9tico-g%C3%A9om%C3%A9trique&oldid=99682278. [En ligne ; Page disponible le 1-juillet-2014].

[259] Wikipédia. Inégalité de kantorovitch — wikipédia, l'encyclopédie libre, juillet 2014. URL http://fr.wikipedia.org/w/index.php?title=In%C3%A9galit%C3%A9_de_Kantorovitch&oldid=102417007. [En ligne ; Page disponible le 1-juillet-2014].

[260] Georges Comte. Lemme de Morse. 2009-2010. URL http://gc83.perso.sfr.fr/Agregation/LemmedeMorse.pdf.

[261] Louis Fauchier-Magnan. Introduction à la théorie de Morse. URL http://cqfd.epfl.ch/webdav/site/cqfd/shared/projets/igat/IntroductionàlathéoriedeMorse-LouisFauchier-Magnan.pdf.

[262] Michel Rascle. Analyse fonctionnelle pour la licence. http://math.unice.fr/~rascle/pdffiles/coursanapp/ana-fonc.pdf.

[263] B. Mauray. Analyse fonctionnelle et théorie spectrale (version longue). 2001–2002. http://www.math.jussieu.fr/~maurey/ts012/poly/lths.pdf.

[264] Mourad Besbes. Chapitre 5 : espaces métriques complets, espaces de banach. http://besbes.mourad.free.fr/Enseignement/Topologie/Chapitre5.pdf.

[265] C. Antonini, JF. Quint, P. Borgnat, J. Bérard, E. Lebeau, E. Souche, A. Chateau, and O. Teytaud. Complété d'un espace métrique. . URL http://www.les-mathematiques.net/a/a/b/node26.php.

[266] Wikipédia. Espace complet — Wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Espace_complet&oldid=91125353. [En ligne ; Page disponible le 1-avril-2013].

[267] Transformation laissant invariante une forme quadratique. URL http://www.ilemaths.net/forum-sujet-500814.html.

[268] Éric Jourdain. Angles. URL https://perso.univ-rennes1.fr/eric.jourdain/GEEU/Cours/Angles.pdf.

[269] David Cimasoni. Cours de géométrie I, semestre de printemps. Printemps 2014. URL http://www.unige.ch/math/folks/cimasoni/GeometrieI.pdf.

[270] Christophe Mourougane. Théorie des groupes et géométrie. 2008-2009. http://perso.univ-rennes1.fr/christophe.mourougane/enseignements/2008-9/THGG/poly.groupes.pdf.

[271] Philippe Picart. Définition de cos et sin par les séries entières. URL http://trucsmaths.free.fr/telech/exp_cos_sin.pdf.

[272] Alfred Gray. Modern differential geometry of curves and surfaces with mathematica. 1998. URL http://webmath2.unito.it/paginepersonali/sergio.console/CurveSuperfici/.

[273] Wikipédia. Pi — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Pi&oldid=133197907. [En ligne ; Page disponible le 3-janvier-2017].

[274] Daniel Perrin. Isométries du plan. URL https://www.math.u-psud.fr/~perrin/CAPES/geometrie/isometries11-12.pdf.

[275] Groupes diédraux. http://theoriedesgroupes.perso.sfr.fr/cours/Diedraux.pdf.

[276] Pierre Lissy. 6 avril 2010. URL http://www.ljll.math.upmc.fr/~lissy/Agreg/developpements/Died.pdf.

[277] Pierre Renfer. Roulettes et colliers. http://les.mathematiques.free.fr/pdf/collier.pdf.

[278] Wikipédia. Angle — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Angle&oldid=130370952. [En ligne ; Page disponible le 15-novembre-2016].

[279] B. Guesmi. Théorème de l'angle inscrit. URL http://s999cdfd874dd6835.jimcontent.com/download/version/1427398317/module/5509391568/name/theoremedel'arccapable.pdf1.pdf.

[280] Frédéric Hélein. Géométrie euclidienne. 2005-2006. URL https://webusers.imj-prg.fr/~frederic.helein/cours/euclide.pdf.

[281] Laurent Dietrich. Les sous-groupes finis de so(3). 16 mai 2011. URL http://www.math.cmu.edu/~ldietric/doc/ssgrso3.pdf.

[282] Sous groupes finis de so(3,r). 2013-2014. URL http://minerve.bretagne.ens-cachan.fr/images/Sous-groupes_finis_SO3.pdf.

[283] Marc Sage. Sons, fréquences, harmoniques, tons : le compromis du piano. . http://www.normalesup.org/~sage/Musique/Harmonie.pdf.

[284] Wikipédia. Chiffrement rsa — wikipédia, l'encyclopédie libre, 2018. URL http://fr.wikipedia.org/w/index.php?title=Chiffrement_RSA&oldid=146774926. [En ligne ; Page disponible le 25-mars-2018].

[285] URL https://www.torproject.org/about/torusers.html.en.

[286] URL https://www.torproject.org/projects/torbrowser.html.en.

[287] Wikipédia. Polynôme cyclotomique — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Polyn%C3%B4me_cyclotomique&oldid=93751495. [En ligne ; Page disponible le 25-juin-2013].

[288] Adam Parusinski. Théorème de Dirichlet (version faible). 2010-2011. http://math.unice.fr/~parus/AGREG/Dev/dirichlet.pdf.

[289] Michel Merle. corps finis. 2005-2006. http://www-math.unice.fr/~merle/Complements/corps_finis.pdf.

[290] Gabriel Peyré. Corps finis. 2007. https://objectifagregation.github.io/.

[291] François Rodier. Corps finis. http://iml.univ-mrs.fr/~rodier/Cours/RappelCorpsfinis.pdf.

[292] Stephane Vento. La loi de réciprocité quadratique. 2005. http://www.proba.jussieu.fr/pageperso/nourdin/LeSiteDeLAgregatif/vento2.pdf.

[293] Wikipedia contributors. Chevalley–warning theorem — Wikipedia, the free encyclopedia, 2017. URL https://en.wikipedia.org/w/index.php?title=Chevalley%E2%80%93Warning_theorem&oldid=792119700. [Online ; accessed 21-May-2018].

[294] Christian Squarcini. Anneaux factoriels. . URL http://christian-squarcini.pagesperso-orange.fr/AgregInterne/Anneauxcorps/3_2.pdf.

[295] Wikipédia. Fonction de möbius — wikipédia, l'encyclopédie libre, mai 2013. URL http://fr.wikipedia.org/w/index.php?title=Fonction_de_M%C3%B6bius&oldid=91236321. [En ligne ; Page disponible le 1-mai-2013].

[296] Wikipédia. Corps fini — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Corps_fini&oldid=92489466. [En ligne ; Page disponible le 1-mai-2013].

[297] Wikipédia. Nombre constructible — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Nombre_constructible&oldid=103130793. [En ligne ; Page disponible le 13-mai-2014].

[298] Tracés usuels « à la règle et au compas ». 2012-2013. URL http://dokeos3.u-cergy.fr/dokeos20/courses/EEMEM2CRPE/scorm/2012_2013SCORMEEMEUE3EC2M2P1_SCORM/page_04.htm.

[299] Wikipédia. Théorème de wantzel — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Wantzel&oldid=103220511. [En ligne ; Page disponible le 13-mai-2014].

[300] Wikipédia. Nombre de fermat — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Nombre_de_Fermat&oldid=103080861. [En ligne ; Page disponible le 18-juin-2014].

[301] Philippe Spindel. Éléments de géométrie différentielle pour la mécanique analytique et le gravitation. Novembre 2010. https://portail.umons.ac.be/EN2/universite/facultes/fs/services/institut_physique/mecanique_et_gravitation/Documents/Textespédagogiques/geodiff.pdf.

[302] William K Allard. The Brouwer fixed point theorem. 2004. http://www.math.duke.edu/~wka/math204/fixed.pdf see also http://www.math.duke.edu/~wka/math204/.

[303] Kenneth Kuttler. Topics in analysis. Janvier 2012. http://www.math.byu.edu/~klkuttle/SobolevSpacesB.pdf Juste un indice : ce document fait 2086 pages . . . j'y arriverais, courage !

[304] Hervé Le Dret. Notes de cours M2 Équations aux dérivées partielles élliptiques. 4 mars 2010. URL https://www.ljll.math.upmc.fr/~ledret/M2Elliptique/chapitre4.pdf.

[305] Bachir Bekka. Théorème du point fixe de Markov-Kakutani : applications à l'existence de mesures de haar et aux chaînes de Markov. 2005. http://agreg-maths.univ-rennes1.fr/documentation/docs/PointFixe.pdf.

[306] Olivier Casterá. Différentielles totales exactes. http://o.castera.free.fr/pdf/Differentielle_totale_exacte.pdf.

[307] Intégration de fractions rationnelles. URL http://blascheck.franck.free.fr/IMG/pdf/IntegFracRat.pdf.

[308] Ophélie Rouby. Théorème de Rothstein-Trager. URL http://perso.eleves.bretagne.ens-cachan.fr/~oroub842/dvlpts/Rothstein_Trager.pdf.

[309] Frédéric Chyzak. Intégration symbolique des fractions rationnelles. Décembre 2004. URL http://www.enseignement.polytechnique.fr/informatique/profs/Frederic.Chyzak/CF-2004/notes-IntRat.pdf.

[310] gb. Cercle circonscrit à une courbe. URL http://www.les-mathematiques.net/phorum/read.php?8,750556,750586.

[311] Mohammad Ghomi. Lecture note on differential geometry. 2007. URL http://people.math.gatech.edu/~ghomi/LectureNotes/LectureNotes5U.pdf.

[312] Wikipedia. Convex curve — wikipedia, the free encyclopedia, 2015. URL https://en.wikipedia.org/w/index.php?title=Convex_curve&oldid=685739052. [Online ; accessed 13-April-2016].

[313] Wikipédia. Théorème de jordan — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Jordan&oldid=124990679. [En ligne ; Page disponible le 6-avril-2016].

[314] Les quadrilatères. URL http://ddata.over-blog.com/xxxyyy/4/39/08/01/cinquieme/5eme.parallelogramme.jeu-set-et-maths.mathematiques.pdf.

[315] Robert Rolland. Géométrie projective. . http://robert.rolland.acrypta.com/telechargements/geometrie/pr0.pdf.

[316] Daniel Bertrand. Géométrie projective. 2009-2010. http://www.math.jussieu.fr/~bertrand/Enseignement/epdf/Algeo09_chap2.pdf.

[317] Nicolas Jacon. Géométrie projective. URL http://njacon.perso.math.cnrs.fr/jacon_Geometrieproj.pdf.

[318] François Labourie. Géométrie affine et projective. 9 février 2019. URL http://math.unice.fr/~labourie/preprints/pdf/geomproj.pdf.

[319] Matthieu Romagny. Droites projectives, homographies. 2005-2006. URL https://perso.univ-rennes1.fr/matthieu.romagny/agreg/droite_projective_Halberstadt.pdf.

[320] Arnaud Bodin. L'inversion. 2012. URL http://math.univ-lille1.fr/~bodin/geometrie/ch_inversion.pdf.

[321] Laurent Dietrich. Le groupe circulaire. 21 mai 2011. URL http://www.math.cmu.edu/~ldietric/doc/groupecirc.pdf.

[322] Florian Bouget. Le groupe circulaire. URL http://florian.bouguet.free.fr/doc/developpements/Groupe_circulaire.pdf.

[323] Florian Bouguet. Le groupe circulaire. URL http://minerve.bretagne.ens-cachan.fr/images/Groupe_circulaire.pdf.

[324] 182 : utilisation des nombres complexes en géométrie. Homographies. 2015-2016. URL http://minerve.bretagne.ens-cachan.fr/images/182_2015-2016.pdf.

[325] Sébastien Pellerin. Action du groupe modulaire sur le demi-plan de Poincaréé. . http://dynamaths.free.fr/docs/lecons/developpement_algebre_6.pdf.

[326] Alexandre Alessandri. Action du groupe modulaire sur le demi-plan de Poincaré. 2011-2012. http://minerve.bretagne.ens-cachan.fr/images/Dvt_psl2.pdf.

[327] Aline Kurtzmann. Espérance conditionnelle. http://www.iecn.u-nancy.fr/~kurtzman/cours/agreg/esperance-conditionnelle.pdf.

[328] Wikipédia. Théorème de projection sur un convexe fermé — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_projection_sur_un_convexe_ferm%C3%A9&oldid=90246160. [En ligne ; Page disponible le 3-mai-2013].

[329] Michel Merle. Espaces de Hilbert. 2004-2005. URL http://www-math.unice.fr/~merle/Algebre_et_geometrie/hilbert.pdf.

[330] Douglas N. Arnold. Functional analysis. 1997. URL http://www.math.umn.edu/~arnold/502.s97/functional.pdf.

[331] Wikipédia. Base de hilbert. . URL https://fr.wikipedia.org/wiki/Base_de_Hilbert?oldid=128533376.

[332] Christiane Schomblond. *Théorie quantique des champs, QED, QCD*. Notes de cours, en français. http://homepages.ulb.ac.be/~cschomb/intfonc+QED+QCD.pdf.

[333] Remsirems. Re : Réponses à quelques questions. 2016. URL http://linuxfr.org/nodes/110155/comments/1675813.

[334] Jonathan Bain. Kochen-Specker and measurement. 2011. http://faculty.poly.edu/~jbain/philqm/lectureslides/08.KS&Measurement.pdf.

[335] Michel Kern. Introduction à la méthode des éléments finie. 2004-2005. URL http://mms2.ensmp.fr/ef_paris/formulation/polycop/f_coursEF.pdf.

[336] Antoine Bensalah. Communication électronique privée. 2017.

[337] Wikipédia. Théorème d'Ascoli — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_d%27Ascoli&oldid=92958118. [En ligne ; Page disponible le 11-mai-2013].

[338] Wikipédia. Théorème de banach-steinhaus — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Banach-Steinhaus&oldid=92377069. [En ligne ; Page disponible le 31-août-2013].

[339] B. Maurey. 9 mars 2011. http://www.math.jussieu.fr/~maurey/IntegProba/Cours/CoursIP101-14.pdf.

[340] Danarmk. 2016. URL http://linuxfr.org/nodes/110155/comments/1675589.

[341] B. Mauray. Convolution, inégalités, approximations et régularisation. 2007. URL https://webusers.imj-prg.fr/~bernard.maurey/agreg/Textes/Convolution.pdf.

[342] Schilling. Minkowsky's inequality for integrals. URL http://math.ucsd.edu/~lni/math240/suppl-1.pdf.

[343] Charles Suquet. Intégration, analyse de fourier,probabilités. 2003-2004. http://math.univ-lille1.fr/~suquet/ens/IFP/Cours/cours04/CoursIFP04.html.

[344] Wikipédia. Théorème de Riesz-Fischer — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Riesz-Fischer&oldid=90074115. [En ligne ; Page disponible le 3-mai-2013].

[345] Gilles Leborgne. Introduction à la théorie des distributions. 16 mai 2013. URL http://www.isima.fr/~leborgne/IsimathDistributions/distrib.pdf.

[346] Olof Hanner. On the uniform convexity of l^p and l^p. 1955. URL https://projecteuclid.org/download/pdf_1/euclid.afm/1485893271.

[347] Wikipédia. Espace Lp — wikipédia, l'encyclopédie libre, 2013. URL http://fr.wikipedia.org/w/index.php?title=Espace_Lp&oldid=92618226. [En ligne ; Page disponible le 10-mai-2013].

[348] Didier Smets. Base d'analyse fonctionnelle. 2011-2012. URL https://www.ljll.math.upmc.fr/smets/MM005/MM005_Chapitre_6.pdf.

[349] Haïm Brézis. *Analyse fonctionnelle*. Masson, Paris, 1983. ISBN 2-225-77198-7. Théorie et applications.

[350] MoebiusCorzer Giraudo, Davide. Show smooth functions of compact support are dense in the Schwartz space. 2012. URL http://math.stackexchange.com/questions/229751/show-smooth-functions-of-compact-support-are-dense-in-the-schwartz-space.

[351] B. Maurey. Holomorphie. 7 février 2005. http://www.math.jussieu.fr/~maurey/agreg/Textes/Holomorphe045.pdf.

[352] Pascal Dingoyan. Chapitre 5. URL http://www.math.jussieu.fr/~dingoyan/lescours/lm367-pageweb/lm367-chapitre5.pdf.

[353] Wikipédia. Anneau principal — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Anneau_principal&oldid=141994549. [En ligne ; Page disponible le 27-octobre-2017].

[354] Corentin. Re : formule de cauchy et dérivation. URL http://www.les-mathematiques.net/phorum/read.php?4,385272,385289.

[355] Philippe Charpentier. Analyse complexe. Septembre 2010. http://www.math.u-bordeaux1.fr/~pcharpen/enseignement/fichiers-master1/Analyse_Complexe.pdf.

[356] Christian Kkein. Séries de laurent, résidus. URL http://math.u-bourgogne.fr/IMB/klein/Complement_dAnalyse_files/chap607.pdf.

[357] Romain Boillaud. Séries entières. http://www.rblld.fr/cours/chap8.pdf.

[358] Wikipédia. Intégrale de fresnel — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Int%C3%A9grale_de_Fresnel&oldid=132170602. [En ligne ; Page disponible le 26-novembre-2016].

[359] David Sirajuddin. Fresnel integrals. 2008. http://itcanbeshown.com/integrals/FresnelIntegrals/fresnel_integrals.pdf.

[360] Noach Dana-Picard. The logarithm of a complex number. URL http://ndp.jct.ac.il/tutorials/complex/node26.html.

[361] Guillaume Carlier. Analyse complexe. 2012-2013. URL https://www.ceremade.dauphine.fr/~carlier/analysecomplexe.pdf.

[362] Wikiversité. Fonctions d'une variable complexe/le logarithme complexe — wikiversité,, 2016. URL https://fr.wikiversity.org/w/index.php?title=Fonctions_d%27une_variable_complexe/Le_logarithme_complexe&oldid=617412. [En ligne ; accédé le 16-décembre-2016].

[363] Jean-Marie Lion. Fonctions holomorphes d'une variable. 5 mai 2004. http://perso.univ-rennes1.fr/jean-marie.lion/coursholo.pdf.

[364] Richard Gomez. Séries de Fourier. 10 octobre 2010. URL http://megamaths.perso.neuf.fr/rg/fourier.pdf.

[365] David Delaunay. Séries de Fourier. 8 juin 2012. http://mp.cpgedupuydelome.fr/pdf/SériesdeFourier.pdf.

[366] Wikipédia. Théorème de jordan — wikipédia, l'encyclopédie libre, Mai 2013. URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Jordan&oldid=92525205. [En ligne ; Page disponible le 25-mai-2013].

[367] B. Maurey. Notes de cours : analyse hilbertienne et de Fourier. 2008-2009. http://www.math.jussieu.fr/~maurey/FourHilb/PolyFH.pdf.

[368] Bachir Bekka. Le théorème d'inversion de Fourier du point de vue des distributions. URL http://agreg-maths.univ-rennes1.fr/documentation/docs/Theoreme-Inversion-Fourier.pdf.

[369] Erdős László. Fourier transform. 2008. URL https://www.mathematik.uni-muenchen.de/~lerdos/WS08/QM/four08.pdf.

[370] Richard Zekri. Analyse de fourier. 15 septembre 2009. URL http://lumimath.univ-mrs.fr/infoetudiant/ANALYSEDEFOURIERMASTER2009.pdf.

[371] Michel Crouzeix. Compléments d'analyse. URL https://perso.univ-rennes1.fr/michel.crouzeix/publis/anep.pdf.

[372] Daniel Li. Notions sur les distributions. . URL http://labomathlens.free.fr/Liens/AF/distrib.pdf.

[373] Claude Zuily. *Éléments de distributions et d'équations aux dérivées partielles*. Dunod, 2002.

[374] Sabin Lessard. Cours de processus stochastiques. 2013. URL http://www.dms.umontreal.ca/~lessards/ProcessusStochastiquesLessard2014.pdf.

[375] Peter Young. Singular Fourier transforms and the integral representation of the Dirac delta function. November, 2009.

[376] Daniel Choï. Méthode des éléments-finis par l'exemple. Avril 2010. URL http://meca.unicaen.fr/~choi/pdf/cours-mef.pdf.

[377] Michael Reiter and Arthur Schuste. Fourier transform and Sobolev spaces. 2008. URL http://www.mat.univie.ac.at/~stein/lehre/SoSem08/sobolev_fourier.pdf.

[378] Norbert Heuer. Ltcc module computation methods. URL http://www.ltcc.ac.uk/courses/AppliedComputationalMethods/ACM2007Chap1.pdf\etautresfichiersdelamêmesérie.

[379] Arkhnor. Densité des fonctions Cinfty à support compact. 2014. URL http://www.ilemaths.net/sujet-densite-des-fonctions-c-infty-a-support-compact-587937.html.

[380] Gilles Leborgne. Complément : espaces de Sobolev fractionnaires. 2007. URL http://ws.isima.fr/~leborgne/IsimathDistributions/sobolev_frac.pdf.

[381] Richard B. Melrose. Lecture notes for 18.155, fall 2002. 2002. URL https://ocw.mit.edu/courses/mathematics/18-155-differential-analysis-fall-2004/lecture-notes/section10.pdf.

[382] Julien Vovelle. Équations différentielles, cours no 2. URL http://math.univ-lyon1.fr/~vovelle/2Cours.pdf.

[383] A. Munnier. Théorie des équations différentielles ordinaires. 2006-2007. URL http://www.iecn.u-nancy.fr/~munnier/files/cours_edo.pdf.

[384] Olivier Debarre. Équations différentielles. URL http://www.math.ens.fr/~debarre/EquaDiff.pdf.

[385] Sylvie Benzoni-Gavage. Propagation d'ondes. URL http://math.univ-lyon1.fr/~benzoni/Ondes.pdf.

[386] Sonia Fliss. L'équation de transport à coefficients variables. URL http://perso.ensta-paristech.fr/~fliss/Sonia_Fliss_web_page/Enseignement_files/Amphi3_2014.pdf.

[387] Sylvie Benzoni-Gavage. Équations différentielles ordinaires. 11 mai 2007. URL http://math.univ-lyon1.fr/~benzoni/EDO-M1.pdf.

[388] Wikipédia. Flot (mathématiques) — wikipédia, l'encyclopédie libre, 2017. URL http://fr.wikipedia.org/w/index.php?title=Flot_(math%C3%A9matiques)&oldid=134130419. [En ligne ; Page disponible le 1-février-2017].

[389] Alexander Grigor'yan. Ordinary differential equation. 2009. URL https://www.math.uni-bielefeld.de/~grigor/odelec2009.pdf.

[390] Christopher P. Grant. Theory of ordinary differential equation. 2007. URL http://www.math.pitt.edu/~bard/bardware/classes/2920/Grant_4july2007.pdf.

[391] John B. Picard-lindelöf theorem with parameter : what regularity ? 2017. URL http://math.stackexchange.com/questions/2126539/picard-lindelöf-theorem-with-parameter-what-regularity.

[392] Kevin Quirin. Développements. 2011-2012. URL http://kevin.quirin.free.fr/Trucs/agreg/developpements.pdf.

[393] Wikipédia. Équations de lotka-volterra — wikipédia, l'encyclopédie libre, 2015. URL http://fr.wikipedia.org/w/index.php?title=%C3%89quations_de_Lotka-Volterra&oldid=113756354. [En ligne ; Page disponible le 5-juillet-2015].

[394] Predrag Cvitanović. Partial differential equations. 2012. URL http://www.cns.gatech.edu/~predrag/courses/PHYS-6124-12/StGoChap6.pdf.

[395] Ralph Chill. Introduction aux équations aux dérivées partielles. 2010-2011. URL http://www.math.univ-metz.fr/~chill/edp.pdf.

[396] Gianluca Bontempi. Un cours d'analyse numérique et de matlab. 2009. http://www.ulb.ac.be/di/map/gbonte/calcul/.

[397] Wikipedia. Two's complement — wikipedia, the free encyclopedia, 2016. URL https://en.wikipedia.org/w/index.php?title=Two%27s_complement&oldid=706335033. [Online ; accessed 6-April-2016].

[398] Wikipédia. Ieee 754 — wikipédia, l'encyclopédie libre, 2015. URL http://fr.wikipedia.org/w/index.php?title=IEEE_754&oldid=120234746. [En ligne ; Page disponible le 12-avril-2016].

[399] Comprendre les nombres à virgules flottantes. URL http://blog.netinfluence.ch/2009/09/24/comprendre-les-nombres-a-virgule-flottante/.

[400] Harald Schmidt. IEEE754 converter. URL http://www.h-schmidt.net/FloatConverter/IEEE754.html.

[401] David Goldberg. What every computer scientist should know about floating-point arithmetic. 1991. URL http://docs.oracle.com/cd/E19957-01/806-3568/ncg_goldberg.html#11655.

[402] Emmanuel Frénod. Calcul scientifique. 2011-2012. URL http://web.univ-ubs.fr/lmam/frenod/IMG/DocEtudiant/MTH1504/calculscientifique.pdf.

[403] Franck Boyer. Agrégation externe de mathématiques. Analyse numérique. 14 octobre 2014. URL http://www.math.univ-toulouse.fr/~fboyer/_media/enseignements/agreg/cours_an_agreg_fboyer_2014.pdf.

[404] Antoine Chambert-Loir. Autour de la méthode de newton. http://perso.univ-rennes1.fr/antoine.chambert-loir/2005-06/agreg/newton.pdf.

[405] Wikipédia. Vitesse de convergence — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=Vitesse_de_convergence&oldid=110252913. [En ligne ; Page disponible le 22-mai-2016].

[406] Martin Pinto Campos. Approximations polynomiales de fonctions et et des intégrales. 21 décembre 2006. D'après le cours de Laurence Halpern http://documents.lamacs.fr/cours/macs1/cours-macs1-pinto1.pdf.

[407] Ming Yang. Matrix decomposition. URL http://users.eecs.northwestern.edu/~mya671/files/Matrix_YM_.pdf.

[408] Benjamin Ambrosio. Chapitre 3 : Méthodes directes de résolution du système $Ax = b$. URL http://lmah.univ-lehavre.fr/~ambrosio/CoursMN/Chapitre3.pdf.

[409] Emmanuel Frénod. Méthodes directes de résolution des systèmes linéaires. URL http://web.univ-ubs.fr/lmam/frenod/IMG/DocEtudiant/MTH1504/methodes-directes.pdf.

[410] Christian Lécot. Analyse numérique matricielle. 2009-2010. URL https://www.lama.univ-savoie.fr/~lecot/data/M1_ANM.pdf.

[411] Wikipédia. Matrice à diagonale dominante — wikipédia, l'encyclopédie libre, 2016. URL http://fr.wikipedia.org/w/index.php?title=Matrice_%C3%A0_diagonale_dominante&oldid=132268176. [En ligne ; Page disponible le 4-mai-2017].

[412] Habib Joulak. Méthodes des différences finies en élasticité. 1974. URL https://ori-nuxeo.univ-lille1.fr/nuxeo/site/esupversions/0e05f3d2-fe12-4d67-b74f-f622869539f2.

[413] Philippe Briand. Probabilités de base. 2005-2006. http://www.lama.univ-savoie.fr/~briand/proba/g12_cours.pdf.

[414] Vincent Bansaye. Variables aléatoires, espérance, indépendance. http://www.cmap.polytechnique.fr/~bansaye/CoursTD2.pdf.

[415] Alfio Marazzi. There are 22 chapters in 22 files. http://www.iumsp.ch/Unites/us/Alfio/polybiostat/ch06.pdf.

[416] Cédric Boutillier. 2016. URL https://github.com/LaurentClaessens/mazhe/issues/16#issue-180279120.

[417] Wikipédia. Inégalité de jensen — wikipédia, l'encyclopédie libre, août 2013. URL http://fr.wikipedia.org/w/index.php?title=In%C3%A9galit%C3%A9_de_Jensen&oldid=90235295. [En ligne ; Page disponible le 9-août-2013].

[418] Gérard Letac. Calcul des probabilités, Deug 2ième année. juin 2001. http://les.mathematiques.free.fr/pdf/proba.zip.

[419] Daniel Saada. Note sur le théorème de transfert pour les variables aléatoires réelles. 18 octobre 2012. http://www.les-mathematiques.net/phorum/file.php?12,file=17435,filename=th_du_tranfert.pdf.

[420] Christophe Sabot. Différents types de convergence de suites de v.a. URL http://math.univ-lyon1.fr/~sabot/Convergence.pdf.

[421] Antonin Macé and Thomas Pradeau. Exposé de maîtrise. Estimations de densités de probabilité. 2007. Encadré par Gilles Stoltz. http://www.eleves.ens.fr/home/amace/travaux/exposemaitrise.pdf.

[422] Wikipedia. Continuous mapping theorem — wikipedia, the free encyclopedia, 2013. URL http://en.wikipedia.org/w/index.php?title=Continuous_mapping_theorem&oldid=545523937. [Online ; accessed 15-June-2013].

[423] Wikipédia. Inégalité de markov — wikipédia, l'encyclopédie libre, 2014. URL http://fr.wikipedia.org/w/index.php?title=In%C3%A9galit%C3%A9_de_Markov&oldid=100208179. [En ligne ; Page disponible le 11-février-2014].

[424] Charles Suquet. Simulation. 2005-2006. http://math.univ-lille1.fr/~suquet/ens/Agr/simul06.pdf.

[425] Yoann Gelineau. Vecteurs gaussiens. . http://math.univ-lyon1.fr/~gelineau/files/vecteurs_gaussiens.pdf.

[426] Wikipédia. Nombre normal — wikipédia, l'encyclopédie libre, avril 2013. URL http://fr.wikipedia.org/w/index.php?title=Nombre_normal&oldid=89848922. [En ligne ; Page disponible le 28-avril-2013].

[427] Wikipédia. Théorème de glivenko-cantelli — wikipédia, l'encyclopédie libre, . URL http://fr.wikipedia.org/w/index.php?title=Th%C3%A9or%C3%A8me_de_Glivenko-Cantelli&oldid=90841113. [En ligne ; Page disponible le 15-juin-2013].

[428] Vincent Rivoirard and Gilles Stoltz. Estimation de densité de probabilités. URL http://www.math.ens.fr/statenaction/PDF/Densite-Principal.pdf.

[429] Nils Berglund. Chaînes de Markov. Décembre 2007. URL http://www.univ-orleans.fr/mapmo/membres/berglund/markov.pdf.

[430] Arnaud Guyader. Espérance conditionnelle et chaînes de Markov. 200-2008. http://www.sites.univ-rennes2.fr/laboratoire-statistique/AGUYADER/doc/proba/poly.pdf.

[431] Massimiliano Gubinelli. Comportement asymptotique des martingales. 2011-2012. http://www.ceremade.dauphine.fr/~mgubi/e1112/pd3.pdf.

[432] Jean Lacroix and Pierre Priouret. Probabilités approfondies. 2005-2006. http://www.proba.jussieu.fr/cours/processus.pdf.

[433] Wikipedia. Optional stopping theorem — wikipedia, the free encyclopedia, 2013. URL http://en.wikipedia.org/w/index.php?title=Optional_stopping_theorem&oldid=547547174. [Online ; accessed 14-August-2013].

[434] Nizar Touzi. Martingales en temps discret et chaîne de Markov. Septembre 2009. URL http://www.cmap.polytechnique.fr/~touzi/MAP432-Poly.pdf.

[435] Zach Weinersmith. 2008. URL http://www.smbc-comics.com/comic/2008-02-21.