Phénomènes de grande dimension

Notes de cours, en travaux, datées du 3 avril 2025
Formation 1A(L3) mathématiques/physique

Département de mathématiques et applications (DMA)
École normale supérieure – PSL
45 rue d'Ulm, Paris 5e

2024 – 2025

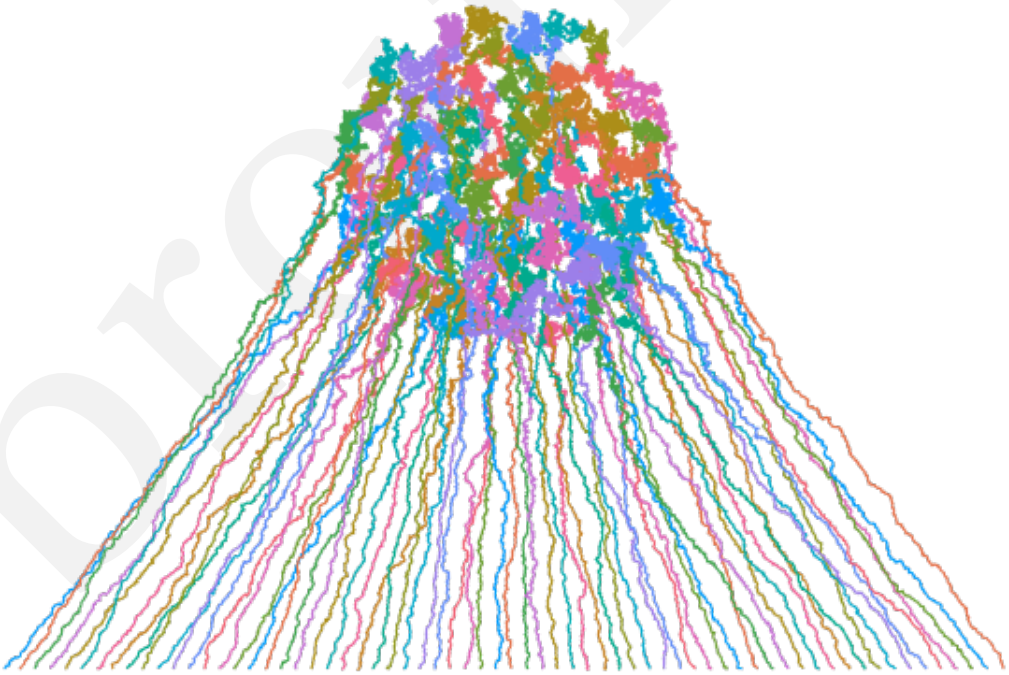


Table des matières

Ce cours est offert par le département de mathématiques et applications de l'École normale supérieure, au second semestre, pour le parcours maths/physique de la première année (1A c'est-à-dire L3).

Ce cours est centré autour de phénomènes de grande dimension de nature probabiliste. Il s'agit au départ du comportement des vecteurs, matrices, et tenseurs aléatoires en grande dimension, à commencer par les théorèmes limites pour les variables indépendantes.

La première partie du cours est assurée par un mathématicien probabiliste (Djalil Chafaï) tandis que la seconde partie du cours est assurée par un physicien théoricien (Giulio Biroli). La dernière séance est assurée par Jean-Philippe Bouchaud, physicien théoricien et membre de l'Académie des sciences.

— **Pré-requis.** Cours de L3 *Intégration et probabilités* et *Physique statistique*[1].
  En particulier, aucune notion de conditionnement ni ne processus stochastique n'est requise.
— **Évaluation.** Sous forme d'oral ou d'écrit en fonction du nombre d'étudiants, pour chaque partie.
— **Emploi du temps.** 2 × 6 semaines avec 2×1,5h de cours + 2h de TD/GT, + 1 séance de cours finale
— **Rétribution.** 2 × 6 ECTS au total.

Ces notes de cours sont plus riches que le cours oral, qui n'en aborde qu'une sélection.

Ces notes ne concernent que la partie mathématique (DC) du cours. Le but de la partie mathématique du cours est de fournir une culture de base autour des phénomènes de grande dimension, entre probabilités, analyse (asymptotique, fonctionnelle, spectrale), et géométrie convexe. Il s'agit d'un premier contact, d'une ouverture vers un vaste champ thématique. Une difficulté vient du fait qu'il s'agit d'un cours de L3 S2 : en principe les étudiants du DMA de L3 S2 ont déjà suivi le cours de L3 S1 *Intégration et probabilités*, mais pas encore le cours de M1 S1 *Processus stochastiques*. Ils connaissent donc la formalisation des probabilités avec la théorie de la mesure et l'intégrale de Lebesgue, mais n'ont pas encore étudié dans ce cadre les notions d'espérance conditionnelle, de martingale, et de chaîne de Markov. Il est tenu compte de tout cela.

Les phénomènes de grande dimension abordés dans ces notes sont relatifs au comportement de quantités d'intérêt qui font intervenir un grand nombre de constituants aléatoires. Voici les thèmes phares :

— Loi des grands nombres, théorème limite central, méthode de Monte–Carlo
— Lois gaussiennes et mesures de Boltzmann–Gibbs
— Entropie de Boltzmann–Shannon, divergence de Kullback–Leibler ou énergie libre ou entropie relative
— Phénomène couche ou couronne mince et théorème limite central pour des corps convexes
— Phénomène de concentration de la mesure et inégalités fonctionnelles de log-Sobolev et de Talagrand
— Principes de grandes déviations de Cramér et Sanov, principes de Laplace–Varadhan et de contraction
— Queues lourdes, invariance d'échelle, lois stables, théorème limite central stable
— Théorème de Wigner pour les matrices aléatoires, et esquisse du théorème de Marchenko–Pastur. Méthode des moments, ensemble gaussien orthogonal (GOE), transformée de Cauchy–Stieltjes, résolvante.

L'entropie et l'énergie sont très présentes, à la fois pour l'analyse quantitative et pour l'analyse asymptotique. La convexité joue également un rôle important, à travers les corps convexes ou la transformée de Legendre.

On l'aura compris, il s'agit avant tout d'un cours sur l'aléatoire, qui met en jeu un paramètre de dimension ou interprété comme tel. L'aléatoire est à la fois au cœur de la théorie des probabilités en mathématiques, mais aussi au cœur de la physique statistique, et de la mécanique statistique, à la confluence des deux.

Du point de vue utilitaire, la grande dimension peut poser problème en faisant exploser le champ des possibles, et on parle alors de fléau ou de malédiction de la dimension. D'un autre côté, la grande dimension peut également simplifier le comportement, et on parle alors parfois de bénédiction de la grande dimension.

Ces thématiques sont liées aux travaux de plusieurs célébrités, dont : Michel Talagrand (concentration et transport, prix Abel 2024), Luis Caffarelli (transport, prix Abel 2023), Srinivasa Varadhan (grandes déviations, prix Abel 2007), Jean Bourgain (corps convexes, médaille Fields 1994), Terence Tao (matrices aléatoires, médaille Fields 2006), Cédric Villani (transport et log-Sobolev, médaille Fields 2010), Alessio Figalli (transport, médaille Fields 2018), Leonid Kantorovich (transport, prix Nobel d'Économie 1975), Eugene Wigner (matrices aléatoires, prix Nobel de Physique 1963), Giorgio Parisi (physique statistique, prix Nobel de Physique 2021).

Un conseil, au-delà de ce cours : n'hésitez pas à suivre la fée Curiosité, même si elle ressemble à sa sœur jumelle Dispersion. Il n'y a pas de frontière nette entre les mathématiques, l'informatique, et la physique, et ce grand triangle se navigue bien pour ceux qui ont le pied marin et le goût du voyage, avec ou sans port d'attache!

Auteur de ces notes de cours :

— 2022 – 2023, 2023 – 2024, 2024 – 2025 : Djalil Chafaï

1. Pour la partie mathématique, avoir suivi le cours de L3 *Physique statistique* est recommandé mais pas indispensable.

Organisation de l'année 2024 – 2025

Créneaux horaires

— 2×1,5 = 3h de CM et 2h de TD par semaine
— Les CMs ont lieu le mercredi 15h15 – 16h45 (Cartan) et le jeudi 13h30 – 15h (Noether)
— Le TD a lieu le jeudi de 15h15 à 17h15 (Noether).

Ressources pédagogiques

Les notes de cours, feuilles de TD, DM, se trouvent sur `https://moodle.psl.eu/`

Calendrier prévisionnel des cours et TD

DC = Djalil Chafaï (TD assuré par Lucas REY), GB = Giulio Biroli, JPB = Jean-Philippe Bouchaud.

— **Semaine 1**
`Me 04/02` : DC1 Vecteurs gaussiens, LGN, TLC, Monte–Carlo, caractérisations de Maxwell et de Boltzmann
`Je 06/02` : DC2 Phénomène couche mince, géométrie de la sphère, principe d'Archimède et TLC pour les corps convexes
— **Semaine 2**
`Me 12/02` : DC3 Phénomène de concentration de la mesure, inégalité de log-Sobolev
`Je 13/02` : DC4 Phénomène de concentration de la mesure, inégalité de Talagrand
— **Semaine 3**
`Me 19/02` : DC5 Principe de grandes déviations de Cramér, transformée de Legendre
`Je 20/02` : DC6 Principe de grandes déviations de Cramér, lemme de Laplace–Varadhan
— **Vacances (semaine du 22 février au 28 février)**
— **Semaine 4**
`Je 05/03` : DC7 Principe de grandes déviations de Sanov, mesure empirique et entropie
`Me 06/03` : DC8 Queues lourdes, lois de puissance, invariance d'échelle
— **Semaine 5**
`Me 12/03` : DC9 Queues lourdes, lois stables, universalité
`Je 13/03` : DC10 Théorème de Wigner et méthode des moments
— **Semaine 6 (pendant la semaine des partiels)**
`Me 19/03` : DC11 Théorème de Wigner et méthode de Pastur
`Je 20/03` : Examen oral DC salle R3
— **Semaine 7**
`Me 26/03` : Pause
`Je 27/03` : GB1
— **Semaine 8**
`Me 02/04` : GB2
`Je 03/04` : GB3
— **Semaine 9**
`Me 09/04` : GB4
`Je 10/04` : GB5
— **Semaine 10**
`Me 16/04` : GB6
`Je 17/04` : GB7
— **Vacances (semaine du 21 au 25 avril)**
— **Semaine 11**
`Me 30/04` : GB8
`Je 01/05` : Férié (fête du travail)
— **Semaine 12**
`Me 07/05` : GB9
`Je 08/05` : Férié (commémoration de la victoire de la second guerre mondiale)
— **Semaine 13**
`Me 14/05` : GB10
`Je 15/05` : GB11
— **Semaine 14**
`Me 21/05` : JPB1
`Je 22/05` : JPB2
— **Semaine de révision (26 au 30 mai) mais examen GB en fin de semaine**
`Ve 30/05` : Examen GB

Modalité d'évaluation 2024 – 2025

L'examen pour la partie mathématique est un oral individuel.

Il est organisé autant que possible pendant la semaine des partiels.

Il consiste en un exposé de 15 minutes au tableau, suivi de 10 minutes de questions, qui pourront porter sur toutes les parties du cours. L'exposé concerne tout ou partie d'un article à lire, à choisir librement parmi un ensemble d'articles mis à disposition sur Moodle, avec un niveau de difficulté indicatif. L'exposé doit mettre en avant autant que possible les idées et les articulations, plutôt que le contenu in extenso, et faire le lien avec le cours et les travaux dirigés. Il est recommandé de se munir d'une montre et de penser à la gestion du tableau.

Liste des articles proposés pour 2024 – 2025 et disponibles sur Moodle :

— **Piste noire**
Gérard Ben Arous et Alice Guionnet
Large deviations for Wigner's law and Voiculescu's non-commutative entropy
Probability Theory and Related Fields 108 517–542 (1997)

— **Piste rouge**
Alice Guionnet et Ofer Zeitouni
Concentration of the spectral measure for large matrices
Electronic Communications in Probability 5 119–136 (2000)

— **Piste rouge bis**
Bernard Maurey
Some deviation inequalities
Geometric and Functional Analysis I(2) 188–197 (1991)

— **Piste bleue**
Sergey Bobkov
An isoperimetric inequality on the discrete cube and an elementary proof of the isoperimetric inequality in Gauss space
The Annals of Probability 25(1) 206–214 (1997)

— **Piste verte**
Sacha Friedli et Yvan Velenik
Statistical Mechanics of Lattice Systems : A Concrete Mathematical Introduction
Chapitre 2 : *The Curie–Weiss Model*
Cambridge University Press (2017)

Chapitre 1

Théorèmes limites standards, méthode de Monte–Carlo, corps convexes

Me 05/02

Abordé en cours :
— Vecteurs gaussiens (Cochran rapidement)
— Maxwell et Boltzmann, énergie libre et mesures de Boltzmann–Gibbs
— TLC et LGN sans preuve avec formulation vecteur aléatoire de TLC
— Esquisse de Monte–Carlo (rapidement).

But : comportement en grande dimension n d'un vecteur aléatoire $(X_1,\dots,X_n)$ de $\mathbb{R}^n$.
Phénomènes : loi des grands nombres et théorème limite central.
Algorithme : méthode de Monte–Carlo.

Pour $x\in\mathbb{R}^n$ on note $|x|_p := (|x_1|^p+\cdots+|x_n|^p)^{1/p}$ si $1\le p<\infty$, $|x|_\infty := \max_{1\le i\le n}|x_i|$.
Boules et sphères : on note $\mathbb{B}_p^n(r) := \{x\in\mathbb{R}^n : |x|_p\le r\}$ et $\mathbb{S}_p^{n-1}(r) := \{x\in\mathbb{R}^n : |x|_p = r\}$.
On omet souvent l'indice p quand $p=2$, et « (r) » quand $r=1$, en particulier $\mathbb{B}^n=\mathbb{B}_2(1)$ et $\mathbb{S}^{n-1}=\mathbb{S}_2^{n-1}(1)$.
$\mathbb{B}_1^n$, $\mathbb{B}_2^n$, $\mathbb{B}_\infty^n$, sont le « diamant », la boule euclidienne, et le cube $[-1,1]^n$ de rayon unité.
On note $L^p(E)$ l'espace de Lebesgue des v.a. $\Omega\to E$ où $(\Omega,\mathscr{A},\mathbb{P})$ est un espace de probabilité implicite.
On omet E lorsque typiquement $E=\mathbb{R}$ ou $E=\mathbb{R}^n$, ou en fonction du contexte.

1.1 Lemme de Borel–Cantelli et convergence complète

Compter des événements revient à sommer leurs indicateurs. Pour des événements (A_n) dans une tribu $\mathscr{A}$,

$$\limsup A_n := \bigcap_m \bigcup_{n\ge m} A_n = \Big\{\sum_n \mathbb{1}_{A_n}=\infty\Big\} = \{\text{appartenir à } A_n \text{ pour une infinité de valeurs de } n\}.$$

Lemme 1.1.1. Borel–Cantelli.

Soient (A_n) des événements dans $(\Omega,\mathscr{A},\mathbb{P})$.
— Si $\sum_n \mathbb{P}(A_n)<\infty$ alors $\mathbb{P}(\limsup A_n)=0$.
— Si $\sum_n \mathbb{P}(A_n)=\infty$ et si les A_n sont indépendants alors $\mathbb{P}(\limsup A_n)=1$.
En particulier, si les A_n sont indépendants, alors $\mathbb{P}(\limsup A_n)\in\{0,1\}$ (loi du zéro-un de Borel).

Il en découle que la convergence en probabilité implique la convergence p.s. d'une sous-suite.

Démonstration. Pour la partie Borel, si X est à valeur dans $[0,\infty]$ et $\mathbb{E}(X)<\infty$ alors $\mathbb{P}(X<\infty)=1$. Or avec $X:=\sum_n \mathbb{1}_{A_n}$, cela donne $\mathbb{E}(X)<\infty\Rightarrow\mathbb{P}(\limsup A_n)=0$, tandis que par Fubini–Tonelli ou convergence monotone,

$$\mathbb{E}(X)=\mathbb{E}\sum_n \mathbb{1}_{A_n}=\sum_n \mathbb{E}\mathbb{1}_{A_n}=\sum_n \mathbb{P}(A_n).$$

Pour la réciproque, par continuité supérieure ou convergence monotone, et indépendance des A_n^c

$$1-\mathbb{P}(\limsup A_n)=\mathbb{P}(\liminf A_n^c)=\mathbb{P}(\cup_m\cap_{n\ge m}A_n^c)=\lim_{m\to\infty}\mathbb{P}(\cap_{n\ge m}A_n^c)=\lim_{m\to\infty}\prod_{n\ge m}(1-\mathbb{P}(A_n)),$$

et il ne reste plus qu'à utiliser le critère de convergence de produit infini en passant au logarithme. □

Théorème 1.1.2. Convergence complète.

Soit (X_n) une suite de variables aléatoires à valeurs dans un espace métrique (E,d), et soit $c \in E$.
Les propriétés suivantes sont équivalentes :

i) $\sum_n \mathbb{P}(d(X_n,c) \geq \varepsilon) < \infty$ pour tout $\varepsilon > 0$ (convergence en probabilité « sommable », vers c).

ii) $Y_n \xrightarrow[n\to\infty]{\text{p.s.}} c$ pour toute suite de variables aléatoires (Y_n) t.q. $X_n \overset{\text{d}}{=} Y_n$ pour tout n.

iii) $Y_n \xrightarrow[n\to\infty]{\text{p.s.}} c$ pour une suite de variables aléatoires (Y_n) indépendantes t.q. $X_n \overset{\text{d}}{=} Y_n$ pour tout n.

Ce mode de convergence ne dépend que des lois marginales. Il s'agit d'une propriété de la suite (μ_n) où μ_n est la loi de X_n. La première condition s'écrit $\sum_n \mu_n(B(c,\varepsilon)^c) < \infty$. On parle de convergence complète de μ_n vers c. Comme nous allons le voir, la convergence complète est naturellement présente dans les problèmes de probabilités en grande dimension, en liaison avec le phénomène de concentration de la mesure. Elle affirme une universalité : convergence presque sûre vers une constante quelque soit le couplage des lois marginales.

Démonstration. (i)⇒(ii). De (i) on tire $Y_n \xrightarrow{\mathbb{P}} c$ mais cela est trop faible. De (i) on a, par Borel–Cantelli (première partie) : pour tout $\varepsilon > 0$, il existe A_ε tel que $\mathbb{P}(A_\varepsilon) = 1$, et sur A_ε, il existe $N = N_\varepsilon$ tel que $d(Y_n,c) < \varepsilon$ pour tout $n \geq N$. En prenant $A := \cap_{r=1}^\infty A_{1/r}$, il vient que $\mathbb{P}(A) = 1$ tandis que sur A, $d(Y_n,c) \to 0$, c'est-à-dire (ii).
(ii)⇒(iii) Immédiat. (iii)⇒(i). Soit $\varepsilon > 0$ et $A_n := \{d(Y_n,c) \geq \varepsilon\}$. De (iii) on tire $\mathbb{P}(\limsup A_n) = 0$, puis de l'indépendance des A_n et de Borel–Cantelli (seconde partie !), on tire $\sum_n \mathbb{P}(A_n) < \infty$. □

1.2 Vecteurs gaussiens et caractérisations de Maxwell et Boltzmann

Les vecteurs gaussiens, ou plutôt leurs lois appelées lois gaussiennes, apparaissent naturellement par le phénomène TLC, comme limite de sommes normalisées de variables aléatoires indépendates et identiquement distribuées de carré intégrable. Du point de vue analytique, cela correspond à un point fixe de convolution renormalisée. Le mouvement brownien est un vecteur gaussien de dimension infinie, limite d'échelle universelle, par le phénomène TLC, des marches aléatoires à incréments de carré intégrable.

Si $X = (X_1,\dots,X_n)^\top$ est un vecteur (colonne) aléatoire de $\mathbb{R}^n$ (à composantes) de carré intégrable, alors son vecteur moyenne et sa matrice de covariance sont donnés par (la matrice K est symétrique et à spectre ≥ 0)

$$\begin{aligned} m &= \mathbb{E}(X) = (\mathbb{E}(X_1),\dots,\mathbb{E}(X_n))^\top \\ K &= \big(\mathbb{E}((X_j - m_j)(X_k - m_k))\big)_{1\leq j,k\leq n} = \mathbb{E}((X-m)(X-m)^\top) = \mathbb{E}(XX^\top) - mm^\top \\ &= \mathbb{E}((X-m)\otimes(X-m)) = \mathbb{E}(X\otimes X) - m\otimes m. \end{aligned}$$

— La covariance K est une moyenne de matrices symétriques de rang 1 semi-définies positives.
— On dit que X est isotrope si $K = \sigma^2 I_n$ pour $\sigma > 0$.
— Si les composantes $X_1,\dots,X_n$ sont indépendantes alors K est diagonale.
— Si $a \in \mathbb{R}^d$ et $A \in \mathscr{M}_{d,n}(\mathbb{R})$ alors $a + AX$ a pour moyenne $a + Am$ et covariance $AKA^\top$.
En effet, si X est centré alors $\mathbb{E}(AX\otimes AX) = \mathbb{E}(AX(AX)^\top) = A\mathbb{E}(XX^\top)A^\top = AKA^\top$.
En particulier, en prenant $d = 1$, si $x \in \mathbb{R}^n$ alors $\langle X, x\rangle$ a pour moyenne $\langle m, x\rangle$ et covariance $\langle Kx, x\rangle$.
— En diagonalisant ou par Choleski, toute matrice symétrique à spectre ≥ 0 est une matrice de covariance.
— $X \in L^2$ ssi la fonction caractéristique $\varphi_X(\cdot) := \mathbb{E}(e^{i\langle X,\cdot\rangle})$ est dérivable deux fois en 0, et dans ce cas

$$m = \mathbb{E}(X) = -i\nabla\varphi_X(0) \quad \text{et} \quad K + mm^\top = \mathbb{E}(XX^\top) = -\text{Hess}\varphi_X(0).$$

Théorème-définition 1.2.1. Vecteurs gaussiens et lois gaussiennes.

Soit $X = (X_1,\dots,X_n)^\top$ un vecteur (colonne) aléatoire de $\mathbb{R}^n$. Les propriétés suivantes sont équivalentes :

(i) toute combinaison linéaire des composantes est gaussienne :

$$\langle t, X\rangle = t_1X_1 + \cdots + t_nX_n \text{ est une variable gaussienne réelle, pour tout } t \in \mathbb{R}^n.$$

(ii) les cumulants d'ordre > 2 sont tous nuls :

$$\varphi_X(t) := \mathbb{E}(e^{i\langle t,X\rangle}) = e^{i\langle t,m\rangle - \frac{1}{2}\langle Kt,t\rangle}, \text{ pour tout } t \in \mathbb{R}^n.$$

(iii) $X \overset{\text{loi}}{=} m + AZ$, où $m \in \mathbb{R}^n$, $A \in \mathscr{M}_n(\mathbb{R})$, $Z = (Z_1,\dots,Z_n)$, $Z_1,\dots,Z_n$ i.i.d. $\mathscr{N}(0,1)$.

On dit alors que X est un vecteur gaussien. Sa loi ne dépend que de m et K : loi gaussienne $\mathcal{N}(m,K)$.

— Nommage : loi gaussienne mais aussi loi normale, d'où la notation $\mathcal{N}$.
— Notation : $\gamma_\sigma^n := \mathcal{N}(0,\sigma^2 I_n)$ (gaussiennes isotropes).
— Cas extrêmes : $\gamma_0^n = \delta_0$ et $\gamma_\infty^n \approx$ Lebesgue.
— Brownien/EDP : le noyau de la chaleur $\gamma_{1/(2t)}^n$ interpole entre δ_0 en $(t=0)$ et Lebesgue $(t=\infty)$.
— Notation : $\gamma^n := \gamma_1^n = \mathcal{N}(0,I_n) = \mathcal{N}(0,1)^{\otimes n} = (\gamma_1)^{\otimes n}$ est la loi gaussienne standard.
— Le (iii) dit que les lois gaussiennes sont les déformations affines de la loi gaussienne standard $\mathcal{N}(0,I_n)$.
— Toute matrice symétrique à spectre positif est la matrice de covariance d'un vecteur aléatoire gaussien.
— Application : Simulation de $\mathcal{N}(m,K)$ par réduction à $\mathcal{N}(0,1)$ via $K = AA^\top$ (diagonalisation ou Choleski).

Démonstration. Rappelons que $\varphi_{\mathcal{N}(a,b^2)}(s) = \mathrm{e}^{\mathrm{i}as - \frac{b^2}{2}s^2}$ (mnémotechnique : vecteur propre de Fourier).
(i)⇒(ii). Les $X_1,\dots,X_n$ sont gaussiennes (prendre $t=e_i$) donc de carré intégrable, donc X possède un vecteur moyenne m et une matrice de covariance K. Si $t\in\mathbb{R}^n$ alors $\langle t,X\rangle = t_1X_1+\cdots+t_nX_n \sim \mathcal{N}(\langle t,m\rangle,\langle Kt,t\rangle)$ donc $\varphi_X(t) = \varphi_{\langle t,X\rangle}(1) = \exp(\mathrm{i}\langle t,m\rangle - \frac{1}{2}\langle Kt,t\rangle)$. Donc les cumulants de $\langle t,X\rangle$ d'ordre >2 sont tous nuls.
(ii)⇒(i). Pour tout $t\in\mathbb{R}^n$, comme $\varphi_X(t) = \varphi_{\langle t,X\rangle}(1)$, on en déduit que $\langle t,X\rangle \sim \mathcal{N}(\langle t,m\rangle,\langle Kt,t\rangle)$.
(iii)⇒(i) Les composantes de X sont des combinaisons linéaires de v.a.r. gaussiennes indépendantes.
(ii)⇒(iii) Soit A telle que $AA^\top = K$ (factorisation de Choleski ou diagonalisation). Alors pour tout $t\in\mathbb{R}^n$,

$$\varphi_{m+AZ}(t) = \mathrm{e}^{\mathrm{i}\langle m,t\rangle}\mathbb{E}(\mathrm{e}^{\mathrm{i}\langle A^\top t,Z\rangle}) = \mathrm{e}^{\mathrm{i}\langle m,t\rangle}\varphi_Z(A^\top t) = \mathrm{e}^{\mathrm{i}\langle m,t\rangle}\mathrm{e}^{-\frac{1}{2}\|A^\top t\|^2} = \mathrm{e}^{\mathrm{i}\langle m,t\rangle}\mathrm{e}^{-\frac{1}{2}\langle AA^\top t,t\rangle} = \varphi_X(t).$$

□

Théorème 1.2.2. Stabilité par transformation affine et indépendance des composantes.

Soit $X = (X_1,\dots,X_n)^\top \sim \mathcal{N}(m,K)$ un vecteur gaussien de $\mathbb{R}^n$.

(i) Pour tout $m\in\mathbb{R}^d$ et tout $A\in\mathcal{M}_{d,n}(\mathbb{R})$, on a $m+AX \sim \mathcal{N}(m,AKA^\top)$.

(ii) Les composantes $X_1,\dots,X_n$ sont indépendantes ssi K est diagonale.

— Contre-exemple : si $Z\sim\mathcal{N}(0,1)$ et $\varepsilon \sim \frac{1}{2}(\delta_{-1}+\delta_1)$ alors $X = (Z,\varepsilon Z)$ n'est pas gaussien car $X_1+X_2 = Z+\varepsilon Z = (1+\varepsilon)Z$ vaut 0 avec probabilité 1/2. La matrice de covariance de X est I_2, diagonale, car $\mathbb{E}(X_1X_2) = \mathbb{E}(\varepsilon)\mathbb{E}(Z^2) = 0$ et $X_1^2 = X_2^2 = Z^2$, mais X_1 et X_2 ne sont pas indépendantes : $\mathrm{Loi}(X_1 \mid X_2 = x) = \pm x$.

Démonstration.

(i) Pour tout $t\in\mathbb{R}^n$, $\langle t,m+AX\rangle = \langle t,m\rangle + \langle A^\top t,X\rangle$ qui est gaussienne car X est gaussien.

(ii) Si $K = \mathrm{diag}(\sigma_1^2,\dots,\sigma_n^2)$ alors pour tout $t\in\mathbb{R}^n$, $\varphi_X(t) = \prod_{i=1}^n \exp(\mathrm{i}t_im_i - \frac{1}{2}\sigma_i^2t_i^2) = \prod_{i=1}^n \varphi_{X_i}(t_i)$ donc $X_1,\dots,X_n$ sont indépendantes. La réciproque n'a rien de gaussien (covariance nulle si indépendance).

□

Théorème 1.2.3. Densité gaussienne.

La loi gaussienne $\mathcal{N}(m,K)$ sur $\mathbb{R}^n$ a une densité ssi K est inversible, donnée par

$$x\in\mathbb{R}^n \mapsto \frac{\exp\Big(-\frac{1}{2}\langle K^{-1}(x-m),x-m\rangle\Big)}{\sqrt{(2\pi)^n\det(K)}},$$

et dans le cas contraire $\mathcal{N}(m,K)$ est portée par le sous-espace affine strict $m+\mathrm{Im}(K)$ de $\mathbb{R}^n$.

— La loi gaussienne standard $\mathcal{N}(0,I_n) = \mathcal{N}(0,1)^{\otimes n}$ a pour densité $x\in\mathbb{R}^n \mapsto (2\pi)^{-\frac{n}{2}}\mathrm{e}^{-\frac{|x|^2}{2}}$.
— La loi gaussienne isotrope $\gamma_\sigma^n := \mathcal{N}(0,\sigma^2I_n) = \mathcal{N}(0,\sigma^2)^{\otimes n}$ a pour densité $x\in\mathbb{R}^n \mapsto (2\pi\sigma^2)^{-\frac{n}{2}}\mathrm{e}^{-\frac{|x|^2}{2\sigma^2}}$.
— Mnémotechnique sur K : la densité fait apparaître K^{-1}, ce qui est normal pour la dispersion, tandis que la fonction caractéristique fait apparaître K. C'est une instance du principe d'incertitude de l'analyse harmonique : dispersion simultanée impossible du signal et de sa transformée de Fourier !

Démonstration. Soit $A \in \mathcal{M}_{n,n}(\mathbb{R})$ telle que $K = AA^\top$, de sorte que $X \overset{\mathrm{d}}{=} m + AZ$, où $Z \sim \mathcal{N}(0, I_n) = \mathcal{N}(0,1)^{\otimes n}$. Si K est inversible, alors A est inversible, et comme Z a pour densité $\prod_{i=1}^n (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x_i^2) = (2\pi)^{-\frac{n}{2}} \exp(-\frac{1}{2}|x|^2)$, il vient par changement de variable affine que $m + AZ$ a pour densité la formule de l'énoncé. Si K n'est par inversible, alors A n'est pas inversible et $m + AZ$ prend ses valeurs dans $m + A\mathbb{R}^n$. Or $\mathrm{Im}(A) = \mathrm{Im}(K)$. □

Théorème 1.2.4. Théorème de Cochran sur les gaussiennes isotropes [a].

a. Se prononce kok-rane.

Si $Z = (Z_1, \dots, Z_n) \sim \gamma_\sigma^n$, et si $\mathbb{R}^n = E_1 \oplus \cdots \oplus E_r$ avec $E_j \perp E_k$ si $j \neq k$, alors les vecteurs aléatoires $\mathrm{proj}_{E_k}(Z) = P_k Z$, $1 \leq k \leq r$, sont gaussiens indépendant avec $P_k Z \sim \mathcal{N}(0, \sigma^2 P_k)$ dans $\mathbb{R}^n$.

Démonstration. Comme les E_i sont $\perp$ et comme la loi de Z est invariante par rotation, on peut se ramener au cas où $E_k = \mathrm{vect}(e_i : i \in I_k)$, $1 \leq k \leq r$, pour une partition $I_1 \cup \cdots \cup I_r = \{1, \dots, n\}$. Le résultat découle alors du théorème 1.2.2 : la matrice P_k de projection orthogonale sur E_k est diagonale et $(P_k)_{i,i} = \mathbb{1}_{i \in I_k}$ pour tout i. □

Corollaire 1.2.5. Studentisation.

Si les $X_1, \dots, X_n$ sont i.i.d. $\mathcal{N}(m, \sigma^2)$ alors la moyenne empirique et la variance empirique

$$m_n := \frac{X_1 + \cdots + X_n}{n} \quad \text{et} \quad \sigma_n^2 := \frac{(X_1 - m_n)^2 + \cdots + (X_n - m_n)^2}{n-1}$$

sont indépendantes, $m_n \sim \mathcal{N}(m, \frac{\sigma^2}{n})$, $\frac{n-1}{\sigma^2}\sigma_n^2 \sim \chi^2(n-1)$, d'où $\sqrt{n}\frac{m_n - m}{\sigma_n} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} =: \mathrm{Student}(n-1)$.

— On qualifie d'empirique ce qui est issu de l'expérience, ici l'échantillonnage d'un phénomène aléatoire.

Démonstration. On pose $\mathbb{R}^n = E_1 \oplus E_2$, $E_1 := \mathbb{R}(1, \dots, 1)$, $E_2 = E_1^\perp$, d'où $\mathrm{proj}_{E_1}(X) = m_n$ et $\mathrm{proj}_{E_2}(X) = X - m_n$. □

Remarque 1.2.6. Caractérisations de la gaussienne.

Voici trois propriétés qui caractérisent la gaussienne sous certaines conditions :

— Moments : si $Z \sim \mathcal{N}(0,1)$ alors

$$\mathbb{E}(Z^{2n-1}) = 0 \quad \text{et} \quad \mathbb{E}(Z^{2n}) = \prod_{k=0}^{n-1}(2k+1) = \frac{(2n)!}{2^n n!} \quad \text{pour tout } n \geq 1.$$

— Intégration par parties gaussienne ou équation de Stein : si $Z \sim \mathcal{N}(0,1)$ alors

$$\mathbb{E}(Zf(Z)) = \mathbb{E}(f'(Z)) \quad \text{pour tout } f \in \mathscr{C}_c^1.$$

— Formule de Wick : si $X \sim \mathcal{N}(0, K)$ est un vecteur gaussien centré de $\mathbb{R}^d$ et si $f_1, \dots, f_{2n}$ est un nombre pair de formes linéaires $\mathbb{R}^d \to \mathbb{R}$ alors

$$\mathbb{E}(f_1(X) \cdots f_{2n}(X)) = \sum_{\mathscr{A}_n} \prod_{r=1}^n \mathbb{E}(f_{i_r} f_{j_r})$$

où $\mathscr{A}_n$ désigne l'ensemble des appariements : suites non-ordonnées de paires non-ordonnées $\{\{i_1, j_1\}, \dots, \{i_n, j_n\}\}$ où chacun des entiers $1, \dots, 2n$ apparaît exactement une fois. Formule d'Isserlis : $\mathrm{Card}(\mathscr{A}_n) = \frac{(2n)!}{2^n n!}$. Chaque appariement peut être numéroté de manière unique de sorte que $i_1 < \cdots < i_n$ et $i_r < j_r$ pour tout $1 \leq r \leq n$. Les éléments de $\mathscr{A}_2$ par exemple sont $\{\{1,2\},\{3,4\}\}$, $\{\{1,3\},\{2,4\}\}$, $\{\{1,4\},\{2,3\}\}$. Si par exemple X est gaussien centré de $\mathbb{R}^2$ avec $\mathbb{E}(X_1^2) = 2$, $\mathbb{E}(X_2^2) = 3$, et $\mathbb{E}(X_1 X_2) = 1$, alors la formule de Wick avec $n = 2$, $f_1(x) = f_2(x) = x_1$ et $f_3(x) = f_4(x) = x_2$, donne

$$\mathbb{E}(X_1^2 X_2^2) = \mathbb{E}(X_1 X_1 X_2 X_2) = 2 \times 3 + 1 \times 1 + 1 \times 1 = 8.$$

Deux caractérisations des gaussiennes issues de la physique statistique dues à Maxwell [1] et Boltzmann [2].

1. James Clerk Maxwell (1831 – 1879).
2. Ludwig Eduard Boltzmann (1844 – 1906).

Théorème 1.2.7. Caractérisation géométrique de Maxwell des gaussiennes isotropes.

Une loi de probabilité μ sur $\mathbb{R}^n$, $n \geq 2$, est produit et invariante par rotation ssi $\mu = \gamma_\sigma^n$ pour un $\sigma \geq 0$.

— Quand $\sigma = 0$, on dispose de la convention très naturelle $\gamma_0^n = \delta_0 = \delta_0^{\otimes n}$.
— Quand $n = 1$, toute loi est produit, et l'invariance par rotation devient la symétrie, or toute loi centrée symétrique n'est pas gaussienne, le théorème n'a donc pas lieu pour $n = 1$. La loi de Rademacher $\frac{1}{2}\delta_{-\sigma} + \frac{1}{2}\delta_\sigma$ est centrée et symétrique, ses deux premiers moments, 0 et σ^2, sont identiques à ceux de γ_σ^1, et on pourrait la considérer comme une gaussienne binaire, pour rendre hommage à Maxwell!
— La motivation de Maxwell était de deviner la distribution des vitesses des molécules d'un gaz isolé et à l'équilibre. Dans ce cas, $n = 3$, et les contraintes d'indépendance des trois composantes et d'invariance par rotation, naturelles du point de vue physique, imposent finalement le caractère gaussien isotrope. Cette découverte a fait que les gaussiennes isotropes sont appelées lois de Maxwell en théorie cinétique des gaz, ancêtre de la physique statistique et de la mécanique statistique. Le caractère gaussien de la distribution des vitesses a été retrouvé ensuite par Boltzmann par une méthode variationnelle, abordée plus loin, tout aussi élémentaire, et d'un impact important bien au-delà du cadre initial.
— Il existe une version pour les matrices aléatoires : si X est une matrice symétrique réelle aléatoire dont les coefficients sont indépendants et dont la loi est invariante par conjugaison par toute matrice orthogonale alors la loi de X a une densité gaussienne de la forme $x \mapsto c\exp(-a\mathrm{Tr}(x^2) + b\mathrm{Tr}(x))$ pour des constantes $a > 0$, $b \in \mathbb{R}$, $c = c(a,b)$, le cas $b = 0$ est appelé Gaussian Orthogonal Ensemble (GOE).

Démonstration. Pour tout $\sigma > 0$, la loi $\gamma_\sigma^n = \mathcal{N}(0, \sigma^2 I_n) = \mathcal{N}(0,\sigma^2)^{\otimes n}$, de densité $\mathrm{e}^{-\frac{|x|^2}{2\sigma^2} - n\log\sqrt{2\pi\sigma^2}}$ est produit et invariante par rotation. Réciproquement, soit μ une loi produit et invariante par rotation. Quitte à remplacer μ par $\mu * \mathcal{N}(0, \varepsilon I_n)$, $\varepsilon > 0$, c'est-à-dire $X + \varepsilon Z$ en terme de variables aléatoires, on peut supposer sans perte de généralité [3] que μ a une densité lisse $f : \mathbb{R}^n \to (0,\infty)$. L'invariance par rotation donne $\log(f(x)) = g(|x|^2)$ et donc

$$\partial_i \log(f(x)) = 2g'(|x|^2)x_i.$$

D'autre part la nature produit donne $\log(f(x)) = h_1(x_1) + \cdots + h_n(x_n)$ et donc

$$\partial_i \log(f(x)) = h_i'(x_i).$$

Donc $\partial_i \log(f(x))$, qui dépend de $|x|$ via $g'(|x|)$, ne dépend que de x_i. Comme $n \geq 2$, il vient que g' est constante (il faut au moins deux coordonnées!). Comme g' est constante, il existe $a, b \in \mathbb{R}$ tels que $g(u) = au + b$ pour tout u, et donc $f(x) = \mathrm{e}^{a|x|^2+b}$. Comme f est une densité, $a < 0$ et $\mathrm{e}^b = (\pi/a)^{-n/2}$, et $\mu = \gamma_\sigma^n$ avec $\sigma^2 = -1/(2a)$. □

Remarque 1.2.8. Entropie à la Boltzmann ou Shannon.

Pour toute loi de probabilité μ de $\mathbb{R}^n$, on pose

$$\mathrm{S}(\mu) := -\int f(x)\log(f(x))\mathrm{d}x$$

si $\frac{\mathrm{d}\mu(x)}{\mathrm{d}x} = f(x)$ et $f\log f \in \mathrm{L}^1(\mathrm{d}x)$, et $\mathrm{S}(\mu) := +\infty$ sinon. On a

$$\mathrm{S}(\mathcal{N}(m,K)) = \log\sqrt{(2\pi\mathrm{e})^n \det K}.$$

C'est l'entropie de Boltzmann (physique statistique) et aussi l'entropie de Shannon [a] (théorie de l'information). Il est utile et éclairant d'introduire l'entropie exponentielle de Shannon

$$\mathrm{N}(\mu) := \frac{\mathrm{e}^{\frac{2}{n}\mathrm{S}(\mu)}}{2\pi\mathrm{e}}, \text{ de sorte que } \mathrm{N}(\mathcal{N}(m,K)) = (\det K)^{1/n}, \text{ et en particulier } \mathrm{N}(\gamma_\sigma^n) = \sigma^2.$$

Liaison entre désordre, information, et volume. Pour en savoir plus : [2, Ch. 10], [81], [27]. La fonction S apparaît en analyse via $\partial_p \|f\|_p^p = \langle f^p, \log(f)\rangle$, en physique statistique et en théorie de l'information via

3. Comme $X_\varepsilon := X + \varepsilon Z$ et Z sont dans L^2 car gaussiens, et centrés, le vecteur X l'est aussi. Si K est la matrice de covariance de X, alors la matrice de covariance de X_ε est $K_\varepsilon := K + \varepsilon^2 I_n$, et donc pour $t \in \mathbb{R}^n$, par indépendance, $\varphi_X(t) = \varphi_{X_\varepsilon}(t)/\varphi_Z(\varepsilon t) = \exp(-\frac{1}{2}\langle (K_\varepsilon - \varepsilon^2 I_n)t, t\rangle) = \varphi_{\mathcal{N}(0,K)}(t)$. Cela ne nécessite pas $\varepsilon \to 0$. Alternativement, quand $\varepsilon \to 0$, d'une part $\varphi_{X_\varepsilon}(t) = \varphi_X(t)\varphi_Z(\varepsilon t) \to \varphi_X(t)$ et d'autre part les deux premiers moments de X_ε convergent vers ceux de X, et comme X_ε est gaussien, on obtient que X l'est aussi.

l'analyse asymptotique combinatoire (découle de la formule de Stirling $n! \sim \sqrt{2\pi n}(n/\mathrm{e})^n$)

$$\frac{1}{n}\log\binom{n}{n_1,\dots,n_r} \longrightarrow \mathrm{S}(p) := -\sum_{i=1}^{r} p_i \log(p_i) \quad \text{c'est-à-dire} \quad \binom{n}{n_1,\dots,n_r} \approx \mathrm{e}^{n\mathrm{S}(p)},$$

quand $n := n_1 + \cdots + n_r \to \infty$ avec $n_i/n \to p_i$, en statistique via la log-vraisemblance, etc. Certains de ces aspects sont revisités dans le chapitre 4 de ces notes. L'entropie S, mesure extensive à la fois du désordre et de l'information, peut également être caractérisée par un nombre restreint d'axiomes naturels pour son interprétation, incluant notamment la sous-additivité.

a. « *My greatest concern was what to call it. I thought of calling it 'information'. But the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. He told me : "You should call it entropy, for two reasons. In first place your uncertainty has been used in statistical mechanics under that name, so it already has a name. In second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage."* ». Claude Shannon (1916 – 2001), à propos de l'entropie qui porte son nom en théorie de l'information, cité dans [81], lui même citant [57].

Théorème 1.2.9. Caractérisation variationnelle de Boltzmann des gaussiennes.

Soit μ une loi de probabilité sur $\mathbb{R}^n$, $n \geq 1$, d'entropie finie : $\mathrm{S}(\mu) < \infty$.
Si $\int |x|^2 \mathrm{d}\mu(x) = \int |x|^2 \mathrm{d}\gamma_\sigma^n(x)$ pour un $\sigma > 0$, alors $\mathrm{S}(\mu) \leq \mathrm{S}(\gamma_\sigma^n)$ avec égalité ssi $\mu = \gamma_\sigma^n$.

— L'entropie étant invariante par translation, la contrainte porte donc ici sur la variance. La gaussienne γ_σ^n réalise le désordre maximal à variance fixée.
— Théorème-H de Boltzmann : loi de conservation de l'énergie (second moment) et monotonie de l'entropie le long de l'équation d'évolution de la densité (équation de Boltzmann), d'où l'équilibre gaussien isotrope en temps grand par caractérisation variationnelle des gaussiennes comme maximum d'entropie à second moment fixé. Ce caractère gaussien de l'équilibre était déjà connu de Maxwell, directement, en utilisant les invariances de l'équilibre et la caractérisation géométrique des gaussiennes isotropes.
— Point de vue de Shannon sur le TLC inspiré du théorème-H de Boltzmann : loi de conservation du second moment et monotonie de l'entropie, d'où convergence vers maximum d'entropie à variance fixée. Le TLC est donc vu comme une équation d'évolution de loi, à temps discret. Pour en savoir plus : [23].
— Les lois obtenues par maximum d'entropie sous contrainte de moment sont essentielles en statistique bayésienne, en rapport avec la dualité entre lois a priori et a posteriori relativement à l'observation. L'entropie y apparaît via la log-vraisemblance qu'on maximise aussi. Pour en savoir plus : [70] et [27].

Démonstration. On écrit, en notant $f_\sigma = \mathrm{e}^{-\frac{|\cdot|^2}{2\sigma^2} - c_n}$ la densité de γ_σ^n, $c_n := n\log\sqrt{2\pi\sigma^2}$, et f la densité de μ,

$$\begin{aligned}
\mathrm{S}(\gamma_\sigma^n) - \mathrm{S}(\mu) &= -\int f_\sigma(x)\Big(-\frac{|x|^2}{2\sigma^2} - c_n\Big)\mathrm{d}x + \int f(x)\log(f(x))\mathrm{d}x \\
&= -\int f(x)\Big(-\frac{|x|^2}{2\sigma^2} - c_n\Big)\mathrm{d}x + \int f(x)\log(f(x))\mathrm{d}x \\
&= -\int f(x)\log(f_\sigma(x))\mathrm{d}x + \int f(x)\log(f(x))\mathrm{d}x \\
&= \int \frac{f(x)}{f_\sigma(x)}\log\frac{f(x)}{f_\sigma(x)} f_\sigma(x)\mathrm{d}x = \mathbb{E}(\Phi(Z_\sigma)) \geq \Phi(\mathbb{E}(Z_\sigma)) = \Phi(1) = 0
\end{aligned}$$

par l'inégalité de Jensen pour la fonction strictement convexe $\Phi(u) = u\log(u)$, avec égalité ssi $f = f_\sigma$. □

— Soit $K \subset \mathbb{R}^n$ un compact non-vide, adhérence de son intérieur. Le maximum d'entropie parmi les lois à support dans K est donné par la loi uniforme sur K, de densité par rapport à la mesure de Lebesgue

$$\frac{1}{|K|}\mathbb{1}_K$$

où $|K|$ est la mesure de Lebesgue ou volume de K. Il s'agit de la caractérisation variationnelle de Boltzmann de la mesure de Boltzmann–Gibbs $\frac{1}{Z}\mathrm{e}^{-V}$ où $V := 0$ sur K et $V := +\infty$ hors de K, tandis que $Z = |K|$.

— La caractérisation variationnelle de Boltzmann se généralise bien au-delà des gaussiennes aux lois de Boltzmann–Gibbs : si $\mathrm{d}\mu_\beta(x) := \frac{1}{Z_\beta}\mathrm{e}^{-\beta V(x)}\mathrm{d}x$ et $\int V(x)\mathrm{d}\mu(x) = \int V(x)\mathrm{d}\mu_\beta(x)$ et $\mathrm{S}(\mu) < \infty$ alors [4]

$$\mathrm{S}(\mu) \leq \mathrm{S}(\mu_\beta) \quad \text{avec égalité ssi } \mu = \mu_\beta.$$

On a alors $\mathrm{S}(\mu_\beta) = \beta\int V\mathrm{d}\mu_\beta + \log Z_\beta$. La preuve est identique : on vérifie que $\mathrm{S}(\mu_\beta) - \mathrm{S}(\mu) = \mathrm{H}(\mu \mid \mu_\beta)$ où H est l'entropie relative ou divergence de Kullback–Leibler (statistique, théorie de l'information) :

$$\mathrm{H}(\nu \mid \mu) = \int \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\log\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\mathrm{d}\mu = \int \log\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\mathrm{d}\nu \geq 0 \quad \text{avec égalité ssi } \nu = \mu$$

et la convention $\mathrm{H}(\nu \mid \mu) = +\infty$ si $\nu \not\ll \mu$ ou si l'intégrale diverge. En pratique, on fixe une énergie moyenne m admissible au sens où il existe $\beta = \beta_m$ tel que $\int V\mathrm{d}\mu_\beta = m$, et on obtient que la loi μ_β maximise l'entropie parmi les lois d'énergie moyenne m et d'entropie finie. Plutôt que de fixer l'énergie moyenne, on peut fixer β. Cela revient à adopter un point de vue lagrangien en incorporant la contrainte dans la fonctionnelle : l'énergie libre de Helmholtz (énergie moyenne − température × entropie) :

$$\mathrm{F}(\nu) := \int V\mathrm{d}\nu - \frac{1}{\beta}\mathrm{S}(\nu).$$

On a alors $\mathrm{F}(\mu_\beta) = -\frac{1}{\beta}\log(Z_\beta)$, tandis que la mesure de Boltzmann–Gibbs μ_β est un minimum global :

$$\mathrm{F}(\mu) - \mathrm{F}(\mu_\beta) = \frac{1}{\beta}\mathrm{H}(\mu \mid \mu_\beta) \geq 0 \quad \text{avec égalité ssi } \mu = \mu_\beta.$$

L'énergie libre peut également être vue comme une transformée de Legendre, cf. remarque 2.5.6, ou comme une fonction génératrice des moments, elle joue un rôle clé en mécanique statistique.

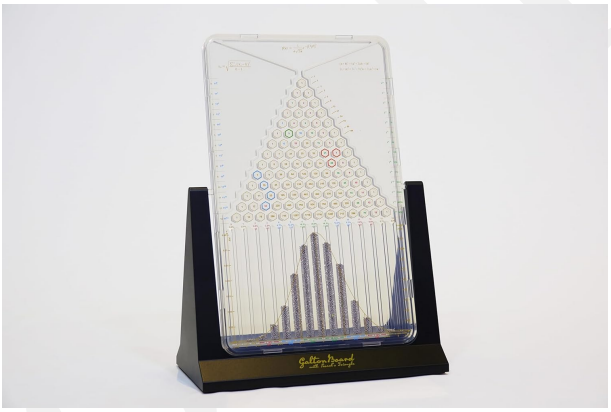


FIGURE 1.1 – Planche de Galton, dispositif expérimental portable illustrant le phénomène TLC.

1.3 Théorème limite central

La variance d'une somme de v.a.r. i.i.d. $X_1, \ldots, X_n$ de variance σ^2 est linéaire en n :

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i,j=1}^{n} \mathrm{Cov}(X_i, X_j) = n\sigma^2.$$

Si $X_1, \ldots, X_n$ sont des v.a.r. i.i.d. de carré intégrable de moyenne m et de variance σ^2 alors

$$\mathbb{E}\Big(\frac{X_1 + \cdots + X_n}{n}\Big) = m \quad \text{et} \quad \mathbb{E}\Big(\Big(\frac{X_1 + \cdots + X_n}{n} - m\Big)^2\Big) = \mathrm{Var}\Big(\frac{X_1 + \cdots + X_n}{n}\Big) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$$

d'où le phénomène loi faible des grands nombres :

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{L^2} m, \quad \text{en particulier} \quad \frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{\mathbb{P}} m.$$

Dilater $\frac{X_1+\cdots+X_n}{n} - m$ par $\sqrt{n}$ stabilise la variance, et même la loi : c'est le phénomène limite central [5] :

4. À retenir au passage : la température dans une mesure de Boltzmann–Gibbs joue le même rôle qu'une variance de gaussienne.

5. Henri Poincaré (1854 – 1912), à propos de la loi normale du théorème limite central, désigné sous le nom de loi des erreurs à l'époque : « *Cette loi ne s'obtient pas par des déductions rigoureuses ; plus d'une démonstration qu'on a voulu en donner est grossière, entre autres celle*

Théorème 1.3.1. Théorème limite central (TLC).

Si $(X_n)_{n\geq 1}$ est une suite de v.a.r. i.i.d. de carré intégrable, de moyenne m et de variance $\sigma^2 > 0$, alors

$$\underbrace{\frac{\sqrt{n}}{\sigma}\Big(\frac{X_1+\cdots+X_n}{n}-m\Big)}_{\substack{\text{point de vue statistique}\\ n=\text{taille échantillon}\\ \text{fluctuation de moyenne empirique}}} = \underbrace{\Big\langle \frac{X-m}{\sigma}, 1_n\Big\rangle}_{\substack{\text{point de vue géométrique}\\ n=\text{dimension}\\ \text{projection unidimensionnelle}\\ 1_n=\left(\frac{1}{\sqrt{n}},\dots,\frac{1}{\sqrt{n}}\right)\in\mathbb{S}^{n-1}}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

Plus généralement, si $(X_n)_{n\geq 1}$ est une suite de vecteurs aléatoires de $\mathbb{R}^d$, indépendants et identiquement distribués, de carré intégrable avec un vecteur moyenne m et une matrice de covariance K, alors

$$\sqrt{n}\Big(\frac{X_1+\cdots+X_n}{n}-m\Big) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,K).$$

— Le TLC est un théorème limite [6] : analyse asymptotique quand $n\to\infty$: $S_n := X_1+\cdots+X_n \approx mn+\sqrt{n}\sigma Z$ Du second ordre au sens où il s'agit d'étudier l'échelle et la distribution des fluctuations pour la LGN. Pour la marche aléatoire $(S_n)_{n\geq 1}$: comportement linéaire dû à la moyenne et correction ou fluctuations asymptotiquement gaussiennes d'ordre $\sqrt{n}$. Outre le cas gaussien, voici deux autres cas remarquables :
  — Si $X_1 \sim \text{Bernoulli}_p(\pm 1) = p\delta_1+(1-p)\delta_{-1}$ alors S_n a la loi de $\varphi(U)$ où $\varphi(x)=2x-n$ et $U\sim\text{Binom}(n,p)$. Dans ce cas S_n prend ses valeurs dans $[-n,n]$, tandis que $(m,\sigma^2)=(p,p(1-p))$. Le cas Rademacher $p=1/2$ donne $(m,\sigma^2)=(0,1)$ comme pour la gaussienne $\mathcal{N}(0,1)$.
  — Si $X_1\sim\text{Gamma}(a,\lambda)$ alors $S_n\sim\text{Gamma}(na,\lambda)$, $(m,\sigma^2)=(a/\lambda,a/\lambda^2)$, $\frac{S_n}{\sqrt{n}}\sim\text{Gamma}(na,\sqrt{n}\lambda)$.
— Stabilité des gaussiennes : si $X_i\sim\mathcal{N}(0,1)$ alors $(m,\sigma^2)=(0,1)$, et $\frac{X_1+\cdots+X_n}{\sqrt{n}}\sim\mathcal{N}(0,1)$, non-asymptotique.
— Phénomène d'universalité : la loi limite $\mathcal{N}(0,1)$ ne dépend pas de la loi des X_n.
— La normalisation en translation/dilatation ou position/échelle du TLC assure la conservation des deux premiers moments, calibrés pour être égaux aux deux premiers moments de la loi limite $\mathcal{N}(0,1)$:

$$\mathbb{E}(Z_n)=0=\mathbb{E}(Z) \quad \text{et} \quad \mathbb{E}(Z_n^2)=1=\mathbb{E}(Z^2), \quad \text{où} \quad Z_n := \frac{\sqrt{n}}{\sigma}\Big(\frac{X_1+\cdots+X_n}{n}-m\Big) \quad \text{et} \quad Z\sim\mathcal{N}(0,1).$$

— Statistiquement, la moyenne empirique $\widehat{m}_n := \frac{X_1+\cdots+X_n}{n}$ est un estimateur de la moyenne m, sans biais car $\mathbb{E}(\widehat{m}_n)=m$, consistant car $\widehat{m}_n\to m$ dans L^2, et de vitesse $\sqrt{n}$ car $\sqrt{n}(\widehat{m}_n-m)\to\mathcal{N}(0,1)$ en loi.
— Géométriquement, peut-on remplacer dans $\langle\frac{X-m}{\sigma},1_n\rangle$ le vecteur $1_n=(\frac{1}{\sqrt{n}},\dots,\frac{1}{\sqrt{n}})\in\mathbb{S}^{n-1}$ par un vecteur $\theta\in\mathbb{S}^{n-1}$ quelconque. La réponse est évidemment négative si θ est un vecteur de la base canonique, en revanche elle est positive si θ est suffisamment délocalisé : TLC à profil de variance abordé en TD.
— Si $Y_1,\dots,Y_n$ sont des v.a.r. i.i.d. avec $\mathbb{E}(Y_1)=0$, $\mathbb{E}(Y_1^2)=1$, alors dans L^2, $Y_1,\dots,Y_n$ sont des vecteurs orthogonaux, de norme unité, et la diagonale de leur parallélépipède est de longueur $\sqrt{n}$:

$$\mathbb{E}\Big(\Big(\frac{Y_1+\cdots+Y_n}{\sqrt{n}}\Big)^2\Big)=1, \quad \text{par exemple} \quad Y_i := \frac{X_i-m}{\sigma}.$$

— Le TLC est en fait un théorème sur un tableau triangulaire de v.a. (la convergence en loi est marginale).
— Le TLC repose sur la finitude du moment d'ordre 2, mais se généralise au-delà en modifiant la normalisation (TLC stable). Par ailleurs, il existe une sorte de réciproque au TLC : si les $(X_n)_{n\geq 1}$ sont des v.a.r. i.i.d. intégrables avec $\mathbb{E}(X_1)=0$ et telles que $\frac{X_1+\cdots+X_n}{\sqrt{n}}\to\mathcal{N}(0,\sigma^2)$ en loi quand $n\to\infty$, pour un certain $\sigma^2>0$, alors les X_i sont de carré intégrable et de variance σ^2. Pour en savoir plus : [62].

qui s'appuie sur l'affirmation que la probabilité des écarts est proportionnelle aux écarts. Tout le monde y croit cependant, me disait un jour M. Lippmann, car les expérimentateurs s'imaginent que c'est un théorème de mathématiques, et les mathématiciens que c'est un fait expérimental. » In Le calcul des Probabilités (1896) p. 149. Comme le dit le physicien Oriol Bohigas (1937 – 2013), « *De nos jours, nous savons que c'est à la fois un fait expérimental et un théorème de mathématiques!* ».

6. Le théorème limite central a été développé notamment par Abraham Moivre (1667 – 1754), Pierre-Simon Laplace (1749 – 1827), Pafnouti Tchebychev (1821 – 1894), Alexandre Liapounov (1857 – 1918), Jarl Lindeberg (1876 – 1932), et Paul Lévy (1886 – 1971). George Pólya (1887 – 1985) publie pendant les années folles un article intitulé « Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem », où il qualifie ce « théorème limite » (Grenzwertsatz) de « zentral » (primordial, principal) en raison de son caractère universel. Si la traduction anglaise semble fixée (« central limit theorem » souvent abrégé en CLT), l'usage francophone, partant de la traduction littérale « théorème limite central » (TLC), attribue souvent la centralité à la limite en parlant de « théorème de la limite centrale », allant parfois même jusqu'à l'audacieux « théorème central limite » (TCL). Note de bas de page adaptée de [26].

— Le TLC contient toujours une loi faible des grands nombres car du TLC on tire

$$\frac{X_1-m+\cdots+X_n-m}{n} \xrightarrow[n\to\infty]{\text{loi}} \delta_0 \quad \text{autrement dit} \quad \frac{X_1+\cdots+X_n}{n} \xrightarrow[n\to\infty]{\mathbb{P}} m.$$

— Version condensée du TLC, analytique, tweetée par Gabriel Peyré :

$$\int (1,x,x^2)f(x)\mathrm{d}x=(1,0,1) \quad \Rightarrow \quad f^{*n}\Big(\frac{x}{\sqrt{n}}\Big) \xrightarrow[n\to\infty]{\text{loi}} \frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\mathrm{d}x.$$

Démonstration. On se ramène pas translation et dilatation au cas $m=0$ et $\Sigma=I_n$. Donnons la preuve du cas $d=1$, celle du cas $d>1$ étant semblable. L'idée est de montrer asymptotiquement, les cumulants d'ordre >2 s'annulent. Comme $\varphi_{X_i}(s)=1+\frac{s^2}{2}+o(s^2)$ car $\mathbb{E}(X_i)=0$ et $\mathbb{E}(X_i^2)=1$, il vient, grâce au caractère i.i.d. des X_i,

$$\varphi_{\frac{X_1+\cdots+X_n}{\sqrt{n}}}(t)=\mathbb{E}(\mathrm{e}^{\mathrm{i}\frac{t}{\sqrt{n}}(X_1+\cdots+X_n)})=\prod_{i=1}^n \mathbb{E}(\mathrm{e}^{\mathrm{i}\frac{t}{\sqrt{n}}X_i})=\varphi_{X_1}\big(\tfrac{t}{\sqrt{n}}\big)^n=(1+\tfrac{t^2}{2n}+o(\tfrac{1}{n}))^n=\tfrac{t^2}{2}+o(1)\xrightarrow[n\to\infty]{}\varphi_{\mathcal{N}(0,1)}(t).$$

□

Remarque 1.3.2. Quelques variantes célèbres du TLC.

1. Profil de variance. Soit $(X_n)_{n\geq 1}$ des v.a.r. indépendantes telles que $\mathbb{E}(X_n)=0$ et $\sigma_n^2:=\mathbb{E}(X_n^2)$. Soit $s_n^2:=\mathrm{Var}(X_1+\cdots+X_n)=\sigma_1^2+\cdots+\sigma_n^2$. Si la condition de Lindeberg est vérifiée : $\forall\varepsilon>0$,

$$\lim_{n\to\infty}\frac{1}{s_n^2}\sum_{k=1}^n \mathbb{E}(X_k^2\mathbb{1}_{\{|X_k|>\varepsilon s_n\}})=0, \quad \text{alors } \frac{X_1+\cdots+X_n}{s_n}\xrightarrow[n\to\infty]{\text{loi}}\mathcal{N}(0,1).$$

   — La réciproque est vraie lorsque la condition de Feller est vérifiée : $\lim_{n\to\infty}\frac{\sigma_k^2}{s_n^2}=0$.

   — Si $\tau_n:=\mathbb{E}(|X_n|^3)$ alors on a un TLC uniforme (inégalité de Berry–Esseen) :

$$\sup_{x\in\mathbb{R}}\Big|\mathbb{P}\Big(\frac{X_1+\cdots+X_n}{s_n}\leq x\Big)-\mathbb{P}(Z\leq x)\Big|\leq C\frac{\mathbb{E}(|X|_3^3)}{\mathbb{E}(|X|_2^2)^{3/2}}=C\frac{\tau_1^2+\cdots+\tau_n^3}{(\sigma_1^2+\cdots+\sigma_n^2)^{3/2}}.$$

   où C est une constante universelle et où $Z\sim\mathcal{N}(0,1)$. Phénomène TLC quand la variance totale s_n^2 est délocalisée sur les coordonnées. Le phénomène TLC n'apparaît pas si la variance totale s_n^2 est très localisée (sparsité). Localisation et sparsité. Pour en savoir plus : [38] et TD.

2. Variables dépendantes.

   — Corps convexes. Soit X un vecteur de $\mathbb{R}^n$ de loi uniforme sur un corps convexe (convexe compact non-vide) K, centré $\mathbb{E}(X)=0$ et isotrope normalisé $\mathbb{E}(XX^\top)=I_n$. Soit $\theta\in\mathbb{S}^{n-1}$. Alors $\langle X,\theta\rangle\to\mathcal{N}(0,1)$ en loi quand $n\to\infty$, du moment que θ est suffisamment délocalisé. Dans le cas du cube [a] $K=(\sqrt{3}[-1,1])^n$, les coordonnées sont i.i.d. et le résultat découle par exemple de l'inégalité de Berry–Esseen, et dans ce cas il est clair que le résultat n'a pas lieu pour tout $\theta\in\mathbb{S}^{n-1}$, car il suffit de prendre par exemple $\theta=e_i$. Pour en savoir plus : [46, 47].

   — Martingales et Markov. Le phénomène TLC a lieu pour les martingales de carré intégrable, ainsi que pour les fonctionnelles additives des processus de Markov ergodiques, et plus généralement encore comme analyse asymptotique des fluctuations dans le théorème ergodique. Pour en savoir plus : cours de M1 S1 *Processus stochastiques* et *Systèmes dynamiques*.

   — Matrices aléatoires. Si M est une matrice aléatoire $n\times n$ remplie de variables aléatoires gaussiennes complexes i.i.d. $\mathcal{N}(0,\frac{1}{2n}I_2)$ de sorte que $\mathbb{E}(M_{jk})=0$ et $\mathbb{E}(|M_{jk}|^2)=\frac{1}{n}$, et si $\lambda_1,\ldots,\lambda_n$ sont ses valeurs propres, alors pour toute fonction $f:\mathbb{C}\to\mathbb{R}$ à support dans le disque unité,

$$\sum_{k=1}^n (f(\lambda_k)-\mathbb{E}(f(\lambda_k)))\xrightarrow[n\to\infty]{\text{loi}}\mathcal{N}\Big(0,\frac{1}{4\pi}\int|\nabla f(x)|^2\mathrm{d}x\Big).$$

   On notera l'absence de normalisation due à la répulsion! Pour en savoir plus : [69].

3. Cadre algébrique. Il est possible d'établir le TLC en utilisant la méthode des moments. En remplaçant la notion de variable aléatoire par la notion d'élément d'algèbre, la notion d'espérance par une forme linéaire, la notion de loi par la notion de moments, la notion d'indépendance par une notion de liberté, Dan Virgil Voiculescu a établi un TLC algébrique, dans lequel la loi du demi-cercle joue le rôle de la loi gaussienne. Pour en savoir plus : [59].

a. La loi uniforme sur $[-1,1]$ a pour densité $\frac{\mathbb{1}_{[-1,1]}}{2}$, de moyenne 0 et de variance $\frac{1}{2}\int_{-1}^1 x^2\mathrm{d}x=\frac{1}{2}[\frac{x^3}{3}]_{-1}^1=\frac{1}{3}$.

Remarque 1.3.3. Quelques approches alternatives célèbres pour le TLC.

1. Principe de remplacement de Lindeberg. Soient $Z_1, \dots, Z_n$ des v.a.r. i.i.d. de loi $\mathcal{N}(0,1)$ et $f : \mathbb{R} \to \mathbb{R}$ une fonction test lisse. On remplace les X_i par les Z_i un à un, via l'interpolation télescopique

$$f\Big(\frac{X_1+\cdots+X_n}{\sqrt{n}}\Big) - f\Big(\frac{Z_1+\cdots+Z_n}{\sqrt{n}}\Big) = \sum_{i=0}^{n-1} (f(R_i) - f(R_{i+1})) \quad \text{où} \quad R_i := \frac{Z_1+\cdots+Z_i+X_{i+1}+\cdots+X_n}{\sqrt{n}}.$$

   Ensuite on utilise une formule de Taylor à l'ordre 2 pour contrôler $f(R_i) - f(R_{i+1})$ et on se sert de l'égalité des deux premiers moments pour se ramener au seul contrôle du reste.

2. Méthode de Stein. La gaussienne $\mathcal{N}(0,1)$ est caractérisée par l'intégration par parties gaussienne

$$\mathbb{E}(f'(Z)) = \mathbb{E}(Zf(Z)) \quad \text{pour } Z \sim \mathcal{N}(0,1) \text{ et toute fonction test } f \in \mathscr{C}^1_c.$$

   Cela donne l'équation de Stein $\mathbb{E}(A_Z(f)) = 0$ où $A_z(f) := f'(z) - zf(z)$ est l'opérateur de Stein. L'idée est alors d'étudier une version où Z est remplacé par $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ et d'en déduire le TLC, dans une version quantitative du type $\sup_{f\in\mathscr{F}} \mathbb{E}(A_{\frac{X_1+\cdots+X_n}{\sqrt{n}}}(f)) \xrightarrow[n\to\infty]{} 0$ pour une classe $\mathscr{F}$ bien choisie.

Pour en savoir plus dans le contexte des probabilités en grande dimension : [77].

Remarque 1.3.4. Marche aléatoire, principe d'invariance de Donsker, mouvement Brownien.

La marche aléatoire $(S_n)_{n\geq 1}$ est définie par

$$S_n := X_1 + \cdots + X_n$$

où $(X_n)_{n\geq 1}$ sont des v.a.r. i.i.d. Supposons que $\mathbb{E}(X_1) = 0$ et $\mathbb{E}(X_1^2) = 1$. Le TLC donne, pour tout $t > 0$,

$$B_t^{(n)} := \frac{S_{\lfloor tn \rfloor}}{\sqrt{n}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,t).$$

Plus généralement, pour tout entier $k \geq 1$ et tous réels $0 \leq t_1 < \cdots < t_k$,

$$(B_{t_1}^{(n)}, \dots, B_{t_k}^{(n)}) \xrightarrow[n\to\infty]{\text{loi}} \text{Loi}(B_{t_1}, \dots, B_{t_k})$$

où $(B_t)_{t\in\mathbb{R}_+}$ est une famille de v.a.r. à accroissements gaussiens indépendants et stationnaires : pour tout $k \geq 1$ et tous $0 =: t_0 < t_1 < \ldots < t_k := t$, on a la somme télescopique en accroissements

$$B_t = B_{t_1} - B_{t_0} + \cdots + B_{t_k} - B_{t_{k-1}}$$

avec $B_{t_1} - B_{t_0}, \dots, B_{t_k} - B_{t_{k-1}}$ indépendantes et de lois respectives $\mathcal{N}(0, t_1 - t_0), \dots, \mathcal{N}(0, t_k - t_{k-1})$.
Cela exprime une convergence en loi de la suite de processus $(B^{(n)})_{n\geq 1}$ vers un processus limite B, qui est universel : ne dépend pas de la loi des ingrédients initiaux. C'est le principe d'invariance de Donsker. Il est possible de construire le processus limite de sorte que ses trajectoires soient continues, et on parle alors de mouvement brownien. Sa loi est une gaussienne sur l'espace de dimension infinie $\mathscr{C}(\mathbb{R}_+, \mathbb{R})$.
On peut relier le principe de Donsker au théorème de Kolmogorov–Smirnov abordé en TD. D'autre part la convergence peut être renforcée (topologie uniforme de Skorokhod). Pour aller plus loin [a] : [48].

a. Et encore plus loin : un théorème de Komlós–Major–Tusnády affirme qu'on peut coupler la marche aléatoire et le mouvement brownien et contrôler leur écart uniformément en temps et presque sûrement, on parle d'approximation forte KMT.

1.4 Loi des grands nombres

Théorème 1.4.1. Loi des grands nombres.

Si $(X_n)_{n\geq 1}$ est une suite de vecteurs aléatoires de $\mathbb{R}^d$ i.i.d. intégrables de vecteur moyenne m, alors

$$\frac{X_1+\cdots+X_n}{n} \underset{n\to\infty}{\longrightarrow} m \quad \text{presque sûrement et dans } L^1.$$

— La LGN est un théorème limite [7], comme le TLC.
En combinant LGN et TLC on obtient l'analyse asymptotique $X_1+\cdots+X_n \approx nm+\sqrt{n}\sigma Z$, $Z\sim\mathcal{N}(0,1)$.
— On parle de loi forte des grands nombres car la convergence est p.s. et la condition de moment optimale.
— Point de vue statistique : la moyenne empirique est un estimateur fortement consistant de la moyenne. Combiné au TLC, cela fournit un test statistique pour tester l'hypothèse $m=0$ par exemple.
— La preuve subsiste pour des v.a.r. indépendantes centrées, bornées dans L^4, pas forcément de même loi.
— Réciproque : si $(X_n)_{n\geq 1}$ sont i.i.d. et $\frac{X_1+\cdots+X_n}{n}\to m$ p.s. quand $n\to\infty$, pour un réel m, alors les X_n sont intégrables et de moyenne m. En effet, $\frac{X_n}{n}=\frac{S_n}{n}-\frac{n-1}{n}\frac{S_{n-1}}{n-1}$ converge p.s. vers 0. Or comme les (X_n) sont indépendantes, le lemme de Borel–Cantelli donne $\sum_n \mathbb{P}(|X_n|\geq \varepsilon n)<\infty$ pour tout $\varepsilon>0$, et comme les (X_n) sont de même loi, on obtient $\sum_n \mathbb{P}(|X_1|\geq \varepsilon n)<\infty$, d'où $\mathbb{E}(|X_1|)<\infty$.
— Lois stables : Si les X_i suivent la loi de Cauchy standard, alors on peut vérifier que $\frac{X_1+\cdots+X_n}{n}$ suit la même loi, et en particulier ne peut pas converger vers 0. Il n'y a pas de contradiction car la loi de Cauchy n'a pas d'espérance. Sa propriété de stabilité par convolution à l'échelle n est analogue de celle de la loi gaussienne $\mathcal{N}(0,1)$ à l'échelle $\sqrt{n}$. L'étude de telles loi stables est abordée plus loin.
— La convergence complète nécessite plus d'intégrabilité : moment exponentiel par exemple.

Démonstration. Cette preuve est essentiellement tirée de [76].
— On se ramène à $m=0$ par translation et à $d=1$ en appliquant le cas $d=1$ à chacune des d coordonnées.
— Cas où les X_i sont indépendantes, centrées, et bornées dans L^4 (pas forcément de même loi). En posant $M_n:=\frac{X_1+\cdots+X_n}{n}$, il vient, par centrage, Cauchy–Schwarz, et bornitude dans L^4,

$$\mathbb{E}(M_n^4)=\frac{1}{n^4}\sum_{i,j,k,l}\mathbb{E}(X_iX_jX_kX_l)=\frac{1}{n^4}(\mathbb{E}(X_1^4)+12n(n-1)\mathbb{E}(X_1^2X_2^2))\leq\frac{1}{n^4}(\mathbb{E}(X_1^4)+12n(n-1)\mathbb{E}(X_1^4)))=O(\tfrac{1}{n^2}),$$

donc $\mathbb{E}(\sum_n M_n^4)<\infty$ donc $\sum_n M_n^4<\infty$ p.s. donc $M_n\to 0$ p.s. comme dans la preuve de Borel–Cantelli. Alternativement, par Markov $\mathbb{P}(M_n^4\geq\varepsilon)=O(\frac{1}{\varepsilon n^2})$, d'où $\sum_n\mathbb{P}(M_n^4\geq\varepsilon)<\infty$, puis Borel–Cantelli.
— Lemme de troncature. Si X est intégrable centrée alors pour tout $\delta>0$, il existe Y centrée bornée telle que $\mathbb{E}(|X-Y|)\leq\varepsilon$. En effet, par convergence dominée, pour tout $\delta>0$, il existe $r>0$ tel que $\mathbb{E}(|X|\mathbb{1}_{|X|>r})\leq\delta$. Soit $Y':=X\mathbb{1}_{|X|\leq r}$ et $Y:=Y'-\mathbb{E}(Y')$. Alors

$$|Y|\leq C:=2r,\quad \mathbb{E}(Y)=0,\quad \mathbb{E}(|X-Y|)\leq\mathbb{E}(|X-Y'|)+|\mathbb{E}(X)-\mathbb{E}(Y')|\leq 2\mathbb{E}(|X-Y'|)\leq 2\delta.$$

— Cas de X_i i.i.d. centrées et intégrables. Soit $Y_i:=X_i\mathbb{1}_{|X_i|>r}-\mathbb{E}(X_i\mathbb{1}_{|X_i|>r})$ avec $r=r(\delta)$, $S_n:=X_1+\cdots+X_n$, $T_n:=X_1+\cdots+X_n$. Alors les (Y_i) sont centrées et bornées, en particulier bornées dans L^4, donc $\frac{T_n}{n}\to 0$ p.s. et dans L^1 par convergence dominée car les $(\frac{1}{n}T_n)$ sont bornées par r. D'autre part

$$\frac{\mathbb{E}(|S_n-T_n|)}{n}\leq\frac{1}{n}\sum_{i=1}^n\mathbb{E}(|X_i-Y_i|)\leq\delta.$$

D'où $\overline{\lim}_{n\to\infty}\frac{\mathbb{E}(|S_n|)}{n}\leq\delta$. Comme δ est arbitraire, cela donne $\frac{S_n}{n}\to 0$ dans L^1 et donc en $\mathbb{P}$ (loi faible).
— Convergence p.s. (loi forte). Il suffit d'établir que pour une constante $C>0$,

$$\overline{\lim_{n\to\infty}}\frac{Z_1+\cdots+Z_n}{n}\leq C\mathbb{E}(Z_1),\quad \text{où}\quad Z_i:=|X_i-Y_i|.$$

Par dilatation on peut supposer que $\mathbb{E}(Z_1)=1$. Par Borel–Cantelli, il suffit d'établir que

$$\sum_k\mathbb{P}\Big(\max_{2^k<n\leq 2^{k+1}}\frac{Z_1+\cdots+Z_n}{n}\geq C\Big)<\infty,$$

7. La LGN a été étudiée et raffinée par de nombreux mathématiciens, comme Jacques Bernoulli (1654 – 1705), Siméon Denis Poisson (1781 – 1840) qui semble lui avoir donné son nom de LGN avec *loi* comme dans *loi de la nature*, Alexandre Khintchine (1894 – 1959) qui semble avoir optimisé l'hypothèse de moment, Andrei Kolmogorov (1903 – 1987), et bien d'autres. On lit dans [50] que Andreï Markov (1856–1922) a introduit ses fameuses chaînes pour démontrer que la loi des grands nombres restait valide au delà de l'indépendance deux à deux et répondre ainsi aux mystiques pour qui cette émergence d'un déterminisme à partir du hasard était une négation du libre arbitre!

et comme $Z_i \geq 0$ il suffit d'établir que

$$\sum_k \mathbb{P}\Big(\sum_{i=1}^{2^{k+1}} Z_i \geq C2^k\Big) < \infty,$$

En posant $U_i := Z_i \mathbb{1}_{Z_i \leq 2^k}$ et $V_i := U_i - \mathbb{E}(U_i)$, et $C = 3$, il vient

$$\mathbb{P}\Big(\sum_{i=1}^{2^{k+1}} Z_i \geq 3 \cdot 2^k\Big) \leq \mathbb{P}(\exists i \leq 2^{k+1} : Z_i \neq U_i) + \mathbb{P}\Big(\sum_{i=1}^{2^{k+1}} U_i \geq 3 \cdot 2^k\Big).$$

Or comme $\sum_{i=1}^{2^{k+1}} \mathbb{E}(U_i) \leq \sum_{i=1}^{2^{k+1}} \mathbb{E}(Z_i) = 2^{k+1}\mathbb{E}(Z_1) = 2^{k+1}$ il vient

$$\mathbb{P}\Big(\sum_{i=1}^{2^{k+1}} Z_i \geq 3 \cdot 2^k\Big) \leq \mathbb{P}(\exists i \leq 2^{k+1} : Z_i \neq U_i) + \mathbb{P}\Big(\sum_{i=1}^{2^{k+1}} V_i \geq 2^k\Big).$$

Comme les Z_i sont de même loi et intégrables, on a

$$\sum_k \mathbb{P}(\exists i \leq 2^{k+1} : Z_i \neq U_i) \leq \sum_k 2^{k+1}\mathbb{P}(Z_1 > 2^k) < \infty,$$

tandis que par Markov et l'inégalité $\mathbb{E}((W - \mathbb{E}W)^2) \leq \mathbb{E}(W^2)$, il vient

$$\mathbb{P}\Big(\sum_{i=1}^{2^{k+1}} V_i \geq 2^k\Big) \leq \frac{1}{2^{2k}} \sum_{i=1}^{2^{k+1}} \mathbb{E}(V_i^2) \leq \frac{2}{2^k}\mathbb{E}(Z_1^2 \mathbb{1}_{Z_1 \leq 2^k}).$$

Enfin, pour tout $t \geq 1$, $\sum_{k \in A_t} \frac{1}{2^k} \leq \frac{2}{t}$ où $A_t := \{k \geq 0 : 2^k \geq t\}$. D'où, par Fubini–Tonelli ou CV monotone,

$$\sum_{k=0}^{\infty} \frac{1}{2^k}\mathbb{E}(Z_1^2 \mathbb{1}_{Z_1 \leq 2^k}) = \mathbb{E}\Big(Z_1^2 \sum_{k \in A_{\max(Z_1,1)}} \frac{1}{2^k}\Big) \leq 2 + 2\mathbb{E}(Z_1) < \infty.$$

□

Remarque 1.4.2. Quelques variantes célèbres de la LGN.

1. Il existe de nombreuses preuve de la LGN, notamment via une marche aléatoire [29].
2. Indépendance. Etemadi a démontré dans les années 1980 que la LGN subsiste pour des v.a.r. intégrables centrées et indépendantes deux à deux. Pour en savoir plus : [15].
3. La LGN revient à la convergence p.s. de la mesure empirique vers la loi commune, ou à la convergence p.s. ponctuelle de la fonction de répartition empirique, et cette convergence est uniforme (théorème de Glivenko–Cantelli). Pour aller plus loin : TD.
4. Auto-normalisation. Sous les hypothèses du TLC, si $\sigma_n^2 := \frac{(X_1 - m_n)^2 + \cdots + (X_n - m_n)^2}{n-1}$ avec $m_n := \frac{X_1 + \cdots + X_n}{n}$, alors $\mathbb{E}(\sigma_n^2) = \sigma^2$, la LGN donne $\sigma_n^2 \to m$ p.s. et par le TLC et le lemme de Slutsky,

   $$\sqrt{n}\frac{m_n - m}{\sigma_n} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

   En statistique, on parle de méthode plug-in : on remplace σ par un estimateur.
5. Dépendance. La LGN existe pour les martingales et pour les fonctionnelles additives des processus de Markov ergodiques La LGN peut être vue comme une instance du théorème ergodique. Pour en savoir plus : cours du M1 S1 *Processus stochastique* et *Systèmes dynamiques*.
6. Si $(X_n)_{n\geq 1}$ est une suite de v.a. à valeurs dans un espace d'états quelconque E (graphe, groupe symétrique, variété, etc) dans lequel la notion d'addition et d'espérance ne font pas sens, on peut toujours utiliser une fonction test $f : E \to \mathbb{R}$ et considérer la LGN pour les v.a.r. $(f(X_n))_{n\geq 0}$. C'est le point de vue adopté pour les fonctions additives des processus de Markov.
7. Il existe des analogues de la LGN et du TLC dans des espaces de dimension infinie comme les Hilbert et les Banach. Pour en savoir plus : [54].

8. Loi du logarithme itéré de Strassen. Sous les hypothèses du TLC, presque sûrement,

$$\text{ValeursAdhérence}_{n\to\infty}\Big(\frac{X_1-m+\cdots+X_n-m}{\sqrt{2n\log(\log(n))}}\Big)=[-\sigma,\sigma].$$

En pratique $n\mapsto\log(\log(n))$ est très plate, à méditer ! Pour en savoir plus : [48].

9. LGN de Marcinkiewicz–Zygmung. Soient $(X_n)_{n\geq 1}$ v.a.r. i.i.d., $\mathbb{E}(|X_1|^\alpha)^{1/\alpha}<\infty$ avec $1<\alpha<2$. Alors

$$\frac{X_1+\cdots+X_n}{n^{1/\alpha}}\xrightarrow[n\to\infty]{\text{p.s.}}0.$$

Recentrage superflu : la moyenne est écrasée par la normalisation. Pour en savoir plus : [48].

10. LGN de Marcinkiewicz–Zygmung multiple avec sa réciproque de Kolmogorov, de Bai–Silverstein. Soient $(X_{ij})_{i,j\geq 1}$ une matrice infinie de v.a.r. i.i.d. et $0<\alpha<2$, $\beta\geq 0$, $M>0$ des constantes. Alors

$$\max_{1\leq j\leq Mn^\beta}\Big|n^{-1/\alpha}\sum_{i=1}^n(X_{ij}-c)\Big|\xrightarrow[n\to\infty]{\text{p.s.}}0\quad\text{ssi}\quad\mathbb{E}(|X_{1,1}|^{\alpha(1+\beta)})<\infty\ \text{ et }\ c=\begin{cases}\mathbb{E}(X_{1,1}) & \text{si } \alpha\geq 1\\ \text{réel quelconque} & \text{si } \alpha<1\end{cases}.$$

De plus, si $\mathbb{E}(|X_{1,1}|^{\alpha(1+\beta)})=\infty$ alors $\overline{\lim}_{n\to\infty}\max_{1\leq j\leq Mn^\beta}|n^{-1/\alpha}\sum_{i=1}^n(X_{ij}-c)|=\infty$ p.s.
L'uniformité exige de l'intégrabilité accrue. Cette LGN est utile pour l'analyse spectrale asymptotique des matrices aléatoires de grande dimension. Pour en savoir plus : [5, Lem. B25] et [14].

1.5 Algorithme de Monte–Carlo

Considérons le problème de l'évaluation numérique approchée de l'intégrale

$$\mu(f):=\int f(x)\mu(\mathrm{d}x)$$

où μ est une loi de probabilité sur $\mathbb{R}^d$ simulable et où $f:\mathbb{R}^d\to\mathbb{R}$ est une fonction μ-intégrable évaluable. Si $X\sim\mu$ alors $\mu(f)=\mathbb{E}(f(X))$ et si les $(X_n)_{n\geq 1}$ sont des copies i.i.d. de X alors par la LGN

$$\frac{f(X_1)+\cdots+f(X_n)}{n}\xrightarrow[n\to\infty]{\text{p.s.}}\mu(f).$$

De plus, si $f(X)$ est de carré intégrable, alors le TLC donne $\frac{f(X_1)+\cdots+f(X_n)}{n}\approx\mu(f)+\frac{\sigma_\mu(f)}{\sqrt{n}}Z$, $Z\sim\mathcal{N}(0,1)$, car

$$\frac{\sqrt{n}}{\sigma_\mu(f)}\Big(\frac{f(X_1)+\cdots+f(X_n)}{n}-\mu(f)\Big)\xrightarrow[n\to\infty]{\text{loi}}\mathcal{N}(0,1)\quad\text{où}\quad\sigma_\mu^2(f):=\int(f(x)-\mu(f))^2\mu(\mathrm{d}x)\leq\int f(x)^2\mu(\mathrm{d}x).$$

Ceci fournit un intervalle de confiance asymptotique : pour tout intervalle $[-q,q]\subset\mathbb{R}$, on obtient

$$\lim_{n\to\infty}\mathbb{P}(\mu(f)\notin[a,b])=\mathbb{P}(Z\in[-q,q])\quad\text{où}\quad[a,b]:=\frac{f(X_1)+\cdots+f(X_n)}{n}+\frac{\sigma_\mu(f)}{\sqrt{n}}[-q,q]\text{ et }Z\sim\mathcal{N}(0,1),$$

utile lorsque q est assez grand pour que $\mathbb{P}(Z\in[-q,q])$ soit petit. La largeur de l'intervalle de confiance $[a,b]$
— diminue quand n augmente, le coût de l'algorithme est typiquement linéaire en n.
— diminue quand $\sigma_\mu(f)$ diminue, voir réduction de variance ci-dessous.
— augmente quand $\mathbb{P}(Z\in[-q,q])$ diminue, et on souhaite que $\mathbb{P}(Z\in[-q,q])$ soit petit.

Remarque 1.5.1. Améliorations et variantes.

— Dépendance en la dimension : la dépendance en d ne se fait que via $\sigma_\mu(f)$.
— Contrôle de l'erreur : la majoration $\sigma_\mu^2(f)\leq\int f^2\mathrm{d}\mu$ indique que $\int f^2\mathrm{d}\mu$ contrôle l'erreur sur l'approximation de $\int f\mathrm{d}\mu$. Une majoration a priori sur f permet de rendre ce contrôle utile.
— Information a priori : la majoration de la variance $\sigma_\mu(f)$ nécessite des informations sur f et μ.
— Réduction de variance : réécrire $f(x)\mu(\mathrm{d}x))=\widetilde{f}(x)\widetilde{\mu}(\mathrm{d}x)$ de sorte que $\mu(f)=\widetilde{\mu}(\widetilde{f})$ et $\sigma(\widetilde{f})\leq\sigma_\mu(f)$.
— MCMC. Il n'est pas toujours facile de simuler μ, c'est-à-dire de générer une suite i.i.d. $(X_n)_{n\geq 0}$

de loi μ. Lorsque μ est une mesure de Boltzmann–Gibbs de densité $\frac{1}{Z}e^{-V}$, l'algorithme de Metropolis–Hastings fournit à partir de V une chaîne de Markov $(M_n)_{n\geq 0}$ qui converge en loi vers μ, et pour laquelle on peut également utiliser une variante de la LGN et du TLC. Il s'agit d'une méthode MCMC = Monte–Carlo Markov Chains. Il en existe de nombreuses variantes : échantillonneur de Gibbs, algorithme de Propp–Wilson, systèmes de particules en interaction, Metropolis-Adjusted Langevin (MALA), Hamiltonian Monte–Carlo (HMC), etc.

Pour en savoir plus : [71] pour la statistique, et [13] via des probabilités discrètes élémentaires.

1.6 Géométrie convexe en grande dimension

Je 06/02

Abordé en cours :
— Loi uniforme sur les convexes comme mesure de Boltzmann–Gibbs log-concave
— Phénomène couche mince pour les vecteurs à coordonnées i.i.d.
Cas du cube et de la boule euclidienne et rôle des points extrémaux
— Gaussiennes et loi uniforme sur la sphère, TLC pour la sphère
— Géométrie de la sphère en grande dimension
Concentration autour des équateurs et orthogonalité asymptotique des vecteurs indépendants
— Principe d'Archimède, TLC pour la boule.

1.6.1 Corps convexes et lois log-concaves

Un corps convexe $K \subset \mathbb{R}^n$ est une partie convexe compacte non-vide, par exemple une boule $\mathbb{B}_p^n$ ou un ellipsoïde $A\mathbb{B}^n$. La loi uniforme sur K, de densité $\frac{1}{|K|}\mathbb{1}_K$, est log-concave : une mesure de Boltzmann–Gibbs

$$\frac{1}{Z}e^{-V} \quad \text{avec } V : \mathbb{R}^n \to (-\infty, +\infty] \text{ convexe,}$$

puisque $V = 0$ sur K et $V = +\infty$ en dehors de K. Étudier les lois log-concaves sur $\mathbb{R}^n$ comme généralisation ou relaxation du concept de loi uniforme sur les corps convexes s'est avéré être un point de vue fructueux, entre probabilités et analyse fonctionnelle. Soit une loi log-concave et X un vecteur la suivant, on dit que X est
— centré lorsque $\mathbb{E}(X) = 0$,
— isotrope normalisé lorsque $\mathrm{Cov}(X) = I_n$
cette notion est relative à la base canonique, à ne pas confondre avec l'invariance par rotation !
— produit (tensoriel) lorsque les composantes de X sont indépendantes.

Si X est produit, alors $\mathrm{Cov}(X)$ est diagonale, et la réciproque est vraie dans le cas gaussien.
Si X a les symétries du cube [8], alors X est centré et $\mathrm{Cov}(X)$ est un multiple de l'identité.

Les composantes d'un vecteur aléatoire log-concave isotrope ne sont pas indépendantes en général. Leur dépendance est géométrique, issue d'une généralisation de l'appartenance à un corps convexe.

Voici quelques exemples élémentaires de lois log-concaves :

$\mathbf{V(x)}$	**Nom ou description**
$\frac{1}{2}\langle K^{-1}(x-m), x-m\rangle$, $K \succ 0$	Gaussienne $\mathcal{N}(m, K)$
$\sum_{i=1}^n U(x_i)$, U convexe	Produit générique
$\sum_{i=1}^n U(x_i) + \sum_{i<j} W(x_i - x_j)$, U, W convexes	Non-produit avec interaction à deux corps
$\lvert x\rvert^2 = \lvert x\rvert_2^2 = \sum_{i=1}^n x_i^2$	Gaussienne $\mathcal{N}(0, \frac{1}{2}I_n)$
$\lvert x\rvert_1 = \sum_{i=1}^n \lvert x_i\rvert$	Double exponentielle produit
$\lvert x\rvert_p^p = \sum_{i=1}^n \lvert x_i\rvert^p$, $p \geq 1$	Schatten
$\sum_{i=1}^n \log(1 + \lvert x_i\rvert^\alpha)$	Produit à queue lourdes
$\langle Ax, x\rangle + \lvert x\rvert_1$, $A \succcurlyeq 0$	Mixte gaussienne-exponentielle

— La loi uniforme sur le cube $\mathbb{B}_\infty^n$ est produit, mais elle n'est pas invariante par rotation.
Cependant elle hérite des symétries du cube, donc centrée, de matrice de covariance multiple de I_n.
— La loi uniforme sur la boule euclidienne $\mathbb{B}_2^n$ n'est pas produit, mais elle est invariante par rotation.
Cette symétrie sphérique continue inclut les symétries du cube.
Nous allons voir qu'en grande dimension, elle ressemble à une gaussienne isotrope, qui est produit.
— La loi uniforme sur $\mathbb{B}_p^n$, $p \in [1, \infty]$, a les symétries du cube, mais n'est produit que pour $p = 1$.
Nous allons voir qu'en grande dimension, elle ressemble à une loi produit.

8. Invariance par changement de signe sur coordonnée quelconque, et par permutation de deux coordonnées quelconques.

La log-concavité est une propriété stable par produit tensoriel, projection, et transformation affine.

Entre probabilités, analyse fonctionnelle, et géométrie convexe, une question importante consiste à déterminer à quel point les lois log-concaves se comportent en grande dimension comme des lois log-concave produit, notamment des lois gaussiennes ou exponentielle double, à quel point le TLC est vrai, etc. Voici quelques phénomènes célèbres pour les lois de probabilité log-concaves, en rapport avec l'histoire et l'actualité :

— le théorème de Dvoretzky sur les sections ellipsoïdales [9]
`https://en.wikipedia.org/wiki/Dvoretzky's_theorem`
— le théorème de Paouris sur l'existence de directions sous-exponentielles,
— le TLC et le phénomène couche mince,
— la conjecture de la variance,
— la conjecture de Kannan-Lovász-Simonovits (KLS) sur l'inégalité de Poincaré,
— la conjecture de l'hyperplan de Bourgain (Bourgain slicing problem)

Ce domaine des mathématiques, dont les héros sont Grothendieck, Dvoretzky, Milman, Maurey, Pisier, Bourgain, Talagrand, etc, porte plusieurs noms, pas tout à fait synonymes : analyse fonctionnelle probabiliste, analyse géométrique asymptotique, théorie locale des espaces de Banach, probabilités de grande dimension, etc.

Dans ce cours, on se contente d'étudier certaines lois log-concaves spéciales pour faire émerger de manière très accessible des phénomènes de grande dimension emblématiques. Pour en savoir plus : [8, 47, 46, 18, 3, 51].

1.6.2 Phénomène couche mince

Théorème 1.6.1. Phénomène couche mince ou couronne mince (thin shell).

Si $X_1, \dots, X_n$ sont des v.a.r. i.i.d. vérifiant $m_1 = \mathbb{E}(X_1) = 0$, $m_2 = \mathbb{E}(X_1^2) = 1$, et $m_4 = \mathbb{E}(X_1^4) < \infty$, alors

$$\sqrt{n}\Big|\frac{|X|}{\sqrt{n}} - 1\Big| \xrightarrow[n\to\infty]{\text{loi}} |\mathcal{N}(0, \tfrac{m_4-1}{4})|.$$

En particulier, pour tout réel α tel que $0 < \alpha < 1/2$,

$$n^\alpha\Big|\frac{|X|}{\sqrt{n}} - 1\Big| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Dessin !

— Ainsi, en grande dimension, la loi de X se concentre dans une mince couronne autour du rayon $\sqrt{n}$: c'est le phénomène couche mince. Cela reste-t-il valable pour les loi log-concaves centrées isotropes normalisées ? La réponse est positive, et peut s'obtenir via le phénomène TLC pour les lois log-concaves.
— En particulier, le phénomène couche mince suggère que la loi uniforme sur un corps convexe en grande dimension ressemble en quelque sorte au cytoplasme d'une cellule végétale déshydratée, toile étirée, tendue, épinglée en les points extrémaux, soulignant la subtilité des phénomènes de grande dimension !
— Examinons le cas où X est uniforme sur le cube $\mathbb{B}_\infty^n(\sqrt{3}) = (\sqrt{3}[-1,1])^n$, corps convexe produit, composantes de X indépendantes uniformes sur $\sqrt{3}[-1,1]$. On a $\mathbb{E}(X) = 0$ et $K := \mathrm{Cov}(X) = \mathbb{E}(XX^\top) = I_n$, et X est isotrope normalisé. Les points extrémaux de $\mathbb{B}_\infty^n(\sqrt{3})$ sont $\{x \in \mathbb{B}_\infty^n(\sqrt{3}) : \exists i : x_i = \pm\sqrt{3}\}$, ce qui inclut notamment les points les plus extrémaux $\{\pm\sqrt{3}\}^n$, qui sont de norme $\sqrt{3n}$. Le théorème 1.6.1 dit qu'en grande dimension, X est concentré autour de la sphère euclidienne $\mathbb{S}^{n-1}(\sqrt{n})$, avec une fluctuation à l'échelle $\sqrt{n}$. Le rayon de cette sphère critique est à une distance bornée des points les plus extrémaux.
— Examinons le cas où X est uniforme sur $\mathbb{B}^n = \mathbb{B}_2^n(1)$, corps convexe non-produit, composantes de X non-indépendantes. Le théorème 1.6.1 ne s'applique pas, mais nous pouvons tenter d'explorer le phénomène. Comme $\mathbb{B}^n$ a les symétries du cube, $\mathbb{E}(X) = 0$ et $K := \mathrm{Cov}(X) = \mathbb{E}(XX^\top) = \sigma^2 I_n$, avec

$$\sigma^2 = \frac{\mathrm{Tr}(K)}{n} = \frac{\mathbb{E}(|X|^2)}{n}.$$

Or $|X|$ a pour densité $r \in [0,1] \mapsto nr^{n-1}$, de second moment $\frac{n}{n+2}$, d'où $\mathbb{E}(|X|^2) = \frac{n}{n+2} \xrightarrow[n\to\infty]{} 1$, et $\sqrt{n}X$ est asymptotiquement (c'est-à-dire en grande dimension) isotrope normalisé. De plus,

$$\mathbb{E}(|X|) = \frac{n}{n+1} \quad \text{et} \quad \mathrm{Var}(|X|) = \frac{n}{n+2} - \Big(\frac{n}{n+1}\Big)^2 = \frac{n}{(n+2)(n+1)^2} = O\Big(\frac{1}{n^2}\Big),$$

donc en grande dimension, $\sqrt{n}X$, de loi uniforme sur $\mathbb{B}^n(\sqrt{n})$, est approximativement isotrope normalisé, concentré autour de la sphère euclidienne $\mathbb{S}^{n-1}(\sqrt{n})$, et on a bien un phénomène couche mince

9. Et son approche probabiliste de Vitali Milman par concentration de la mesure !

au-delà du cas i.i.d.! De plus, le phénomène TLC a bien lieu également pour $\sqrt{n}X$, cf. corollaire 1.6.9 plus loin. Notons au passage qu'en grande dimension, la loi uniforme sur le corps convexe isotrope normalisé $\mathbb{B}^n(\sqrt{n})$ s'écrase sur l'ensemble de ses points extrémaux $\mathbb{S}^{n-1}(\sqrt{n})$.

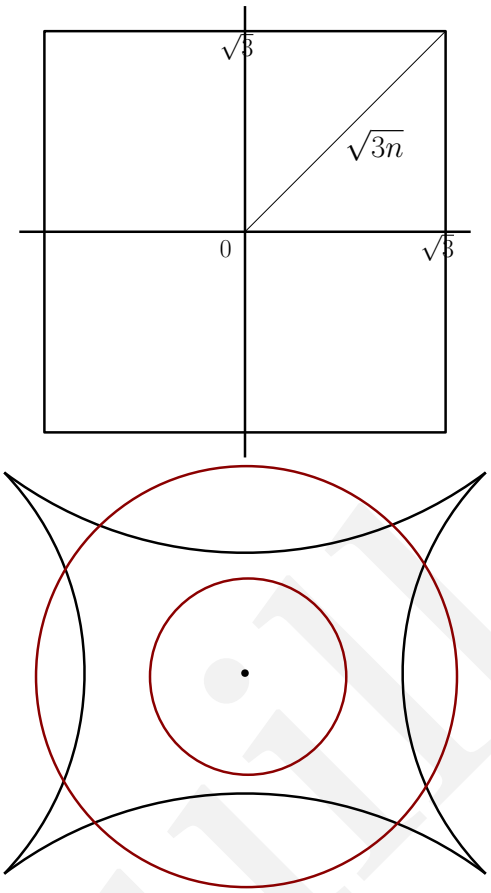


FIGURE 1.2 – Vue d'artiste schématique du phénomène couche mince (thin shell).

Démonstration. La LGN donne

$$\frac{|X|^2}{n} = \frac{X_1^2+\cdots+X_n^2}{n} \xrightarrow[n\to\infty]{\text{p.s.}} m_2 = 1 \quad \text{ou de manière équivalente} \quad \Big|\frac{|X|}{\sqrt{n}} - 1\Big| \xrightarrow[n\to\infty]{\text{p.s.}} 0.$$

Le TLC pour les $(X_n^2)_{n\geq 1}$ donne $\sqrt{n}\big(\frac{|X|^2}{n}-1\big) \to \mathcal{N}(0, m_4-1)$ en loi, et il en découle par la méthode delta pour la fonction $\sqrt{\cdot}$ et le point 1 que $\sqrt{n}\big|\frac{|X|}{\sqrt{n}}-1\big| \to |\mathcal{N}(0, \frac{m_4-1}{4})|$ en loi. Par conséquent, pour tout α tel que $0<\alpha<1/2$, par le lemme de Slutsky, $\frac{n^\alpha}{\sqrt{n}}\sqrt{n}\big|\frac{|X|}{\sqrt{n}}-1\big| \to \delta_0$ en loi, et donc en probabilité car la limite est constante. □

Remarque 1.6.2. Cube entre deux boules euclidiennes en grande dimension.

Pour le cube $\mathbb{B}_\infty^n = [-1,1]^n$, le ratio de volume avec la plus petite boule euclidienne qui le contient vérifie

$$\frac{|\mathbb{B}^n(\sqrt{n})|}{|\mathbb{B}_\infty^n|} = \frac{|\mathbb{B}^n(\sqrt{n})|}{2^n} = \frac{\pi^{\frac{n}{2}} n^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)2^n} \underset{n\to\infty}{\sim} \frac{\pi^{\frac{n}{2}} n^{\frac{n}{2}}}{\sqrt{2\pi\frac{n}{2}}(\frac{n}{2\mathrm{e}})^{\frac{n}{2}}2^n} = \frac{1}{\sqrt{\pi n}}\Big(\frac{\pi\mathrm{e}}{2}\Big)^{\frac{n}{2}} \xrightarrow[n\to\infty]{} \infty.$$

Par ailleurs, concernant la plus grande boule euclidienne qu'il contient, cela donne

$$\frac{|\mathbb{B}_\infty^n|}{|\mathbb{B}^n|} = \frac{2^n}{|\mathbb{B}^n|} = \frac{\Gamma(\frac{n}{2}+1)2^n}{\pi^{\frac{n}{2}}} \underset{n\to\infty}{\sim} \frac{\sqrt{2\pi\frac{n}{2}}(\frac{n}{2\mathrm{e}})^{\frac{n}{2}}2^n}{\pi^{\frac{n}{2}}} = \sqrt{\pi n}\Big(\frac{2n}{\pi\mathrm{e}}\Big)^{\frac{n}{2}} \xrightarrow[n\to\infty]{} \infty.$$

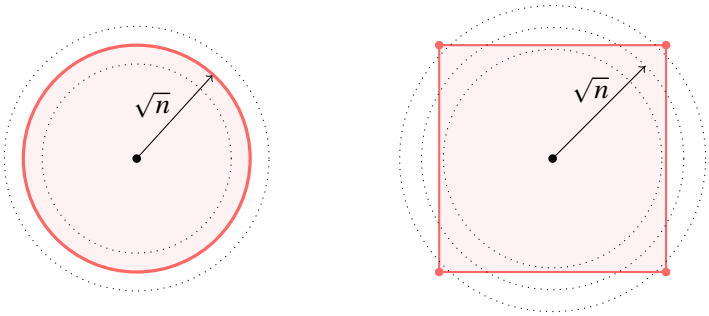


FIGURE 1.3 – Illustration du phénomène couche mince (thin shell). Dans le cas de la boule euclidienne $\mathbb{B}_2^n(\sqrt{n})$, asymptotiquement isotrope, à gauche, la mesure uniforme s'écrase en grande dimension n sur la sphère $\mathbb{S}^{n-1}(\sqrt{n})$ qui en constitue l'ensemble des points extrémaux. Plus la dimension n augmente, plus la proportion du bord de la boule par rapport au centre augmente. Dans le cas du cube $\mathbb{B}_\infty^n(\sqrt{3n})$, isotrope, à droite, la mesure uniforme se concentre en grande dimension n autour de la sphère $\mathbb{S}^{n-1}(\sqrt{n})$, qui est à distance bornée de l'ensemble des points extrémaux $\sqrt{3n}\{\pm 1\}^n$. Plus la dimension n augmente, plus l'aspect « pointu » des points extrémaux du cube s'exacerbe et s'éloigne de l'intuition liée aux basses dimensions visuelles 1, 2, et 3.

Remarque 1.6.3. Volume de la boule unité et de la boule normalisée.

$$|\mathbb{B}^n| = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} \underset{n\to\infty}{\sim} \frac{\pi^{\frac{n}{2}}}{\sqrt{2\pi\frac{n}{2}}(\frac{n}{2\mathrm{e}})^{\frac{n}{2}}} \xrightarrow[n\to\infty]{} 0 \quad \text{et} \quad |\mathbb{B}^n(\sqrt{n})| \underset{n\to\infty}{\sim} \frac{\pi^{\frac{n}{2}} n^{\frac{n}{2}}}{\sqrt{2\pi\frac{n}{2}}(\frac{n}{2\mathrm{e}})^{\frac{n}{2}}} = \frac{(2\pi\mathrm{e})^{\frac{n}{2}}}{\sqrt{\pi n}} \xrightarrow[n\to\infty]{} \infty.$$

1.6.3 Géométrie de la sphère et gaussiennes

Dans des cas symétriques rigides, la loi uniforme a une représentation à base d'ingrédients indépendants.

Théorème 1.6.4. Gaussiennes et loi uniforme sur les sphères.

i) Les composantes d'un vecteur aléatoire Z de $\mathbb{R}^n$ sont i.i.d. de loi $\mathcal{N}(0,1)$ ssi

$$\frac{Z}{|Z|} \quad \text{et} \quad |Z|$$

sont indépendantes de loi Uniforme$(\mathbb{S}^{n-1})$ et $\chi(n)$ respectivement.

ii) Si $X = (X_1, \dots, X_n)$ suit la loi uniforme sur $\mathbb{S}^{n-1}(\sqrt{n})$ alors pour tout $k \geq 1$ fixé,

$$\mathrm{proj}_{\mathbb{R}^k}(X_1, \dots, X_n) = (X_1, \dots, X_k) \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0, I_k).$$

iii) Si X suit la loi uniforme sur $\mathbb{S}^{n-1}(\sqrt{n})$ alors pour tout $\theta \in \mathbb{S}^{n-1}$,

$$\langle X, \theta \rangle \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

Dessin !

— Le (i) est une représentation de la loi uniforme sur $\mathbb{S}^{n-1}$ avec des ingrédients i.i.d.
— Dans (i) l'uniformité de $Z/|Z|$ et son indépendance de $|Z|$ vient de l'invariance par rotation de Z.
— Dans (i) la division de Z par $|Z|$ permet de coller à la géométrie de la sphère, mais rend les composantes de $Z/|Z|$ dépendantes. Cependant, comme les composantes de Z sont indépendantes, la somme normalisée $|Z|^2/n = (Z_1^2 + \cdots + Z_n^2)/n$ est déterministe en grande dimension par la LGN, c'est aussi le cas de $|Z|/\sqrt{n}$, d'où l'indépendance en grande dimension (asymptotique $n \to \infty$) des composantes de $Z/|Z|$.
— L'invariance par rotation de la loi de Z n'est compatible avec l'indépendance des composantes de Z que dans le cas gaussien isotrope d'après la caractérisation géométrique de Maxwell (théorème 1.2.7).
— Le (i) est à comparer avec le (et à distinguer du) théorème de Cochran.
— Le (ii), connu sous le nom d'observation/lemme/théorème de Poincaré/Borel, est encore plus ancien!
— Le (iii) est un TLC pour des variables aléatoires dépendantes dont la dépendance est géométrique : appartenance à la sphère. La sphère $\mathbb{S}^{n-1}(\sqrt{n})$ n'est pas un corps convexe. En revanche le (ii) affirme

qu'en grande dimension, la loi uniforme sur $\mathbb{S}^{n-1}(\sqrt{n})$ ressemble à la gaussienne, qui est log-concave. Cela suggère d'étudier le TLC pour le corps convexe $\mathbb{B}^n$, ce qui qui fait plus loin dans le chapitre.
— Bien qu'isotrope, la sphère $\mathbb{S}^{n-1}(\sqrt{n})$ n'est pas un corps convexe, mais au vu du phénomène couche mince, elle en est la caricature en grande dimension en quelque sorte, concentrée sur le rayon $\sqrt{n}$.
— Toutes les projections de dimension k se valent en raison de l'invariance par rotation de la loi. En particulier, par (iii), pour $\theta = e_1$, $X_1 \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1)$, et pour $\theta = \left(\frac{1}{\sqrt{n}},\dots,\frac{1}{\sqrt{n}}\right)$, $\frac{X_1+\cdots+X_n}{\sqrt{n}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1)$.
— $\mathbb{S}^{n-1}(\sqrt{n})$ est la plus petite sphère contenant le cube discret : $\{\pm 1\}^n \subset \mathbb{S}^{n-1}(\sqrt{n})$. Cela conduit à approcher certains modèles discrets par des modèles continus gaussiens. Exemple : verres de spins.
— Un théorème démontré par Gérard Letac en 1981 affirme que si X est un vecteur aléatoire de $\mathbb{R}^n$, $n \geq 3$, tel que $\mathbb{P}(X=0)=0$, X a des composantes indépendantes, et $X/|X|$ suit la loi uniforme sur $\mathbb{S}^{n-1}$, alors X est gaussien isotrope. Ce théorème est plus fort que la caractérisation de Maxwell (théorème 1.2.7).

Démonstration.

i) Découle d'un passage en coordonnées sphériques [10] : $\mathrm{e}^{-\frac{|x|^2}{2}}\mathrm{d}x = r^{n-1}\mathrm{e}^{-\frac{r^2}{2}}\mathrm{d}r\mathrm{d}\theta$.

ii) Nous avons $X \overset{\text{loi}}{=} \sqrt{n}Z/|Z|$ avec $Z \sim \mathcal{N}(0,I_n)$ par (i), or $(Z_1,\dots,Z_k) \sim \mathcal{N}(0,I_k)$ tandis que la LGN donne $|Z|/\sqrt{n} = \sqrt{(Z_1^2+\cdots+Z_n^2)/n} \xrightarrow[n\to\infty]{} 1$ en probabilité. Donc $\sqrt{n}(Z_1,\dots,Z_k)/|Z| \xrightarrow[n\to\infty]{} \mathcal{N}(0,I_k)$ en loi par le lemme de Slutsky. Alternativement, il est possible de coupler toutes les lois uniformes sur les sphères en considérant une unique suite infinie $Z_1,Z_2,\dots$ de v.a.r. i.i.d. de loi $\mathcal{N}(0,1)$, et d'utiliser la LGN forte pour obtenir $|Z|/\sqrt{n} = \sqrt{(Z_1^2+\cdots+Z_n^2)/n} \to 1$ p.s. d'où $\sqrt{n}(Z_1,\dots,Z_k)/|Z| \xrightarrow[n\to\infty]{} (Z_1,\dots,Z_k)$ p.s. donc en loi.

iii) Par invariance par rotation, on peut prendre $\theta = e_1$, et le résultat découle alors du (ii) avec $k=1$.

□

Corollaire 1.6.5. Concentration autour des équateurs et orthogonalité en grande dimension.

i) Soit $\theta \in \mathbb{S}^{n-1}$, $H_\theta := (\mathbb{R}\theta)^\perp$, et $E_\theta := H_\theta \cap \mathbb{S}^{n-1}$ l'équateur (grand cercle quand $n=3$) orthogonal à θ.
Soit X un vecteur aléatoire de loi uniforme sur $\mathbb{S}^{n-1}$.
Alors en grande dimension, X est concentré dans un voisinage de l'équateur E_θ : pour tout $r \geq 0$,

$$\mathbb{P}(\sqrt{n}\mathrm{dist}_{\mathbb{R}^n}(X,H_\theta) \geq r) \xrightarrow[n\to\infty]{} \mathbb{P}(|Z| \geq r) \leq \mathrm{e}^{-\frac{r^2}{2}}, \quad Z \sim \mathcal{N}(0,1).$$

ii) Soient X et Y deux vecteurs aléatoires indépendants de loi uniforme sur $\mathbb{S}^{n-1}$.
Alors en grande dimension, ils sont approximativement orthogonaux : pour tout $r \geq 0$,

$$\mathbb{P}(\sqrt{n}|\langle X,Y\rangle| \geq r) \xrightarrow[n\to\infty]{} \mathbb{P}(|Z| \geq r) \leq \mathrm{e}^{-\frac{r^2}{2}}, \quad Z \sim \mathcal{N}(0,1).$$

Dessin !

— Tous les équateurs se valent car la loi de X est invariante par rotation (cela peut dérouter l'intuition).
— Le phénomène de concentration gaussien est exploré dans le chapitre suivant.
— Le ii) se généralise à un nombre arbitraire fixé N de vecteurs i.i.d. de loi uniforme sur $\mathbb{S}^{n-1}$.
Le cas où N dépend de n donne lieu à un phénomène de seuil étudié dans [20].
— L'absence de préfacteur 2 devant l'expoentielle peut surprendre. Il s'avère que la queue de distribution gaussienne est encore plus petite ! Des développements récents se trouvent dans [4].

Démonstration. La loi $\mathcal{N}(0,1)$ est concentrée autour de l'origine : pour tout $r \geq 0$, $\mathbb{P}(|Z| \geq r) = \mathrm{erfc}(\frac{r}{\sqrt{2}}) \leq \mathrm{e}^{-\frac{r^2}{2}}$.

i) Découle du fait que $\mathrm{dist}_{\mathbb{R}^n}(X,H_\theta) = |\langle X,\theta\rangle|$ et du théorème précédent (théorème 1.6.4). Intuition géométrique : en grande dimension, X est loin de θ, c'est-à-dire presque orthogonal, donc proche de E_θ (dessin !).

ii) En combinant le théorème de Fubini–Tonelli, l'indépendance [11] de X et Y, l'invariance par rotation de la loi de X, et le théorème précédent (théorème 1.6.4), il vient

$$\begin{aligned}\mathbb{P}(\sqrt{n}|\langle X,Y\rangle| \geq r) &= \mathbb{E}(\mathbb{1}_{|\langle\sqrt{n}X,Y\rangle|\geq r}) = \int\left(\int \mathbb{1}_{|\langle\sqrt{n}x,y\rangle|\geq r}\mathbb{P}_X(\mathrm{d}x)\right)\mathbb{P}_Y(\mathrm{d}y)\\ &= \int \mathbb{P}(|\langle\sqrt{n}X,y\rangle| \geq r)\mathbb{P}_Y(\mathrm{d}y) = \mathbb{P}(|\langle\sqrt{n}X,e_1\rangle| \geq r) \xrightarrow[n\to\infty]{} \mathbb{P}(|Z| \geq r) \leq \mathrm{e}^{-\frac{r^2}{2}}.\end{aligned}$$

10. Au passage, cela fournit une astuce géométrico-probabiliste pour retrouver sans effort la formule de la densité de la loi $\chi(n)$.
11. Avec un peu plus de culture en probabilités, cela revient à conditionner par $Y=\theta$ et à exploiter l'indépendance de X et Y. Cet exemple illustre un fait général : en cas d'indépendance, conditionner est superflu et l'usage du théorème Fubini–Tonelli suffit.

Alternativement, on peut écrire $X = \frac{Z}{|Z|}$ et $Y = \frac{Z'}{|Z'|}$ avec Z et Z' indépendants de loi $\mathcal{N}(0, I_n)$, et utiliser le TLC pour variables i.i.d., la loi des grands nombres, et le lemme de Slutsky, pour obtenir

$$\sqrt{n}\langle X, Y\rangle = \sqrt{n}\frac{\sum_{i=1}^n Z_i Z_i'}{(\sqrt{n}+o(1))^2} = \frac{\sum_{i=1}^n Z_i Z_i'}{\sqrt{n}}(1+o(1)) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1).$$

□

Remarque 1.6.6. Isopérimétrie de Lévy–Gromov.

Un équateur partage la sphère $\mathbb{S}^{n-1}$ en deux hémisphères, et les hémisphères sont des calottes sphériques particulières. Plus généralement, à une calotte sphérique est associée une calotte sphérique complémentaire, et l'une des deux est de mesure $\geq 1/2$ pour la loi uniforme sur $\mathbb{S}^{n-1}$. La plus grande peut être vue comme un voisinage d'une hémisphère, et en particulier d'un équateur, et il découle du corollaire 1.6.5 que sa mesure est très rapidement proche de 1. Vers 1919, Paul Lévy a découvert un raffinement géométrique de ce phénomène : les calottes sphériques ont une mesure de bord minimale parmi les parties de la sphère de même mesure, et ce sont les seules, c'est-à-dire qu'elles constituent les ensembles extrémaux pour le problème isopérimétrique sphérique. Le caractère adimensionnel de l'isopérimétrie sphérique est gaussien : en projetant comme dans le théorème 1.6.4, l'isopérimétrie sphérique donne l'isopérimétrie gaussienne, pour laquelle les ensembles extrémaux sont les demi-espaces, projetés des calottes sphériques en grande dimension. D'autre part, l'extrémalité isopérimétrique des calottes sphériques fait sens intuitivement au vu du fait que les boules euclidiennes sont les ensembles extrémaux pour le problème isopérimétrique sur $\mathbb{R}^n$ (équipé de la mesure de Lebesgue). Vitali Milman a utilisé vers 1971 le phénomène de concentration de la mesure gaussien lié à l'isopérimétrie sphérique pour fournir une preuve probabiliste du théorème de Dvoretzky sur l'existence de sections presque ellipsoïdales des corps convexes en grande dimension, un résultat emblématique du domaine. Cette approche novatrice a séduit Mikhaïl Gromov, qui a notamment démontré ensuite vers 1980 que le phénomène découvert par Paul Lévy restait valable pour les variétés riemanniennes compactes à courbure de Ricci minorée par une constante positive, ce qui est une comparaison à la sphère. Vers 1996, ce théorème de Gromov a été revisité et étendu grâce aux semi-groupes de Markov par Dominique Bakry et Michel Ledoux, en s'appuyant sur une traduction fonctionnelle de type Sobolev obtenue par Sergey Bobkov, et explorée à nouveau vers 2000 notamment par Franck Barthe et Bernard Maurey.

1.6.4 Principe d'Archimède et TLC pour la boule euclidienne

Géométrie et probabilités élémentaires :

Lemme 1.6.7. Loi des projections sphériques.

Si $X = (X_1, \dots, X_n)$ suit la loi uniforme sur $\mathbb{S}^{n-1}$ alors pour tout $1 \leq k \leq n$,

$$|(X_1, \dots, X_k)|^2 = X_1^2 + \cdots + X_k^2 \sim \mathrm{Beta}\Big(\frac{k}{2}, \frac{n-k}{2}\Big).$$

— Toutes les projections de dimension k se valent car la loi est invariante par rotation.

— Lorsque $n \geq 2$ et $k = 1$, par invariance par rotation, $|\langle X, \theta\rangle|^2 \overset{\text{loi}}{=} |X_1|^2 \sim \mathrm{Beta}(\frac{1}{2}, \frac{n-1}{2})$, pour tout $\theta \in \mathbb{S}^{k-1}$. Donc la projection $\langle X, \theta\rangle = \theta_1 X_1 + \cdots + \theta_n X_n$ sur un diamètre quelconque $\theta \in \mathbb{S}^{n-1}$ a pour densité

$$t \in \mathbb{R} \mapsto \frac{\Gamma(\frac{n}{2})}{\sqrt{\pi}\Gamma(\frac{n-1}{2})}(1-t^2)^{\frac{n-3}{2}}\mathbb{1}_{t\in[-1,1]}.$$

C'est la formule de Funk–Hecke. Il s'agit d'une loi Beta sur $[-1,1]$. En particulier :

— si $n = 2$, c'est la loi de l'arsinus de densité $\frac{1}{\pi\sqrt{1-t^2}}\mathbb{1}_{t\in[-1,1]}$

— si $n = 3$, c'est la loi uniforme de densité $\frac{1}{2}\mathbb{1}_{t\in[-1,1]}$ (cas particulier du principe d'Archimède)

— si $n = 4$, c'est la loi du demi-cercle de densité $\frac{2\sqrt{1-t^2}}{\pi}\mathbb{1}_{t\in[-1,1]}$

— si $n \to \infty$, comme $(1-\frac{t^2}{n})^n \to \mathrm{e}^{-\frac{t^2}{2}}$ on trouve $\sqrt{n}\langle X, \theta\rangle \underset{n\to\infty}{\overset{\text{loi}}{\longrightarrow}} \mathcal{N}(0,1)$, ce qui rejoint le théorème 1.6.4, et il en découle en particulier que $\langle X, \theta\rangle$ converge en loi vers δ_0 quand $n \to \infty$ (donc en probabilité).

Démonstration. Par le théorème 1.6.4, $X \overset{\text{loi}}{=} \frac{Z}{|Z|}$ avec $Z \sim \mathcal{N}(0, I_n)$, donc $Y := (X_1, \dots, X_k)$ vérifie

$$|Y|^2 \overset{\text{loi}}{=} \frac{Z_1^2 + \cdots + Z_k^2}{(Z_1^2 + \cdots + Z_k^2) + (Z_{k+1}^2 + \cdots + Z_n^2)}.$$

Or il est connu que si A et B sont indépendantes avec $A \sim \chi^2(a)$ et $B \sim \chi^2(b)$ alors $\frac{A}{A+B} \sim \text{Beta}(\frac{a}{2}, \frac{b}{2})$. □

De la sphère (euclidienne) à la boule (euclidienne) par projection :

Théorème 1.6.8. Principe d'Archimède et représentation de la loi uniforme sur la boule euclidienne.

i) Principe d'Archimède :
Si $n \geq 3$ et si $X = (X_1, \dots, X_n)$ est un vecteur aléatoire de $\mathbb{R}^n$ de loi uniforme sur la sphère $\mathbb{S}^{n-1}$ alors sa projection $(X_1, \dots, X_{n-2})$ sur $\mathbb{R}^{n-2}$ suit la loi uniforme sur la boule unité $\mathbb{B}^{n-2}$.

ii) Principe d'Archimède renversé ou représentation de la loi uniforme sur la boule :
Si $n \geq 1$ et $Z_1, \dots, Z_n$ sont i.i.d. de loi $\mathcal{N}(0,1)$ indépendantes de $E \sim \text{Exp}(1)$, alors

$$\frac{(Z_1, \dots, Z_n)}{\sqrt{Z_1^2 + \cdots + Z_n^2 + 2E}} \quad \text{suit la loi uniforme sur la boule } \mathbb{B}^n.$$

— Il s'agit bien de l'Archimède [12] de la Grèce antique, celui donc de la poussée du même nom. Mais on prendra garde à ne pas confondre ce qu'on appelle ici principe d'Archimède, théorème sur la sphère et le cylindre, avec le principe du même Archimède lié à la poussée d'Archimède et au fameux Eurêka.
— Toutes les projections de dimension k se valent car la loi est invariante par rotation.
— Pour le cas $n = 3$, considéré historiquement par Archimède, la projection de la loi uniforme sur la sphère de $\mathbb{R}^3$ sur un diamètre suit la loi uniforme sur le diamètre. Cela coïncide avec la formule de Funk–Hecke.
— On ne peut pas remplacer $n-2$ par $n-1$ dans le principe d'Archimède, contre-exemple : cas $n = 2$ de la formule de Funk–Hecke, pour lequel la projection suit la loi de l'arcsinus et non pas la loi uniforme.

Démonstration.

i) Quand $n = 3$, il s'agit d'établir que la projection sur un diamètre est uniforme, cela coïncide avec la formule de Funk–Hecke, et découle également du résultat antique d'Archimède (cylindre et sphère) : la surface d'une calotte sphérique est égale à l'élévation entre son bord et son pôle (fois 2π). Le lien entre géométrie et probabilité se fait ici, et comme souvent, via la proportionnalité entre surfaces ou volumes.
Pour le cas général, on pose $Y := (X_1, \dots, X_{n-2})$, qui prend ses valeurs dans $\mathbb{B}^{n-2}$. La loi de Y hérite de celle de X l'invariance par rotation. Son uniformité s'obtient en examinant la loi de $|Y|$. Le lemme 1.6.7 avec $k = n-2$ donne $|Y|^2 \sim \text{Beta}(\frac{n-2}{2}, 1)$, qui a pour densité $r \in [0,1] \mapsto \frac{n-2}{2} r^{\frac{n-4}{2}}$, et donc $|Y|$ a pour densité $r \in [0,1] \mapsto \frac{n-2}{2} 2r(r^2)^{\frac{n-4}{2}} = (n-2)r^{n-3}$, qui est bien la loi du rayon de la loi uniforme sur $\mathbb{B}^{n-2}$.

ii) Découle du (i) et du théorème 1.6.4 via $Z_{n+1} + Z_{n+2} \sim (\chi^2(1))^{*2} = \text{Gamma}(\frac{2}{2}, \frac{1}{2}) = \text{Exp}(\frac{1}{2})$.

□

Dessin!

Corollaire 1.6.9. TLC pour la loi uniforme sur la boule euclidienne.

Si X suit la loi uniforme sur la boule $\mathbb{B}^n(\sqrt{n})$ alors pour toute direction $\theta \in \mathbb{S}^{n-1}$ on a

$$\langle X, \theta \rangle \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

— C'est l'analogue pour les boules du TLC pour les sphères du théorème 1.6.4.
C'est un TLC pour un corps convexe non-produit.
Cela complète les TLC déjà démontrés : sphères (non-convexe), et cube (produit).

12. Archimède de Syracuse (-287 – -212) a démontré que si on place une sphère dans un cylindre ajusté alors l'aire du cylindre est égale à l'aire de la sphère elle-même. Archimède était si fier de ce résultat qu'il a fait graver la figure sur sa pierre tombale. C'est cette gravure qui a permis à son admirateur Cicéron (-106 – -43) d'identifier sa tombe et de la restaurer, en -75, près d'un siècle et demi après son meurtre lors du siège de Syracuse par un soldat romain ignorant. En fait Archimède est même plus précis : si on coupe le tout en deux par un plan perpendiculaire au cylindre, les parties supérieure et inférieure vérifient encore la propriété. Pour en savoir plus : [55, 64, 10].
Voir aussi `https://en.wikipedia.org/wiki/On_the_Sphere_and_Cylinder`

— Toutes les directions se valent car la loi est invariante par rotation. Ce TLC pour le corps convexe $\mathbb{B}^n$ est donc valable quelque soit la direction θ. Comme déjà mentionné, certaines directions sont impossibles pour certains corps convexes comme le cube par exemple.

— En particulier, pour $\theta = e_1$, on obtient $X_1 \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1)$ et pour $\theta = (\frac{1}{\sqrt{n}},\dots,\frac{1}{\sqrt{n}})$, $\frac{X_1+\cdots+X_n}{\sqrt{n}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1)$.

Démonstration. Par le principe d'Archimède renversé (théorème 1.6.8), $X \overset{\text{loi}}{=} \sqrt{n}(Z_1^2+\cdots+Z_n^2+2E)^{-1/2}(Z_1,\dots,Z_n)$ où les $Z_1,\dots,Z_n$ sont i.i.d. $\mathcal{N}(0,1)$ indépendantes de $E \sim \text{Exp}(1)$. Donc pour tout $\theta \in \mathbb{S}^{n-1}$,

$$\langle X,\theta\rangle \overset{\text{loi}}{=} \sqrt{n}\frac{\langle (Z_1,\dots,Z_n),\theta\rangle}{\sqrt{Z_1^2+\cdots+Z_n^2+2E}}.$$

Or $\langle (Z_1,\dots,Z_n),\theta\rangle \sim \mathcal{N}(0,|\theta|^2=1)$, $\sqrt{Z_1^2+\cdots+Z_n^2+2E} = \sqrt{n}(1+o(1))$ p.s. par LGN, et par le lemme de Slutsky,

$$\langle X,\theta\rangle \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1).$$

□

$$\text{Unif}(\mathbb{B}^n(\sqrt{n})) \approx \text{Unif}(\mathbb{S}^{n-1}(\sqrt{n})) \approx \mathcal{N}(0,I_n) \approx \text{Unif}(\{-1,1\}^n)$$

FIGURE 1.4 – Équivalences en grande dimension n.

Remarque 1.6.10. Loi uniforme sur $\mathbb{S}_p^{n-1}$ et $\mathbb{B}_p^n$.

Si la loi uniforme sur $\mathbb{S}_p^{n-1}$ et $\mathbb{B}_p^n$ pour $p=2$ est représentable avec des gaussiennes, quid du cas $p \neq 2$?

i) Les composantes d'un vecteur aléatoire X de $\mathbb{R}_+^n$ sont i.i.d. de loi $\text{Exp}(1) = \text{Gamma}(1,1)$ ssi

$$\frac{X}{|X|_1} \quad \text{et} \quad |X|_1 = X_1+\cdots+X_n$$

sont indépendantes, de loi $\text{Uniforme}(\mathbb{S}_1^{n-1}\cap\mathbb{R}_+^n)$ et $\text{Gamma}(n,1) = (\text{Exp}(1))^{*n}$. Plus généralement, pour tout p, les composantes d'un vecteur aléatoire X de $\mathbb{R}_+^n$ sont i.i.d. de loi L_p de densité

$$x \mapsto \frac{p}{\Gamma(p^{-1})}\mathrm{e}^{-x^p}\mathbb{1}_{x\geq 0} \quad \text{ssi} \quad \frac{X}{|X|_p} \quad \text{et} \quad |X|_p^p = X_1^p+\cdots+X_n^p$$

sont indépendantes de lois $\text{Uniforme}(\mathbb{S}_p^{n-1}\cap\mathbb{R}_+^n)$ et L_p^{*n}. Pour en savoir plus : [73].

ii) $X_i^p \sim \text{Exp}(1)$. La loi uniforme sur $\mathbb{S}_p^{n-1}$ et $\mathbb{B}_p^n$ s'obtiennent respectivement avec des signes aléatoires (symétrisation), puis en rajoutant une variable aléatoire indépendante au dénominateur comme dans le principe d'Archimède renversé. Cette représentation à base d'ingrédients indépendants permet d'établir le TLC et couche mince pour $\mathbb{B}_p^n$ pour tout p. Pour en savoir plus : [9].

Remarque 1.6.11. Algorithme de simulation.

La représentation de la loi uniforme sur $\mathbb{S}_p^{n-1}$ et $\mathbb{B}_p^n$ à base d'ingrédients indépendants permet l'analyse mathématique, mais aussi la simulation. Pour $p=2$ via des v.a.r. i.i.d. gaussiennes, pour $p=1$ via des v.a.r. i.i.d. exponentielles, etc. Dans le cas spécial $p=1$, la loi uniforme sur le simplexe $\mathbb{S}_1^{n-1}\cap\mathbb{R}_+^n$ n'est rien d'autre que $\text{Dirichlet}(1,\dots,1)$, simulable via le réordonnement de v.a.r. i.i.d. uniformes sur $[0,1]$ (statistique d'ordre), et il s'agit là d'une représentation alternative à base d'ingrédient indépendants. C'est l'occasion de souligner qu'au bout du compte, la simulation, même en grande dimension, se réduit à celle d'une suite de 0 et de 1 de loi de Bernoulli symétrique, trinité universelle : $0, 1, \infty$.

Remarque 1.6.12. Loi uniforme sur la sphère et projection stéréographique.

Soit π la projection stéréographique de la sphère $\mathbb{S}^{n-1}$ de $\mathbb{R}^n$, $n \geq 2$, sur l'hyperplan de $\mathbb{R}^n$ d'équation $x_n = 0$ orthogonal à e_n, identifié à $\mathbb{R}^{n-1}$. Pour tout $x \in \mathbb{S}^{n-1}$, on a [a]

$$\pi(x) = \Big(\frac{x_1}{1-x_n}, \dots, \frac{x_{n-1}}{1-x_n}\Big).$$

Le pôle nord e_n est bien singulier et joue le rôle de point à l'infini, tandis que les points de l'équateur $\mathbb{S}^{n-1} \cap \mathbb{R}^{n-1} = \{x \in \mathbb{S}^{n-1} : x_n = 0\}$ sont fixes. Comme $|\pi(x)|^2 = \frac{1-x_n^2}{(1-x_n)^2} = \frac{1+x_n}{1-x_n}$, il vient, pour tout $y \in \mathbb{R}^{n-1}$,

$$\pi^{-1}(y) = \Big(\frac{2y_1}{1+|y|^2}, \dots, \frac{2y_1}{1+|y|^2}, \frac{|y|^2-1}{|y|^2+1}\Big).$$

La projection stéréographique π est un difféomorphisme entre $\mathbb{S}^{n-1} \setminus \{e_n\}$ et $\mathbb{R}^{n-1}$, qui est de plus conforme (préserve les angles). Soit X de loi uniforme sur $\mathbb{S}^{n-1}$. Un calcul de jacobien (changement de variable) montre que $\pi(X)$ suit la loi à queue lourde radiale (loi de Cauchy quand $n = 2$) de densité

$$x \in \mathbb{R}^{n-1} \mapsto \frac{1}{Z(1+|x|^2)^{n-1}} \quad \text{où} \quad Z = \frac{\pi^{\frac{n}{2}}}{2^{n-2}\Gamma(\frac{n}{2})}.$$

La loi de $\pi(X)$ est invariante par rotation. Peut-on exploiter $\lim_{n\to\infty}(1+\frac{|x|^2}{2n})^{-(n-1)} = \mathrm{e}^{-\frac{|x|^2}{2}}$, ce qui revient à comparer la loi de $\pi(X)$ à la loi gaussienne $\mathcal{N}(0, \frac{1}{2n} I_{n-1})$ dans $\mathbb{R}^{n-1}$? En fait $|\pi(X)|$ a pour densité

$$r \mapsto \frac{r^{n-2}}{Z_n(1+r^2)^{n-1}}, \quad \text{où} \quad Z_n := \frac{\Gamma(\frac{n}{2}-1)\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})^2}$$

(reliée à la loi de Student). Par conséquent

$$\mathbb{E}(|\pi(X)|) = \int_0^\infty \frac{r^{n-1}}{Z_n(1+r^2)^{n-1}}\mathrm{d}r = \frac{\Gamma\left(\frac{n}{2}-1\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)^2} \xrightarrow[n\to\infty]{} 1$$

et

$$\mathbb{E}(|\pi(X)|^2) = \int_0^\infty \frac{r^n}{Z_n(1+r^2)^{n-1}}\mathrm{d}r = \frac{2^{n-2}\Gamma\left(\frac{n-3}{2}\right)\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)\Gamma(n-1)} \xrightarrow[n\to\infty]{} 1$$

en particulier $\mathrm{Var}(|\pi(X)|) \xrightarrow[n\to\infty]{} 0$, ce qui évoque le phénomène couche mince.

a. Pour en mesurer l'impact symbolique : c'est l'unique formule figurant dans le mémoire d'habilitation de quelques pages de Bernhard Riemann (1826 – 1866), achevé à Göttingen en 1854, et intitulé *À propos des hypothèses sous-jacentes à la géométrie.*

Dessin

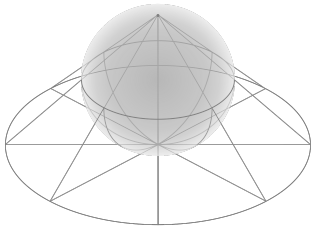


FIGURE 1.5 – Une projection stéréographique, symbole du portail Wikipédia de la géométrie.

FIGURE 1.6 – À gauche, un portrait d'Archimède sur l'avers de la médaille Fields, son nom en grec, une date, et une citation latine du poète Marcus Manilius : *« Transire suum pectus mundoque potiri (s'élever au-dessus de soi-même et conquérir le monde) »*. À droite, au revers de la médaille, une phrase en latin : *« Congregati ex toto orbe mathematici ob scripta insignia tribuere (les mathématiciens rassemblés du monde entier ont récompensé pour des contributions exceptionnelles) »* et à l'arrière-plan, la tombe d'Archimède, avec la gravure de son théorème « De la sphère et du cylindre » disposée derrière un rameau, ce qui correspond au cas $n = 3$ du théorème 1.6.8. La tranche de la médaille, non reproduite ici, porte le nom du lauréat.

1.6.5 Théorème de Carathéodory et méthode empirique de Maurey

Rappelons que l'enveloppe convexe d'une partie A d'un espace vectoriel est définie par

$$\mathrm{co}(A) = \bigcup_{n=1}^{\infty} \mathrm{co}_n(A) \quad \text{où} \quad \mathrm{co}_n(A) = \Big\{ \sum_{i=1}^n \lambda_i x_i : x_i \in A, \lambda_i \in [0,1], \sum_{i=1}^n \lambda_i = 1 \Big\}.$$

C'est donc l'ensemble des moyennes des lois de probabilité discrètes, à nombre fini d'atomes, portées par A.

Théorème 1.6.13. Représentation probabiliste de Carathéodory de l'enveloppe convexe.

Pour tout $A \subset \mathbb{R}^d$, nous avons $\mathrm{co}(A) = \bigcup_{n=1}^{d+1} \mathrm{co}_n(A)$.

Démonstration. Soient $x = \sum_{i=1}^n \lambda_i x_i \in \mathrm{co}(A)$, $n \geq 1$, $x_i \in A$, $\lambda_i \in [0,1]$, $\sum_{i=1}^n \lambda_i = 1$. À présent, si $n > d+1$, alors les $n-1 > d$ vecteurs $x_1 - x_n, \ldots, x_{n-1} - x_n$ de $\mathbb{R}^d$ sont linéairements dépendants, donc $\sum_{i=1}^n \mu_i x_i = 0$ pour des réels $\mu_1, \ldots, \mu_{n-1}$ non tous nuls, avec $\mu_n := -\sum_{i=1}^{n-1} \mu_i$. Par conséquent, nous avons $x = \sum_{i=1}^n (\lambda_i + \theta \mu_i) x_i$ pour tout réel θ. Nous pouvons maintenant sélectionner θ pour exprimer x comme une combinaison convexe de $n-1$ points. Le résultat désiré s'obtient enfin en répétant l'argument $n-(d+1)$ fois.

Cette preuve est constructive. Le cas du simplexe $A = \{0, e_1, \ldots, e_d\} \subset \mathbb{R}^d$ montre que $d+1$ est optimal. □

Théorème 1.6.14. Variante empirique et adimensionnelle.

Si E est un espace de Hilbert séparable et $A \subset E$ borné, alors il existe une constante $c = c(A)$ t.q. pour tout $x \in \mathrm{co}(A)$ et tout $n \geq 1$, il existe $x_1, \ldots, x_n \in A$ t.q. $|x - \frac{1}{n}\sum_{i=1}^n x_i| \leq \frac{c}{\sqrt{n}}$.

Ce théorème est utile notamment pour étudier les discrétisations des corps convexes.

Lorsque H est de dimension infinie, la définition de l'espérance des v.a. à valeur dans H et sa compabilité avec l'indépendance se fait avec une base hilbertienne ou avec le théorème de représentation de Riesz.

Démonstration. Nous adoptons la méthode empirique de (Bernard) Maurey ou méthode probabiliste de (Paul) Erdős. Elle devient constructive par Monte Carlo. Comme $x \in \mathrm{co}(A)$, il existe une mesure de probabilité μ portée par A et de moyenne x. Soient $(X_i)_{i \geq 1}$ des v.a. i.i.d. de loi μ. Alors $\mathbb{E}(X_i) = x$, et pour tout $n \geq 1$,

$$\mathbb{E}\Big(\Big|x - \frac{1}{n}\sum_{i=1}^n X_i\Big|^2\Big) = \frac{1}{n^2} \sum_{i,j=1}^n \mathbb{E}(\langle X_i - x, X_j - x\rangle) = \frac{1}{n}\mathbb{E}(|X_1 - x|^2) = \frac{c^2}{n}.$$

Par conséquent, il existe au moins ω tel que $|x - \frac{1}{n}\sum_{i=1}^n X_i(\omega)| \leq \frac{c}{\sqrt{n}}$. Il suffit alors de poser $x_i = X_i(\omega)$. □

Chapitre 2

Phénomène de concentration de la mesure

Me 12/02

Abordé en cours :
— Inégalité de Hoeffding en admettant transformée de Laplace sur $[a,b]$
— Inégalité de Hoeffding auto-normalisée
— ISL gaussienne en admettant la régularisation
— ISL et Laplace des Lipschitz sous-gaussienne (méthode de Herbst) en admettant la régularisation
— ISL et concentration de la moyenne empirique sous ISL

2.1 Inégalité de Hoeffding et concentration de la mesure

Le phénomène de concentration de la mesure en gros et en bref : si $(X_1,\dots,X_n)$ est un vecteur aléatoire à composantes suffisamment peu dépendantes et si $f_n(X_1,\dots,X_n)$ en est une fonction suffisamment intégrable et suffisamment peu dépendante en chaque composante, alors $f_n(X_1,\dots,X_n)$ se rapproche de son espérance $\mathbb{E}(f_n(X_1,\dots,X_n))$ en grande dimension n. La mesure qui se concentre en grande dimension est la loi de $f_n(X_1,\dots,X_n)$. La moyenne empirique $f(X_1,\dots,X_n)=\frac{X_1+\cdots+X_n}{n}$ est l'exemple le plus emblématique. Plus généralement, pour capturer le phénomène de concentration, une approche consiste à utiliser l'inégalité de Markov pour contrôler une déviation d'une fonction croissante de la fonctionnelle aléatoire d'intérêt, ce qui revient à contrôler un moment : d'ordre 2 pour l'inégalité de Tchebychev, exponentiel pour l'inégalité de Chernoff.

Une instance simple et emblématique du phénomène de concentration de la mesure est la suivante :

Théorème 2.1.1. Inégalité de Hoeffding pour une somme de variables bornées.

Si $S_n := X_1+\cdots+X_n$ où $X_1,\dots,X_n$ sont indépendantes, X_i à valeurs dans $[a_i,b_i]$, alors pour tout $t\geq 0$,

$$\mathbb{P}(S_n-\mathbb{E}(S_n)\geq t)\leq \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\Big).$$

— Contient la « gaussienne binaire » $\frac{1}{2}(\delta_{-1}+\delta_1)$, loi de Rademacher, Bernoulli symétrique sur $\{-1,1\}$.
— Oscillation : $\operatorname{osc}(X_i)=\max(X_i)-\min(X_i)\leq b_i-a_i$, diamètre de l'ensemble $X(\Omega)$, $\leq 2\|X_i\|_\infty$ et plus fin.
— La preuve est aussi importante que le théorème.
— On passe de l'inégalité de déviation à la concentration en appliquant la déviation aux $-X_i$:

$$\mathbb{P}(|S_n-\mathbb{E}(S_n)|\geq t)\leq \mathbb{P}(S_n-\mathbb{E}(S_n)\geq t)+\mathbb{P}(-S_n-\mathbb{E}(-S_n)\geq t)\leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\Big).$$

Démonstration. Inégalité sur la transformée de Laplace. Si X est à valeurs dans $[a,b]$ alors pour tout $\theta\geq 0$,

$$\mathbb{E}(\mathrm{e}^{\theta X})\leq \mathrm{e}^{\theta^2\frac{(b-a)^2}{8}+\theta\mathbb{E}(X)}.$$

En effet, par translation on se ramène au cas $\mathbb{E}(X)=0$ et on alors forcément $a\leq 0\leq b$. Ensuite, pour tout $x\in[a,b]$, on a $\mathrm{e}^{\theta x}\leq \frac{b-x}{b-a}\mathrm{e}^{\theta a}+\frac{x-a}{b-a}\mathrm{e}^{\theta b}$ par convexité de $u\mapsto \mathrm{e}^{\theta u}$ pour la combinaison convexe $x=\frac{b-x}{b-a}a+\frac{x-a}{b-a}b$. En posant $p=-a/(b-a)$ et $f(u)=\log((1-p)\mathrm{e}^{-pu}+p\mathrm{e}^{(1-p)u})=-pu+\log(1-p+p\mathrm{e}^u)$, il vient, grâce à $\mathbb{E}(X)=0$,

$$\mathbb{E}(\mathrm{e}^{\theta X})\leq \frac{b}{b-a}\mathrm{e}^{\theta a}-\frac{a}{b-a}\mathrm{e}^{\theta b}=\mathrm{e}^{f(\theta(b-a))}.$$

À présent, on a $f'(u)=-p+p\mathrm{e}^u/(1-p+p\mathrm{e}^u)$ et[1] $f''(u)=p(1-p)\mathrm{e}^u/(1-p+p\mathrm{e}^u)^2\leq\frac{1}{4}$, et comme $f(0)=f'(0)=0$, on obtient $f(u)\leq u^2/8$, d'où l'inégalité annoncée sur la transformée de Laplace de X à valeurs dans $[a,b]$.

1. Il suffit d'utiliser l'inégalité $\frac{AB}{(A+B)^2}\leq\frac{1}{4}$ avec $A=1-p$ et $B=p\mathrm{e}^u$.

Inégalité de déviation. Pour tout $t \geq 0$, en introduisant le paramètre libre $\theta > 0$, en utilisant l'inégalité de Markov, l'indépendance, et la majoration de la transformée de Laplace précédente, on obtient [2] :

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq t) = \mathbb{P}(\theta(S_n - \mathbb{E}(S_n)) \geq \theta t) \leq \mathrm{e}^{-\theta t}\mathbb{E}(\mathrm{e}^{\theta(S_n - \mathbb{E}(S_n))}) = \mathrm{e}^{-\theta t}\prod_{i=1}^{n}\mathbb{E}(\mathrm{e}^{\theta(X_i - \mathbb{E}(X_i))}) \leq \mathrm{e}^{-\theta t + \frac{\theta^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2}.$$

Le membre de droite est minimal pour le choix (optimal) $\theta = 4t/\sum_{i=1}^{n}(b_i - a_i)^2$, d'où le résultat. □

Quand les X_i sont i.i.d. et dans $[a,b]$, on a $(b_1 - a_1)^2 + \cdots + (b_n - a_n)^2 = n(b-a)^2$, d'où, pour $t = nr$, $r \geq 0$,

$$\mathbb{P}\Big(\Big|\frac{X_1 + \cdots + X_n}{n} - \mathbb{E}(X_1)\Big| \geq r\Big) \leq 2\exp\Big(-\frac{2nr^2}{(b-a)^2}\Big),$$

qui exprime une concentration quand $n \to \infty$, mais aussi, par dilatation ou avec $t = \sqrt{n}r$, $r \geq 0$,

$$\mathbb{P}\Big(\sqrt{n}\Big|\frac{X_1 + \cdots + X_n}{n} - \mathbb{E}(X_1)\Big| \geq r\Big) \leq 2\exp\Big(-\frac{2r^2}{(b-a)^2}\Big),$$

qui est une version non-asymptotique, on dit quantitative, du TLC. Le membre de droite ne dépend pas de n.

L'inégalité de Hoeffding possède de nombreuses variantes et extensions, comme l'inégalité de Bernstein, démontrée en TD, pour les sommes de variables aléatoires de variance finie (plutôt que bornées).

Plus généralement, on parle de concentration sous-gaussienne, sous-poissonienne, et sous-exponentielle quand la borne est en $\exp(-cr^2)$, $\exp(-cr\log r)$, $\exp(-cr)$, par comparaison à la queue de distribution de ces lois. Idem pour les majorations sur la transformée de Laplace. Dans le fil des travaux de Chernoff, Hoeffding, et Bernstein, mais aussi Kolmogorov et Prokhorov, les inégalités de concentration ont été développées pour les besoins de la statistique, de l'analyse des algorithmes, et de l'optimisation combinatoire randomisée : inégalité de Dvoretzky–Kiefer–Wolfovitz, de Azuma, de McDiarmid, d'Efron–Stein, etc.

Corollaire 2.1.2. Inégalité de Hoeffding pour somme de variables symétriques auto-normalisée.

Si $X_1, \ldots, X_n$ sont réelles indépendantes, symétriques, et sans atome en 0, alors pour tout réel $t \geq 0$,

$$\mathbb{P}(T_n \geq t) \leq \mathrm{e}^{-\frac{t^2}{2}} \quad \text{où} \quad T_n := \frac{X_1 + \cdots + X_n}{\sqrt{X_1^2 + \cdots + X_n^2}}.$$

— Ce résultat suivant est dû à Bradley Efron (1969), cf. [34].
— Comme les hypothèses sont vérifiées par les $-X_i$, on obtient $\mathbb{P}(|T_n| \geq r) \leq 2\mathrm{e}^{-\frac{t^2}{2}}$.
— On apprécie l'absence d'hypothèse de support ou de moment sur les X_i. Queues lourdes admises.
— La preuve est plus importante que le résultat : exploitation de l'aléa des signes pour concentrer.
— X est symétrique lorsque X et $-X$ ont même loi, et X a un atome en x lorsque $\mathbb{P}(X = x) > 0$.
— La symétrie assure que le signe de X_i est de loi de Rademacher $\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$, indépendant de la valeur absolue $|X_i|$, tandis que la normalisation $X_1^2 + \cdots + X_n^2$ ne dépend que des valeurs absolues. Si $X_i = \varepsilon_i R_i$ avec $R_i \geq 0$ et ε_i de loi $p\delta_1 + (1-p)\delta_{-1}$, indépendante de R_i, alors $\mathbb{E}(X_i) = \mathbb{E}(\varepsilon_i)\mathbb{E}(R_i) = (2p-1)\mathbb{E}(R_i)$.
— L'absence d'atome en 0 assure que $X_1^2 + \cdots + X_n^2 > 0$ p.s. et permet donc de diviser par cette quantité.
— Si $\mathbb{E}(X_1^2) < \infty$ alors par la LGN, $X_1^2 + \cdots + X_n^2 \sim_{n\to\infty} n$ p.s. et par lemme de Slutsky et le TLC, $T_n \to \mathcal{N}(0,1)$ en loi quand $n \to \infty$. Si les X_i ne sont pas de carré intégrable, l'analyse est différente, cf. chapitre 5.
— L'inégalité de Cauchy–Schwarz indique que T_n prend ses valeurs dans $[-\sqrt{n}, \sqrt{n}]$. Lorsque $X_1, \ldots, X_n$ sont i.i.d. de loi $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, alors $X_1^2 + \cdots + X_n^2 = n$, et on retrouve le S_n du théorème 2.1.1.
— Dans un esprit géométrique, $T_n = \langle U_n, \theta_n \rangle$ où $U_n := \sqrt{n}T_n$ et $\theta_n := (1, \ldots, 1)/\sqrt{n} \in \mathbb{S}^{n-1}$. Par conséquent, si les X_i sont de loi $\mathcal{N}(0,1)$, alors U_n suit la loi uniforme sur la sphère $\mathbb{S}^{n-1}(\sqrt{n})$, et on retrouve un comportement sous-gaussien compatible avec le TLC pour la loi uniforme sur la sphère (théorème 1.6.4). Le vecteur aléatoire U_n a les symétries du cube $[-1,1]^n$, qui viennent des signes $\varepsilon_1, \ldots, \varepsilon_n$.
— Dans un esprit statistique, pour tout $r \geq 0$,

$$\mathbb{P}\Big(\Big|\sqrt{n}\frac{\overline{X}_n}{\widehat{\sigma}_n}\Big| \geq r\Big) \leq 2\exp\Big(-\frac{nr^2}{2(n-1+r^2)}\Big) \quad \text{où} \quad \overline{X}_n := \frac{X_1 + \cdots + X_n}{n} \quad \text{et} \quad \widehat{\sigma}_n^2 := \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n-1},$$

Car $(n-1)\widehat{\sigma}_n^2 = X_1^2 + \cdots + X_n^2 - \frac{(X_1 + \cdots + X_n)^2}{n}$, et pour tout $r \geq 0$, $\Big\{\sqrt{n}\frac{\overline{X}_n}{\widehat{\sigma}_n} \geq r\Big\} = \Big\{T_n \geq r\sqrt{\frac{n}{n-1+r^2}}\Big\}$. Lorsque les X_i sont $\mathcal{N}(0,1)$, alors la studentisation qui découle du théorème de Cochran indique que $\sqrt{n}\overline{X}_n/\widehat{\sigma}_n$ suit une loi de Student $t(n-1)$ qui est à queue lourde, de densité $u \mapsto (1 + u^2/(n-1))^{n/2}$.

2. On parle parfois d'inégalité de déviation à droite par transformée de Laplace, ou de borne de Chernoff.

Démonstration. Comme pour tout i les variables X_i et $-X_i$ ont même loi sans atome en 0, la variable $\varepsilon_i := \mathrm{sign}(X_i)$ vérifie $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$ (loi de Rademacher, Bernoulli symétrique sur $\{-1,1\}$), et $(|X_1|,\dots,|X_n|)$ et $(\varepsilon_1,\dots,\varepsilon_n)$ sont indépendants. Par le théorème de Fubini–Tonelli [3], en notant $T_n := \psi(|X_1|,\dots,|X_n|,\varepsilon_1,\dots,\varepsilon_n)$,

$$\begin{aligned}\mathbb{P}(T_n \geq t) = \mathbb{E}(\mathbb{1}_{T_n\geq t}) &= \iint \mathbb{1}_{\psi\geq t}\mathrm{d}\mathbb{P}_{|X_1|,\dots,|X_n|}\mathrm{d}\mathbb{P}_{\varepsilon_1,\dots,\varepsilon_n} \\ &= \int\Big(\int \mathbb{1}_{\psi\geq t}\mathrm{d}\mathbb{P}_{\varepsilon_1,\dots,\varepsilon_n}\Big)\mathrm{d}\mathbb{P}_{|X_1|,\dots,|X_n|} = \int \varphi(|X_1|,\dots,|X_n|)\mathrm{d}\mathbb{P}_{|X_1|,\dots,|X_n|}\end{aligned}$$

où

$$\varphi(r_1,\dots,r_n) := \mathbb{E}(\mathbb{1}_{\psi(r_1,\dots,r_n,\varepsilon_1,\dots,\varepsilon_n)\geq t}) = \mathbb{P}(\psi(r_1,\dots,r_n,\varepsilon_1,\dots,\varepsilon_n)\geq t) = \mathbb{P}\left(\frac{\varepsilon_1 r_1+\cdots+\varepsilon_n r_n}{\sqrt{r_1^2+\cdots+r_n^2}}\geq r\right).$$

Or $\varphi(r_1,\dots,r_n) = \mathbb{P}(Z_1+\cdots+Z_n\geq t)$ où $Z_i := \varepsilon_i r_i/\sqrt{r_1^2+\cdots+r_n^2}$, et $\sum_{i=1}^n \mathrm{osc}(Z_i)^2 = \sum_{i=1}^n 4r_i^2/(r_1^2+\cdots+r_n^2) \leq 4$, $\mathbb{E}(Z_i)=0$. Par l'inégalité de Hoeffding du théorème 2.1.1, $\mathbb{P}(Z_1+\cdots+Z_n\geq t)\leq \exp(-t^2/2)$, uniforme en r_i ! □

Sur le versant géométrique, le phénomène de concentration de la mesure a été exploré à l'origine notamment par Vitali Milman [4], en rapport avec la concentration gaussienne autour des équateurs de la mesure uniforme sur la sphère (mise en évidence par le corollaire 1.6.5 du théorème 1.6.4). Cela s'exprime également pour les fonctions Lipschitz sous la mesure gaussienne comme suit : si $F:\mathbb{R}^n\to\mathbb{R}$ est une fonction Lipschitz avec $\|F\|_{\mathrm{Lip}}\leq 1$, par exemple $F(x_1,\dots,x_n)=\frac{x_1+\cdots+x_n}{\sqrt{n}}$, alors pour tout $r\geq 0$,

$$\gamma^n\Big(\Big|F-\int F\mathrm{d}\gamma^n\Big|\geq r\Big)\leq 2\mathrm{e}^{-\frac{r^2}{2}}.$$

Cette estimée quantitative est indépendante de la dimension n, on dit adimensionnelle. Elle s'étend à des gaussiennes de dimension infinie [5], et est emblématique de l'analyse gaussienne. Par ailleurs, et comme nous allons le voir, la concentration de la mesure gaussienne pour les fonctions Lipschitz découle de l'inégalité de Sobolev logarithmique, une inégalité fonctionnelle qui s'avère être adimensionnelle, vraie pour les gaussiennes, et reliée au problème du transport optimal (et aux processus de Markov). Suite à l'impulsion de Vitali Milman, la concentration de la mesure géométrique et fonctionnelle a été popularisée et développée notamment par Mikhaïl Gromov, Michel Talagrand, Bernard Maurey, Gilles Pisier, Michel Ledoux, Sergey Bobkov, et bien d'autres.

2.2 Lemme de Johnson – Lindenstrauss sur la réduction de dimension

Lemme 2.2.1. de Johnson – Lindenstrauss pour la réduction de dimension.

Pour tous $n, N\geq 1$ et toute partie $S\subset\mathbb{R}^n$ de cardinal N, tout $0<\varepsilon<1$, et tout $k>\frac{24}{\varepsilon^2}\log N$,
il existe $A\in\mathscr{M}_{k,n}(\mathbb{R})$ telle que pour tous $x,y\in S$, $x\neq y$,

$$\sqrt{1-\varepsilon}\leq\frac{|Ax-Ay|}{|x-y|}\leq\sqrt{1+\varepsilon}.$$

En d'autres termes, il est possible de plonger quasi-isométriquement N points de $\mathbb{R}^n$ dans $\mathbb{R}^{O(\log N)}$.

Démonstration. La propriété s'écrit aussi $||A(x-y)|^2-|x-y|^2|\leq\varepsilon|x-y|^2$. Si $R_1,\dots,R_k$ sont les lignes de A, alors

$$|A(x-y)|^2=\sum_{i=1}^k\langle R_i,x-y\rangle^2.$$

Si A est une matrice aléatoire avec $R_1,\dots,R_k$ i.i.d. $\mathscr{N}(0,\frac{1}{k}I_n)$, alors $\langle R_i,x-y\rangle\sim\mathscr{N}(0,\frac{1}{k}|x-y|^2)$, tandis que $\langle R_i,x-y\rangle^2$ est χ^2, et cela suggère d'exploiter le phénomène de concentration de la mesure de la loi du χ^2 pour obtenir que $|A(x-y)|^2$ est proche de sa moyenne $|x-y|^2$, pour de petites déviations par rapport à la moyenne.

Soient $Z_1,\dots,Z_k$ des v.a.r. i.i.d. de loi $\mathscr{N}(0,\sigma^2)$. Pour tout $\theta<\frac{1}{2\sigma^2}$,

$$\mathbb{E}(\mathrm{e}^{\theta(Z_i^2-\sigma^2)})=\frac{\mathrm{e}^{-\theta\sigma^2}}{\sqrt{2\pi}\sigma}\int\mathrm{e}^{\theta u^2-\frac{u^2}{2\sigma^2}}\mathrm{d}u=\frac{\mathrm{e}^{-\theta\sigma^2}}{\sqrt{1-2\theta\sigma^2}}.$$

3. En termes probabilistes, cela revient à conditionner par les valeurs absolues, et exploiter leur indépendance avec les signes.
4. Inspiré par l'inégalité isopérimétrique de Paul Lévy, pour démontrer le théorème de Dvoretzky sur les corps convexes.
5. Comme par exemple la mesure de Wiener, gaussienne sur $\mathscr{C}(\mathbb{R}_+,\mathbb{R}^d)$, loi des trajectoires du mouvement brownien.

Par inégalité de Markov, pour tous $0<\varepsilon<1$ et $0<\theta<\frac{1}{2\sigma^2}$,

$$\mathbb{P}\Big(\frac{1}{k}\sum_{i=1}^{k}(Z_i^2-\sigma^2)\geq\varepsilon\sigma^2\Big)\leq \mathrm{e}^{-k(\theta\sigma^2(1+\varepsilon)+\log\sqrt{1-2\theta\sigma^2})}.$$

La valeur optimale de $\theta=\frac{\varepsilon}{2\sigma^2(1+\varepsilon)}<\frac{1}{2\sigma^2}$ et l'inégalité $\varepsilon-\log(1+\varepsilon)\geq\frac{\varepsilon^2}{4}$, $0<\varepsilon<1$, donnent

$$\mathbb{P}\Big(\frac{1}{k}\sum_{i=1}^{k}(Z_i^2-\sigma^2)\geq\varepsilon\sigma^2\Big)\leq \mathrm{e}^{-\frac{k}{8}\varepsilon^2}.$$

La même méthode pour $-(Z_i^2-\sigma^2)$ avec $\theta=\frac{\varepsilon}{2\sigma^2(1-\varepsilon)}$ et $-\varepsilon-\log(1-\varepsilon)\geq\frac{\varepsilon^2}{4}$, $0<\varepsilon<1$, donne

$$\mathbb{P}\Big(-\sum_{i=1}^{k}(Z_i^2-\sigma^2)\geq\varepsilon\sigma^2\Big)\leq \mathrm{e}^{-\frac{k}{2}(\theta\sigma^2(\varepsilon-1)+\log\sqrt{1+2\theta\sigma^2})}\leq \mathrm{e}^{-\frac{k}{8}\varepsilon^2}.$$

D'où, par la borne de l'union, pour tout $0<\varepsilon<1$, la majoration suivante qui ne dépend pas de σ^2 :

$$\mathbb{P}\Big(\Big|\frac{1}{k}\sum_{i=1}^{k}Z_i^2-\sigma^2\Big|\geq\varepsilon\sigma^2\Big)\leq 2\mathrm{e}^{-\frac{k}{8}\varepsilon^2}.$$

Maintenant, comme $R_1,\dots,R_k$ sont i.i.d. de loi $\mathcal{N}(0,\frac{1}{k}I_n)$, il vient que pour tout $x\in\mathbb{R}^n$, $|Ax|^2=\sum_{i=1}^{k}\langle R_i,x\rangle^2$ est de même loi que $\frac{1}{k}\sum_{i=1}^{k}Z_k^2$ avec $\sigma^2=|x|^2$, donc pour tout $0<\varepsilon<1$,

$$\mathbb{P}\Big(\big||A(x)|^2-|x|^2\big|\geq\varepsilon|x|^2\Big)\leq 2\mathrm{e}^{-\frac{k}{8}\varepsilon^2}.$$

La borne ne dépend pas de x. Par conséquent, en utilisant la borne de l'union, pour tout $0<\varepsilon<1$,

$$\mathbb{P}\Big(\exists x,y\in S:\big||A(x-y)|^2-|x-y|^2\big|>\varepsilon|x-y|^2\Big)\leq|S^2|\max_{x,y\in S}\mathbb{P}\Big(\big||A(x-y)|^2-|x-y|^2\big|>\varepsilon|x-y|^2\Big)\leq 2N^2\mathrm{e}^{-\frac{k}{8}\varepsilon^2}.$$

Si $k>\frac{8\log(2N^2)}{\varepsilon^2}$, alors la probabilité à gauche est <1, donc il existe ω pour lequel $||A(x-y)|^2-|x-y|^2|\leq\varepsilon|x-y|^2$ pour tous $x,y\in S$. Cette méthode probabiliste devient constructive avec Monte Carlo. La probabilité de trouver un tel ω diminue lorsque ε ou k diminuent, ou lorsque N augmente, ce qui fait sens. □

2.3 Inégalité de Sobolev logarithmique gaussienne

Pour toute mesure de probabilité μ sur E et $f:E\to\mathbb{R}_+$ telle que $f\in L^1(\mu)$, on pose, avec $0\log(0):=0$,

$$\mathrm{Ent}_\mu(f):=\int f\log(f)\mathrm{d}\mu-\int f\mathrm{d}\mu\log\int f\mathrm{d}\mu=\mathrm{H}(\nu\,|\,\mu)\int f\mathrm{d}\mu\quad\text{où}\quad\mathrm{d}\nu:=\frac{f}{\int f\mathrm{d}\mu}\mathrm{d}\mu.$$

L'inégalité de Jensen pour la fonction strictement convexe $u\in\mathbb{R}_+\mapsto u\log(u)$ assure que $\mathrm{Ent}_\mu(f)\in[0,+\infty]$ et $\mathrm{Ent}_\mu(f)=0$ ssi $f=0$ (presque sûrement pour μ). Selon le point de vue, la fonctionnelle Ent_μ est une…

— entropie relative ou divergence de Kullback–Leibler (théorie de l'information, statistique)
— énergie libre de Helmholtz (physique statistique).

Lemme 2.3.1. Formule variationnelle pour l'entropie relative.

Pour toute mesure de probabilité μ sur E et tout $f:E\to\mathbb{R}_+$, $f\in L^1(\mu)$, on a la linéarisation

$$\mathrm{Ent}_\mu(f)=\sup\Big\{\int fg\mathrm{d}\mu:\int\mathrm{e}^g\mathrm{d}\mu\leq1\Big\},\ \text{supremum atteint pour } g=\log(f)-\log\int f\mathrm{d}\mu.$$

En particulier, l'inégalité ≤1 peut être remplacée par l'égalité $=1$.

Démonstration. Découle de l'inégalité de convexité élémentaire $uv\leq u\log(u)-u+\mathrm{e}^v$ valable pour $u\geq0$ et $v\in\mathbb{R}$. Alternativement, et dans un registre plus fonctionnel, après s'être ramené par homogénéité à $\int f\mathrm{d}\mu=1$, il est possible d'utiliser l'inégalité de Jensen pour la fonction concave log et la mesure de probabilité $f\mathrm{d}\mu$:

$$\int fg\mathrm{d}\mu=\int f\log(f)\mathrm{d}\mu+\int\log\Big(\frac{\mathrm{e}^g}{f}\Big)f\mathrm{d}\mu\leq\int f\log(f)\mathrm{d}\mu+\log\int\frac{\mathrm{e}^g}{f}f\mathrm{d}\mu\leq\int f\log(f)\mathrm{d}\mu.$$

Voir aussi la remarque 2.5.6 pour l'interprétation en terme de transformée de Legendre (dualité convexe). □

Lemme 2.3.2. Tensorisation de l'entropie relative.

Si $\mu_1,\dots,\mu_n$ sont des mesures de probabilités sur $E_1,\dots,E_n$ et si $\mu=\mu_1\otimes\cdots\otimes\mu_n$ est la mesure produit sur l'espace produit $E=E_1\times\cdots\times E_n$ alors pour tout $f:E\to\mathbb{R}_+$, $f\in L^1(\mu)$, on a

$$\mathrm{Ent}_\mu(f)\le\sum_{i=1}^n\int\mathrm{Ent}_{\mu_i}(f)\mathrm{d}\mu,$$

où $\mathrm{Ent}_{\mu_i}(f)$ désigne l'entropie relative calculée sur la i-ème coordonnée en fixant les autres.

Démonstration. Par récurrence sur n, il suffit de traiter le cas $n=2$. Soit $g:E\to\mathbb{R}$ telle que $\int \mathrm{e}^g\mathrm{d}\mu=1$. On a

$$g=g_1+g_2\quad\text{avec}\quad g_1:=g-\log\int\mathrm{e}^g\mathrm{d}\mu_1\quad\text{et}\quad g_2:=\log\int\mathrm{e}^g\mathrm{d}\mu_1,$$

de sorte que $\int\mathrm{e}^{g_1}\mathrm{d}\mu_1=1$ et $\int\mathrm{e}^{g_2}\mathrm{d}\mu_2=1$. La formule variationnelle du lemme 2.3.1 pour μ_i et g_i donne

$$\int fg_1\mathrm{d}\mu_1+\int fg_2\mathrm{d}\mu_2\le\mathrm{Ent}_{\mu_1}(f)+\mathrm{Ent}_{\mu_2}(f),\quad\text{d'où}\quad\int fg\mathrm{d}\mu\le\int\mathrm{Ent}_{\mu_1}(f)\mathrm{d}\mu_1+\int\mathrm{Ent}_{\mu_2}(f)\mathrm{d}\mu_2,$$

et il ne reste plus qu'à utiliser à nouveau la formule variationnelle du lemme 2.3.1, cette fois pour μ et g. □

Théorème 2.3.3. Inégalité de Sobolev logarithmique (ISL) gaussienne.

Pour tout $n\ge1$ et toute fonction $f\in L^2(\gamma^n)\cap\mathscr{C}^2(\mathbb{R}^n,\mathbb{R})$, l'inégalité suivante a lieu dans $[0,+\infty]$:

$$\mathrm{Ent}_{\gamma^n}(f^2)\le2\int|\nabla f|^2\mathrm{d}\gamma^n.$$

De plus la constante 2 est optimale au sens où l'égalité est atteinte pour $f^2(x)=\mathrm{e}^{\langle\lambda,x\rangle}$, $\lambda\in\mathbb{R}^n$.

— ISL compare deux quantités, homogènes d'ordre 2, qui mesurent la non-constance de f.
— On dit aussi « log-Sobolev », « log-Sob », et même parfois « sobolog ».
— La classe de fonctions test et les deux membres sont homogènes par multiplication par une constante.
— L'inégalité n'est utile que lorsque le second membre est fini : $|\nabla f|\in L^2(\gamma^n)$.
— Des techniques de régularisation permettent d'étendre l'inégalité à l'espace de Sobolev $H^2(\gamma^n)=W^{1,2}(\gamma^n)$.
— C'est une inégalité fonctionnelle, tout comme l'inégalité de Sobolev, qui a inspiré son nom. En l'écrivant

$$\int f^2\log(f^2)\mathrm{d}\gamma^n\le\int f^2\mathrm{d}\gamma^n\log\int f^2\mathrm{d}\gamma^n+2\int|\nabla f|^2\mathrm{d}\gamma^n,$$

elle permet d'affirmer que $f^2\log(f^2)\in L^1(\gamma^n)$ dès que $f^2\in L^1(\gamma^n)$ et $|\nabla f|^2\in L^1(\gamma^n)$.
— Par un changement de variable affine, on obtient une ISL pour $\mathscr{N}(m,\Sigma)$, de constante $\|\Sigma\|_{\mathrm{op}}^2$. Au-delà des gaussiennes, un théorème de Bakry et Émery (processus de Markov) mais aussi de Caffarelli (transport optimal) affirme qu'une mesure de Boltzmann–Gibbs $\mathrm{d}\mu(x)=\frac{1}{Z}\mathrm{e}^{-V(x)}\mathrm{d}x$ vérifie une ISL lorsqu'il existe une constante $\rho>0$ telle que $(\mathrm{Hess}V)(x)\ge\rho$ au sens des formes quadratiques, pour tout $x\in\mathbb{R}^n$. Cela est équivalent à dire que μ est à densité log-concave par rapport à la gaussienne isotrope $\gamma^n_{1/\rho}$. En mécanique statistique, plusieurs techniques ont été développées pour l'établissement d'une ISL pour les mesures de Boltzmann–Gibbs non-produit décrivant des systèmes de spins en interactions. Par ailleurs, les ISL sont étudiées en analyse fonctionnelle en rapport avec la géométrie des corps convexes.
— La linéarisation de ISL via $f^2=(1+\varepsilon g)^2$ donne une inégalité de Poincaré, cf. TD

$$\int f^2\mathrm{d}\gamma^n-\Big(\int f\mathrm{d}\gamma^n\Big)^2\le\int|\nabla f|^2\mathrm{d}\gamma^n.$$

Démonstration. Étape 1 : inégalité à deux points. Si $\nu=\frac12(\delta_{-1}+\delta_1)$ désigne la loi uniforme sur $\{-1,1\}$, alors

$$\mathrm{Ent}_\nu(g^2)\le\frac{(g(1)-g(-1))^2}{2}\quad\text{pour tout } g:\{-1,1\}\to\mathbb{R},$$

qui peut également se réduire, par homogénéité, à l'inégalité optimale univariée (égalité atteinte quand $u=1$)

$$u\log(u)+(2-u)\log(2-u)\le(\sqrt{u}-\sqrt{2-u})^2,\quad0\le u\le2.$$

Étape 2 : TLC Soit à présent $f \in \mathscr{C}^2_c(\mathbb{R},\mathbb{R})$ et $g : \{-1,1\}^n \to \mathbb{R}$ donnée par

$$g(x_1,\dots,x_n) = f\Big(\frac{x_1+\cdots+x_n}{\sqrt{n}}\Big).$$

Soit $\mu := \nu^{\otimes n}$ la loi uniforme sur le cube $\{-1,1\}^n$. Par tensorisation (lemme 2.3.2) et par l'inégalité sur $\{-1,1\}$,

$$\mathrm{Ent}_\mu(g^2) \le \frac{1}{2}\int \sum_{i=1}^n (g(x^{i,+}) - g(x^{i,-}))^2 \mathrm{d}\mu$$

où $x_j^{i,\pm} := x_j$ si $j \neq i$ et $:= \pm 1$ si $j = i$. Une formule de Taylor à l'ordre 1 pour f en $\frac{x_1+\cdots+x_n}{\sqrt{n}}$ donne

$$g(x^{i,+}) - g(x^{i,-}) = \frac{2}{\sqrt{n}} f'\Big(\frac{x_1+\cdots+x_n}{\sqrt{n}}\Big) + o\Big(\frac{1}{\sqrt{n}}\Big)$$

avec un o uniforme en x car f est $\mathscr{C}^2_c$ et donc à dérivée seconde bornée. D'où, grâce au TLC,

$$\mathrm{Ent}_{\gamma_1}(f^2) \le 2\int f'^2 \mathrm{d}\gamma^1.$$

L'inégalité pour $\gamma^n = (\gamma^1)^{\otimes n}$ s'en déduit en tensorisant à nouveau !

Étape 3 : approximation. Si $f \in \mathscr{C}^2(\mathbb{R}^n,\mathbb{R}) \cap L^2(\gamma^n)$, $|\nabla f| \in L^2(\gamma^n)$, $\int f^2 \mathrm{d}\gamma^n = 1$, et si $\eta_m \in \mathscr{C}^2(\mathbb{R}^n,\mathbb{R})$ avec $\mathbb{1}_{|x|\le m} \le \eta_m \le \mathbb{1}_{|x|\le m+1}$ et $|\nabla\eta| \le \mathbb{1}_{m\le|x|\le m+1}$, $m \ge 1$, alors en considérant la troncature lisse à support compact $f_m := \eta_m f \in \mathscr{C}^2_c \subset L^2(\gamma^n)$, et en procédant par le lemme de Fatou, l'ISL, puis par convergence dominée,

$$\int f^2\log(f^2)\mathrm{d}\gamma^n \le \varliminf_{m\to\infty}\int f_m^2\log(f_m^2)\mathrm{d}\gamma^n \le \varliminf_{m\to\infty}\int |\eta_m\nabla f + f\nabla\eta_m|^2\mathrm{d}\gamma^n + \int f_m^2\mathrm{d}\gamma^n \log\int f_m^2\mathrm{d}\gamma^n = \int |\nabla f|^2\mathrm{d}\gamma^n.$$

Si $\int f^2\mathrm{d}\gamma^n \neq 1$, on se réduit au cas précédent par homogénéité :

$$\mathrm{Ent}_{\gamma^n}(f^2) = \Big(\int f^2\mathrm{d}\gamma^n\Big)\int g^2\log(g^2)\mathrm{d}\gamma^n \quad \text{avec} \quad g := \frac{f}{\sqrt{\int f^2\mathrm{d}\gamma^n}} \quad \text{qui vérifie bien} \quad \int g^2\mathrm{d}\gamma^n = 1.$$

Enfin, si $|\nabla f| \notin L^2(\gamma^n)$ alors le membre de droite de l'ISL est infini ce qui la rend à la fois vraie et inutile.

Détaillons enfin un argument alternatif à la troncature à support compact. Soit $\varphi_m \in \mathscr{C}^2(\mathbb{R},\mathbb{R})$ telle que $\varphi' \ge 0$, $\|\varphi'\|_\infty \le 1$, $\|\varphi''\|_\infty \le 1$, $\varphi_m(x) = x$ pour tout $x \in [-m,m]$, $|\varphi_m| \le (m+1)$, $\varphi_m(x) = -(m+1)$ si $x \le -(m+1)$ et $\varphi_m(x) = m+1$ si $x \ge m+1$. Supposons qu'on a établi l'ISL pour $f \in \mathscr{C}^2_b(\mathbb{R}^n,\mathbb{R}) : \mathscr{C}^2$, bornée, et à dérivées première et seconde bornées. Soit $f \in \mathscr{C}^2(\mathbb{R}^n,\mathbb{R}) \cap L^2(\gamma^n)$ telle que $|\nabla f|^2 \in L^2(\gamma^n)$, et $\int f^2\mathrm{d}\gamma^n = 1$. Alors pour tout $m \ge 1$, $f_m := \varphi_m \circ f \in \mathscr{C}^2_b(\mathbb{R}^n,\mathbb{R})$, et en procédant par le lemme de Fatou, l'ISL, puis par convergence dominée,

$$\int f^2\log(f^2)\mathrm{d}\gamma^n \le \varliminf_{m\to\infty}\int f_m^2\log(f_m^2)\mathrm{d}\gamma^n \le \varliminf_{m\to\infty}\int \varphi_m'(f)^2|\nabla f|^2\mathrm{d}\gamma^n + \int f_m^2\mathrm{d}\gamma^n\log\int f_m^2\mathrm{d}\gamma^n = \int |\nabla f|^2\mathrm{d}\gamma^n.$$

Cette méthode de troncature dans l'espace d'arrivée plutôt qu'en support dans l'espace de départ a l'avantage d'être également efficace pour l'inégalité de Poincaré. Notons au passage qu'en quelque sorte $\eta_m = \varphi_m'$. □

2.4 Inégalité de Sobolev logarithmique et concentration sous-gaussienne

Rappelons que $f : \mathbb{R}^n \to \mathbb{R}$ est Lipschitz quand $\|f\|_{\mathrm{Lip}} := \sup_{x\neq y} \frac{|f(x)-f(y)|}{|x-y|} < \infty$. Une fonction Lipschitz est toujours uniformément continue et en particulier continue, et un théorème de Hans Rademacher affirme qu'elle est dérivable presque partout. Quand f est $\mathscr{C}^1$ alors $\|f\|_{\mathrm{Lip}} = \sup_{x\in\mathbb{R}^n} |\nabla f(x)| = \||\nabla f|\|_\infty$.

Théorème 2.4.1. ISL ⇒ Transformée de Laplace sous-gaussienne.

Soit μ une mesure de probabilité sur $\mathbb{R}^n$, $n \ge 1$, qui vérifie une inégalité de Sobolev logarithmique

$$\exists c \in \mathbb{R}_+,\ \ \forall f \in L^2(\mu)\cap\mathscr{C}^2(\mathbb{R}^n,\mathbb{R}),\ \ \mathrm{Ent}_\mu(f^2) \le c\int |\nabla f|^2\mathrm{d}\mu.$$

Alors les fonctions Lipschitz et intégrables ont une transformée de Laplace sous-gaussienne :

$$\forall f:\mathbb{R}^n\to\mathbb{R} \text{ Lipschitz et dans } L^1(\mu), \forall\theta\in\mathbb{R},\ L(\theta):=\int \exp(\theta f)\mathrm{d}\mu \le \exp\left(\theta^2\frac{c}{4}\|f\|_{\mathrm{Lip}}^2+\theta\int f\mathrm{d}\mu\right).$$

— Grâce au théorème 2.3.3, le résultat est donc valable pour $\mu=\gamma^n$ avec $c=1$.
De plus dans ce cas l'inégalité devient une égalité lorsque $f=\langle\cdot,u\rangle$ avec $u\in\mathbb{R}^n$, $|u|=1$.
— Le résultat s'écrit de manière équivalente sous forme de majoration adimensionnelle pour f recentrée :

$$\int \exp\Big(\theta\Big(f-\int f\mathrm{d}\mu\Big)\Big)\mathrm{d}\mu \le \exp\Big(\theta^2\frac{c}{4}\|f\|_{\mathrm{Lip}}^2\Big).$$

— L'inégalité sur L est un analogue Lipschitz(gaussien) de celle de Hoeffding pour variables bornées.
— Notons que f est Lipschitz donc sous-linéaire : $|f(x)|\le\|f\|_{\mathrm{Lip}}|x|+|f(0)|$.
— L'argument de tensorisation dans la preuve du théorème 2.3.3 indique que si μ,ν vérifient ISL avec constantes c_μ et c_ν alors $\mu\otimes\nu$ vérifie ISL avec constante $\max(c_\mu,c_\nu)$. En particulier si μ vérifie ISL avec constante c alors $\mu^{\otimes N}$ vérifie ISL avec la même constante c, pour tout N. Ce comportement adimensionnel de ISL implique une borne adimensionnelle sur la transformée de Laplace des fonctions Lipschitz.
— Contrairement à ISL, la sous-gaussianité de la transformée de Laplace des fonctions Lipschitz ne se tensorise pas, au sens où on ne sait pas déduire la propriété pour la mesure de probabilité produit $\mu\otimes\nu$ à partir de celles pour μ et ν, avec le maximum des constantes. D'où l'intérêt de ISL quand elle a lieu !
— La mesure de probabilité $\frac{1}{Z_\alpha}\mathrm{e}^{-|x|^\alpha}\mathrm{d}x$ sur $\mathbb{R}$, $\alpha>0$, $Z_\alpha:=\int_{\mathbb{R}}\mathrm{e}^{-|x|^\alpha}\mathrm{d}x<\infty$, vérifie ISL ssi $\alpha\ge 2$, et une inégalité de Poincaré ssi $\alpha\ge 1$, cf. [2, Chapitre 6]. La gaussienne correspond à $\alpha=2$.
— En utilisant le changement de fonction $g=f^2$, l'ISL peut-être réécrite en terme d'entropie relative comme suit : pour toute mesure de probabilité ν sur $\mathbb{R}^n$ telle que $\nu\ll\mu$ et $g:=\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\in L^1(\mu)\cap\mathscr{C}^2(\mathbb{R}^n,\mathbb{R})$,

$$\mathrm{H}(\nu\,|\,\mu)\le\frac{4}{2}\mathrm{I}(\nu\,|\,\mu)\quad\text{où}\quad \mathrm{I}(\nu\,|\,\mu):=4\int|\nabla\sqrt{g}|^2\mathrm{d}\mu=\int\frac{|\nabla g|^2}{g}\mathrm{d}\mu=\int|\nabla\log(g)|^2\mathrm{d}\nu.$$

Le second membre est l'information de Fisher de ν par rapport à μ.
— Le concept d'ISL a été dégagé par Leonard Gross vers 1973, auteur également de la preuve de l'ISL gaussienne en utilisant le TLC pour le cube discret. La variante ci-dessous basée sur la tensorisation de l'entropie est due à Sergey Bobkov. L'ISL gaussienne était déjà présente dans les travaux de Paul Federbush.

Démonstration. Observons qu'on peut supposer sans perte de généralité que f est bornée et $\mathscr{C}^\infty$ en utilisant une approximation par troncature et régularisation, détaillée plus loin. On peut également supposer que $\theta>0$ quitte à remplacer f par $-f$, que f est centrée pour μ, et que $\|f\|_{\mathrm{Lip}}=1$ par translation et dilatation.

L'idée est maintenant la suivante [6] : pour tout $\theta>0$, l'inégalité de Sobolev logarithmique pour μ du théorème 2.3.3 avec $\mathrm{e}^{\theta f}$ au lieu de f^2 donne, via $|\nabla\mathrm{e}^{\frac{1}{2}\theta f}|=|\frac{\theta}{2}\nabla f|\mathrm{e}^{\theta f}$ et $\||\nabla f|\|_\infty=\|f\|_{\mathrm{Lip}}\le 1$, que

$$\theta L'(\theta)-L(\theta)\log L(\theta)\le\tfrac{c}{4}\theta^2L(\theta),\quad\text{c'est-à-dire}\quad K'\le\tfrac{c}{4}\ \text{où}\ K(\theta):=\tfrac{1}{\theta}\log L(\theta).$$

Le résultat découle alors de $K(0)=(\log L)'(0)=L'(0)/L(0)=\mu(f)$, qui vient de $L(0)=1$ et $L'(0)=\mu(f)$.

Détaillons enfin l'argument d'approximation. Si f est Lipschitz, on définit, pour tous $k\ge1$ et $\varepsilon>0$, la fonction $f_{k,\varepsilon}:=\max(-k,\min(f,k))*\rho_\varepsilon$ où $\rho_\varepsilon\in\mathscr{C}^\infty(\mathbb{R}^n,\mathbb{R})$ vérifie [7]

$$\mathrm{supp}(\rho_\varepsilon)\subset\{x\in\mathbb{R}^n:|x|\le\varepsilon\},\quad\rho_\varepsilon\ge0,\quad\text{et}\int_{\mathbb{R}^n}\rho_\varepsilon(x)\mathrm{d}x=1.$$

alors $\|f_{k,\varepsilon}\|_{\mathrm{Lip}}\le\|f\|_{\mathrm{Lip}}$ et $f_{k,\varepsilon}\to f$ ponctuellement et dans $L^1(\mu)$ quand $k\to\infty$ et $\varepsilon\to0$. En effet, on a tout d'abord $f_{k,\varepsilon}=f_k*\rho_\varepsilon$, $f_k:=\max(-k,\min(f,k))$, et $\|f_k\|_{\mathrm{Lip}}\le\|f\|_{\mathrm{Lip}}$. Maintenant $\|f_{k,\varepsilon}\|_{\mathrm{Lip}}\le\|f\|_{\mathrm{Lip}}$ car

$$|f_{k,\varepsilon}(x)-f_{k,\varepsilon}(y)|\le\int_{\mathbb{R}^n}|f_k(x-z)-f_k(y-z)|\rho_\varepsilon(z)\mathrm{d}z\le\|f\|_{\mathrm{Lip}}|x-y|.$$

Pour la convergence ponctuelle de $f_{k,\varepsilon}$ vers f, on observe tout d'abord que si $|x-y|\le\varepsilon$ alors

$$|f_k(y)-f(x)|\le|f_k(y)-f_k(x)|+|f_k(x)-f(x)|\le\varepsilon\|f\|_{\mathrm{Lip}}+|f(x)|\mathbb{1}_{|f(x)|\ge k}\underset{\substack{k\to\infty\\ \varepsilon\to0}}{\longrightarrow}0,$$

6. On parle d'argument de Herbst. Ira Herbst l'a communiqué à Leonard Gross, inventeur du concept d'ISL en 1975, et Oscar Rothaus, qui l'on publié en 1998. Michel Ledoux l'a ensuite popularisé dans ses écrits synthétiques sur la concentration de la mesure.

7. C'est un noyau régularisant (*mollifier*). Exemple : $\rho_\varepsilon(x)=\varepsilon^{-n}\rho(\varepsilon^{-1}|x|)$ avec $\rho(x):=c^{-1}\exp\Big(-\frac{1}{1-x^2}\Big)\mathbb{1}_{|x|<1}$.

et donc, pour tout $x \in \mathbb{R}^n$,

$$|f_{k,\varepsilon}(x) - f(x)| \leq \int_{|x-y|\leq\varepsilon} |f_k(y) - f(x)|\rho_\varepsilon(x-y)\mathrm{d}y \leq \varepsilon\|f\|_{\mathrm{Lip}} + |f(x)|\mathbb{1}_{|f(x)|\geq k} \underset{\substack{k\to\infty\\ \varepsilon\to 0}}{\longrightarrow} 0.$$

La convergence dans $L^1(\mu)$ de $f_{k,\varepsilon}$ vers f s'obtient par convergence dominée

$$\begin{aligned}\int |f_{k,\varepsilon}(x) - f(x)|\mu(\mathrm{d}x) &\leq \iint |f_k(y) - f(x)|\rho_\varepsilon(x-y)\mathrm{d}y\mu(\mathrm{d}x)\\ &\leq \iint_{|x-y|\leq\varepsilon} |f_k(y) - f_k(x)|\rho_\varepsilon(x-y)\mathrm{d}y\mu(\mathrm{d}x) + \iint |f_k(x) - f(x)|\rho_\varepsilon(x-y)\mathrm{d}y\mu(\mathrm{d}x)\\ &\leq \varepsilon\|f\|_{\mathrm{Lip}} + \int |f(x)|\mathbb{1}_{f(y)\geq k}(x)\mu(\mathrm{d}x) \underset{\substack{k\to\infty\\ \varepsilon\to 0}}{\longrightarrow} 0.\end{aligned}$$

Au bout du compte, si le résultat est vrai pour $f_{k,\varepsilon}$ alors, par le lemme de Fatou (première inégalité),

$$\begin{aligned}\int \mathrm{e}^{\theta f}\mathrm{d}\mu = \int \lim_{\substack{k\to\infty\\ \varepsilon\to 0}} \mathrm{e}^{\theta f_{k,\varepsilon}}\mathrm{d}\mu &\leq \varliminf_{\substack{k\to\infty\\ \varepsilon\to 0}} \int \mathrm{e}^{\theta f_{k,\varepsilon}}\mathrm{d}\mu \leq \varliminf_{\substack{k\to\infty\\ \varepsilon\to 0}} \exp\Big(\theta^2\frac{c}{2}\|f_{k,\varepsilon}\|^2_{\mathrm{Lip}} + \theta\int f_{k,\varepsilon}\mathrm{d}\mu\Big)\\ &\leq \exp\Big(\theta^2\frac{c}{2}\|f\|^2_{\mathrm{Lip}} + \theta\int f\mathrm{d}\mu\Big).\end{aligned}$$

□

Corollaire 2.4.2. ISL ⇒ Concentration sous-gaussienne.

Soit μ une mesure de probabilité sur $\mathbb{R}^n$ vérifiant une inégalité de Sobolev logarithmique de constante c comme dans le théorème 2.4.1. Soit $f : \mathbb{R}^n \to \mathbb{R}$ Lipschitz et dans $L^1(\mu)$. Si $X \sim \mu$ alors pour tout $r \geq 0$,

$$\mathbb{P}\Big(\Big|f(X) - \mathbb{E}(f(X))\Big| \geq r\Big) \leq 2\exp\left(-\frac{r^2}{c\|f\|^2_{\mathrm{Lip}}}\right).$$

Plus généralement, si $X_1, \ldots, X_N$, $N \geq 1$, sont i.i.d. de loi μ, alors pour tout $r \geq 0$,

$$\mathbb{P}\left(\Big|\frac{f(X_1)+\cdots+f(X_N)}{N} - \mathbb{E}(f(X_1))\Big| \geq r\right) \leq 2\exp\left(-\frac{Nr^2}{c\|f\|^2_{\mathrm{Lip}}}\right).$$

— La preuve donne en fait des inégalités de déviation, sans valeur absolue ni préfacteur 2 devant exp.
— Si $\mu = \gamma^n$ et si $f(x) = \langle x, \theta\rangle$, $|\theta| = 1$, alors $c = 1$, $\sqrt{N}\Big(\frac{f(X_1)+\cdots+f(X_N)}{N} - \mathbb{E}(f(X_1))\Big) \sim \gamma^1$, et $\|f\|_{\mathrm{Lip}} = 1$.
— Il est possible de réécrire l'inégalité sous une forme adimensionnelle :

$$\mathbb{P}\left(\sqrt{N}\Big|\frac{f(X_1)+\cdots+f(X_N)}{N} - \mathbb{E}(f(X_1))\Big| \geq r\right) \leq 2\exp\left(-\frac{r^2}{c\|f\|^2_{\mathrm{Lip}}}\right).$$

— La concentration n'est pas directement tensorisable, d'où l'avantage de passer par ISL dans la preuve.
— Une conséquence est l'intégrabilité exponentielle pour $Y = f(X)$, cf. TD :

$$\mathbb{E}(\mathrm{e}^{\theta(Y-\mathbb{E}(Y))^2}) = \theta\int_0^\infty r\mathrm{e}^{\theta r^2}\mathbb{P}(|Y - \mathbb{E}(Y)| \geq r)\mathrm{d}r < \infty \quad \text{dès que } \theta < \frac{1}{c\|f\|^2_{\mathrm{Lip}}}.$$

L'intégrabilité exponentielle du carré joue un rôle important dans l'analyse gaussienne.

Démonstration. Première inégalité. On se ramène à $\|f\|_{\mathrm{Lip}} = 1$ et $\mu(f) = \int f\,\mathrm{d}\mu = 0$ pas translation et dilatation. Ensuite, pour tous $r \geq 0$ et $\theta > 0$, par l'inégalité de Markov et le théorème 2.4.1,

$$\mu(f \geq r) = \mu\left(\mathrm{e}^{\theta f} \geq \mathrm{e}^{\theta r}\right) \leq \mathrm{e}^{-\theta r}\int \mathrm{e}^{\theta f}\,\mathrm{d}\mu \leq \mathrm{e}^{-\theta r + \frac{c}{4}\theta^2} \leq \mathrm{e}^{-\frac{r^2}{c}},$$

où la dernière inégalité est obtenue avec le choix optimal $\theta = 2r/c$. En utilisant le résultat pour f et $-f$ il vient

$$\mu\Big(\Big|f - \int f\mathrm{d}\mu\Big| \geq r\Big) \leq 2\exp\left(-\frac{r^2}{2c\|f\|^2_{\mathrm{Lip}}}\right).$$

Deuxième inégalité. On observe que la fonction $x \in (\mathbb{R}^n)^N \mapsto F(x) := \frac{1}{N}(f(x_1)+\cdots+f(x_N))$ est Lipschitz avec

$$\|F\|_{\mathrm{Lip}} \le \frac{\|f\|_{\mathrm{Lip}}}{N} \sup_{x\neq y} \frac{\sum_{i=1}^N |x_i-y_i|}{\sqrt{\sum_{i=1}^N |x_i-y_i|^2}} \le \frac{\|f\|_{\mathrm{Lip}}}{\sqrt{N}}.$$

De plus $\mathbb{E}(F(X_1,\dots,X_N)) = \mathbb{E}(f(X_1))$, et $(X_1,\dots,X_N) \sim \mu^{\otimes N}$ vérifie une inégalité de log-Sobolev de même constante $2c$ (ne dépend pas de N, procéder par tensorisation comme dans la preuve du théorème 2.3.3). □

2.5 Inégalité de transport de Talagrand

Je 13/02

Abordé en cours :
— L'intégralité jusqu'à la fin du chapitre sauf preuve de la formulation duale de Kantorovich–Rubinstein

Soit (E,d) un espace polonais comme $(\mathbb{R}^n,|\cdot|)$, muni de sa tribu borélienne, et soit $\mathscr{P}_1(E)$ l'ensemble des mesures de probabilités μ sur E possédant un moment d'ordre 1 : pour un $x_0 \in E$,

$$\int d(x,x_0)\mu(\mathrm{d}x) < \infty, \quad \text{propriété alors valable pour tout } x_0 \in E \text{ par l'inégalité triangulaire.}$$

La distance de couplage, de transport, de Kantorovitch, ou de Wasserstein sur $\mathscr{P}_1(E)$ est définie par[8]

$$\mathrm{W}_1(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \iint_{E\times E} d(x,y)\pi(\mathrm{d}x,\mathrm{d}y) = \inf_{\substack{(X,Y)\\ X\sim\mu, Y\sim\nu}} \mathbb{E}(d(X,Y)), \quad \text{pour tous } \mu,\nu\in\mathscr{P}_1(E),$$

où $\Pi(\mu,\nu)$ est l'ensemble[9] des mesures de probabilités sur l'espace produit $E\times E$ de lois marginales μ et ν. L'ensemble $\Pi(\mu,\nu)$ est convexe, non-vide car $\mu\otimes\nu\in\Pi(\mu,\nu)$. En interprétant π comme un plan de transport[10], la quantité $\mathrm{W}_1(\mu,\nu)$ est un coût optimal du transport entre μ et ν.

Lemme 2.5.1. Distance de couplage / de transport / de Kantorovich / de Wasserstein.

W_1 est une distance sur $\mathscr{P}_1(E)$.

Démonstration. La symétrie est immédiate. Pour tous $\mu,\nu\in\mathscr{P}_1(E)$ et $x,y,z\in E$, on a $d(x,y)\le d(x,z)+d(z,y)$, et comme Π n'est pas vide, $\mathrm{W}_1(\mu,\nu)\le \mathrm{W}_1(\mu,\delta_z)+\mathrm{W}_1(\delta_z,\nu)$, qui est $<\infty$ car μ et ν ont un moment d'ordre 1. Cet argument donne l'inégalité triangulaire pour W_1, en considérant un triplage ou collage : une mesure de probabilité sur $E\times E\times E$ dont les projections sur les deux premiers et les deux derniers facteurs sont prescrites et compatibles, dont on trouve une preuve dans [79, Le. 7.6] ou [72, Sec. 5.1] par désintégration de mesure[11].

Il est clair que $\mathrm{W}_1(\mu,\mu)=0$. Enfin, si $\mathrm{W}_1(\mu,\nu)=0$, alors la définition variationnelle de $\mathrm{W}_1(\mu,\nu)$ donne une suite (X_n,Y_n) telle que $\mathbb{E}(d(X_n,Y_n))\to 0$, donc pour tout $f:E\to\mathbb{R}$ telle que $\|f\|_\infty<\infty$ et $\|f\|_{\mathrm{Lip}}<\infty$,

$$\left|\int f\mathrm{d}(\mu-\nu)\right| = |\mathbb{E}(f(X_n)-f(Y_n))| \le \mathbb{E}(|f(X_n)-f(Y_n)|) \le \|f\|_{\mathrm{Lip}}\mathbb{E}(d(X_n,Y_n)) \to 0,$$

donc $\mu=\nu$ sont égales sur les fonctions test Lipschitz bornées, donc $\mu=\nu$. □

Si μ et ν sont des mesures de probabilités sur E, rappelons que $\mathrm{H}(\nu\mid\mu) := \mathrm{Ent}_\mu(f)\in[0,+\infty]$ désigne l'entropie relative de ν par rapport à μ, où $f := \mathrm{d}\nu/\mathrm{d}\mu$, avec la convention $\mathrm{H}(\nu\mid\mu)=+\infty$ si $\nu\not\ll\mu$.

Théorème 2.5.2. Inégalité de transport de Talagrand et concentration : argument de Marton.

Supposons que $\mu\in\mathscr{P}_1(E)$ vérifie une inégalité de transport de Talagrand :

$$\exists c\in\mathbb{R}_+,\ \ \forall\nu\in\mathscr{P}_1(E),\ \ \mathrm{W}_1(\mu,\nu)\le\sqrt{c\mathrm{H}(\nu\mid\mu)}.$$

8. Comment construire une v.a. qui suit une loi prescrite arbitraire sur un espace mesuré quelconque ? Il est possible de procéder par approximation si la tribu ou la topologie qui l'engendre se ramène au discret fini, ce que permet par exemple le caractère polonais de E.

9. Lorsque E est fini et μ et ν sont uniformes, cet ensemble est celui des matrices doublement stochastiques.

10. Ces distances ont été étudiées par Leonid Vitalievitch Kantorovitch (1912 – 1986), prix Nobel d'économie 1975, un exploit pour un mathématicien soviétique, mais aussi Cédric Villani (1973 –), médaillé Fields 2010, comme relaxation convexe du problème du transport optimal de Gaspard Monge (1746 – 1818). Citons également Yann Brenier (1957 –), Luis Caffarelli (1948 –), et Alessio Figalli (1984 –).

11. La désintégration de mesure est liée à la notion de conditionnement, mais ici seule une formulation élémentaire pour les mesures est nécessaire, sous forme de lemme, en particulier nul besoin ici de variables aléatoires ni d'espérance conditionnelle.

Alors pour tout $f : E \to \mathbb{R}$ Lipschitz, il existe $r_* \in \mathbb{R}_+$ tel que pour tout $r \geq r_*$,

$$\mu\Big(f \geq \int f \mathrm{d}\mu + r\|f\|_{\mathrm{Lip}}\Big) \leq \exp\Big(-\frac{(r-r_*)^2}{c}\Big).$$

— Ce théorème est du à Katalin Marton (1941 – 2019).
— Pour l'esthétique, c'est ici la déviation qui incorpore $\|f\|_{\mathrm{Lip}}$ plutôt que la borne exponentielle.
— Rappel : $f : E \to \mathbb{R}$ est Lipschitz ssi $\|f\|_{\mathrm{Lip}} := \sup_{x \neq y} \frac{|f(x)-f(y)|}{d(x,y)} < \infty$.
— Comme $\mu \in \mathscr{P}_1(E)$ et f est Lipschitz, on a $f \in L^1(\mu)$ automatiquement.

Démonstration. Soient $A, B \subset E$ des boréliens, $\mu(A) > 0$ et $\mu(B) > 0$, et soient les mesures de probabilités [12]

$$\mu_A := \frac{\mu(\cdot \cap A)}{\mu(A)} \quad \text{et} \quad \mu_B := \frac{\mu(\cdot \cap B)}{\mu(B)}, \quad \text{de densités} \quad f_A := \frac{\mathbb{1}_A}{\mu(A)} \quad \text{et} \quad f_B := \frac{\mathbb{1}_B}{\mu(B)}.$$

L'inégalité triangulaire pour W_1 donne $\mathrm{W}_1(\mu_A, \mu_B) \leq \mathrm{W}_1(\mu_A, \mu) + \mathrm{W}_1(\mu_B, \mu)$, ce qui permet d'utiliser l'inégalité de transport de Talagrand $\mathrm{W}_1(\mu, \nu) \leq \sqrt{c\mathrm{Ent}_\mu(\mathrm{d}\nu/\mathrm{d}\mu)}$, ce qui donne

$$\mathrm{W}_1(\mu_A, \mu_B) \leq \sqrt{c\mathrm{Ent}_\mu(f_A)} + \sqrt{c\mathrm{Ent}_\mu(f_B)} = \sqrt{-c\log\mu(A)} + \sqrt{-c\log\mu(B)}.$$

À présent, choisissons l'ensemble B à partir de l'ensemble A en posant, pour un réel $r \geq 0$ arbitraire,

$$B := (A_r)^c := \{x \in E : \mathrm{dist}(x, A) \geq r\}.$$

Pour un tel choix, la distance entre les supports des mesures de probabilités μ_A et μ_B est $\geq r$, donc par la définition « couplage » de W_1, on a $r \leq \mathrm{W}_1(\mu_A, \mu_B)$, et donc, si $r \geq r_* := \sqrt{-c\log\mu(A)}$, alors

$$\mu((A_r)^c) = \mu(B) \leq \exp\left(-\frac{(r-r_*)^2}{c}\right).$$

Si à présent $f : E \to \mathbb{R}$ est Lipschitz, alors par l'inégalité triangulaire, pour tout réel $r > 0$,

$$\Big\{f \leq \int f \mathrm{d}\mu\Big\}_r \quad \subset \quad \Big\{f < \int f \mathrm{d}\mu + r\|f\|_{\mathrm{Lip}}\Big\}.$$

En prenant à présent $A = \{f \leq \mu(f)\}$, on obtient, pour tout $r \geq r_*$, la borne gaussienne

$$\mu\Big(f - \int f \mathrm{d}\mu > r\|f\|_{\mathrm{Lip}}\Big) \leq \exp\left(-\frac{(r-r_*)^2}{c}\right).$$

□

L'inégalité de déviation fournie par le théorème 2.5.2 n'est pas satisfaisante en raison notamment de la contrainte $r \geq r_*$. Le dépassement de cette difficulté passe par une reformulation de W_1 sous forme de supremum, qui ouvre la voie à une équivalence entre inégalité de transport de Talagrand et concentration sous-gaussienne pour les fonctions Lipschitz, exprimée par le théorème 2.5.5 ci-dessous.

La définition variationnelle de W_1 ci-dessus met en jeu une expression linéaire par rapport à π optimisée sur une contrainte convexe sur π (de dimension infinie). Une telle structure permet une analyse par dualité.

Théorème 2.5.3. Formulation variationnelle duale de W_1 de Kantorovich–Rubinstein.

Pour tous $\mu, \nu \in \mathscr{P}_1(E)$, on a $\mathrm{W}_1(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \left(\int f \mathrm{d}\mu - \int f \mathrm{d}\nu\right)$.

— Par translation, on peut supposer par exemple que $\int f \mathrm{d}\mu = 0$ car μ et ν sont de même masse.

12. En termes probabilistes élémentaires, ce sont des lois conditionnelles : $\mu_A = \mu(\cdot \mid A)$ et $\mu_B = \mu(\cdot \mid B)$.

Démonstration. Si $f : E \to \mathbb{R}$ est Lipschitz avec $\|f\|_{\mathrm{Lip}} \le 1$ alors pour tous $x, y \in E$, $f(x) - f(y) \le d(x,y)$, et donc pour tout couplage $\pi \in \Pi(\mu,\nu)$ sur $E \times E$ de μ et ν, on a $\int f\mathrm{d}\mu - \int f\mathrm{d}\nu \le \iint d(x,y)\pi(\mathrm{d}x,\mathrm{d}y)$, d'où

$$\sup_{\|f\|_{\mathrm{Lip}}\le 1}\Big(\int f\mathrm{d}\mu - \int f\mathrm{d}\nu\Big) \le \inf_{\pi\in\Pi(\mu,\nu)} \iint d(x,y)\pi(\mathrm{d}x,\mathrm{d}y) = \mathrm{W}_1(\mu,\nu).$$

L'égalité provient du théorème de dualité de Kantorovich : si E et F sont des espaces polonais munis de leurs tribus boréliennes, si μ et ν sont des mesures de probabilités sur E et F, et si $c : E \times E \to [0, +\infty]$ est semi-continue inférieurement (c'est-à-dire avec des ensembles de sous-niveau fermés), alors

$$\inf_{\pi\in\Pi(\mu,\nu)} \iint c(x,y)\pi(\mathrm{d}x,\mathrm{d}y) = \sup_{(\phi,\psi)\in\Phi_c}\Big(\int \phi\mathrm{d}\mu + \int \psi\mathrm{d}\nu\Big)$$

où $\Phi_c := \{(\phi,\psi) \in L^1(\mu)\times L^1(\nu) : \phi(x)+\psi(y) \le c(x,y) \text{ pour } \mu\otimes\nu \text{ presque tous } (x,y) \in E\times F\}$. Ce théorème est une conséquence d'un théorème de représentation de Riesz, cf. [79, Th. 1.3 p.19 et sec. 1.1.7 p. 25] et [40, Sec. 2.6].

Comme présenté dans [79, Th. 1.14 p. 34], considérons le cas $E = F$ et $c = d$ une distance. La distance $d_n := d/(1+d/n) \le d$ est bornée, et $d_n \nearrow d$ quand $n \to \infty$, en particulier, si $\|f\|_{\mathrm{Lip}} \le 1$ pour d_n alors c'est aussi le cas pour d. Cela permet de supposer que d est bornée, et dans ce cas, toutes les fonctions Lipschitz sont bornées donc intégrables pour toutes les mesures de probabilités. De plus, dans le théorème de dualité de Kantorovich, on peut se restreindre à $(\phi,\psi) = (f,-f)$ avec $\|f\|_{\mathrm{Lip}} \le 1$. En effet, si $f : E \to \mathbb{R}$, la fonction c-concave conjuguée $f^c : E \to \mathbb{R}\cup\{-\infty\}$ définie par $f^c(x) := \inf_{y\in E}(d(x,y) - f(y))$ vérifie $\|f\|_{\mathrm{Lip}} \le 1$ comme infimum de fonctions Lipschitz. De plus si $\|f\|_{\mathrm{Lip}} \le 1$ alors $f^c = -f$. Ensuite, pour tout $(\phi,\psi) \in \Phi_c$ et tout $(x,y) \in E\times E$, on a $\psi(y) \le d(x,y) - \phi(x)$, donc $\psi(y) \le \phi^c(y)$, d'où

$$\begin{aligned}\int \phi(x)\mu(\mathrm{d}x) + \int \psi(y)\nu(\mathrm{d}y) &\le \int \phi(x)\mu(\mathrm{d}x) + \int \phi^c(y)\nu(\mathrm{d}y)\\ &\le \int (\phi^c)^c(x)\mu(\mathrm{d}x) + \int \phi^c(y)\nu(\mathrm{d}y) = \int \phi^c(y)(\nu-\mu)(\mathrm{d}y).\end{aligned}$$

□

Remarque 2.5.4. Variation totale.

Si d est la distance atomique $d(x,y) = \mathbb{1}_{x\neq y}$, alors $\mathbb{E}(d(X,Y)) = \mathbb{E}(\mathbb{1}_{X\neq Y}) = \mathbb{P}(X \neq Y)$, $\|f\|_{\mathrm{Lip}} = \mathrm{osc}(f)$, et par le théorème 2.5.3, on obtient la distance en variation totale :

$$\mathrm{W}_1(\mu,\nu) = \inf_{(X,Y)} \mathbb{P}(X \neq Y) = \sup_{\|f\|_\infty \le \frac{1}{2}} \int f\mathrm{d}(\mu-\nu).$$

Le théorème suivant renforce le théorème 2.5.2 en affirmant une équivalence !

Théorème 2.5.5. Inégalité de transportation de Talagrand W_1 : caractérisation de Bobkov–Götze.

Pour tout $\mu \in \mathscr{P}_1(E)$ et toute constante $c \in \mathbb{R}_+$, les propriétés suivantes sont équivalentes :

(i) Borne sous-gaussienne pour la transformée de Laplace des fonctions Lipschitz :

$$\forall f : E \to \mathbb{R} \text{ Lipschitz et dans } L^1(\mu),\ \forall \theta \in \mathbb{R},\quad L(\theta) := \int \exp(\theta f)\mathrm{d}\mu \le \exp\Big(\theta^2\frac{c}{4}\|f\|_{\mathrm{Lip}}^2 + \theta\int f\mathrm{d}\mu\Big)$$

(ii) Inégalité de transport de Talagrand :

$$\forall \nu \in \mathscr{P}_1(E),\quad \mathrm{W}_1(\nu,\mu) \le \sqrt{c\mathrm{H}(\nu\mid\mu)}.$$

— En particulier, en combinant cela avec le théorème 2.4.1, on obtient que pour tout $n \ge 1$, la gaussienne $\mu = \gamma^n$ sur $E = \mathbb{R}^n$ vérifie une inégalité de Talagrand de constante $c = 2$.
— Le théorème 2.5.5 a été publié en 1999 par Sergey Bobkov et Friedrich Götze. Depuis une trentaine d'années, les inégalités fonctionnelles, leurs liens, et leurs conséquences ont fait et font encore l'objet d'intenses recherches en analyse fonctionnelle probabiliste et géométrique, avec les travaux de Sergey Bobkov, Michel Ledoux, Cédric Villani, Sourav Chatterjee, Ronen Eldan, Ramon van Handel, et bien d'autres.
— La concentration sous-gaussienne des fonctions Lipschitz (et donc l'inégalité de transport de Talagrand W_1) n'est pas tensorisable, contrairement aux inégalités de log-Sobolev, cf. [2, Ch. 8]. Cela signifie qu'on ne parvient pas à en déduire une inégalité pour $\mu^{\otimes n}$ avec une constante indépendante de n.

— Le (i) s'écrit aussi $\log \int \mathrm{e}^f \mathrm{d}\mu \le \frac{c}{4}\|f\|_{\mathrm{Lip}}^2 + \int f \mathrm{d}\mu$.
— Pour tout $p \ge 1$, il est possible de définir l'ensemble $\mathscr{P}_p(E)$ des mesures possédant un moment fini d'ordre p, et sur cet ensemble la distance $\mathrm{W}_p(\mu,\nu) := (\inf_{(X,Y)} \mathbb{E}(d(X,Y)^p)^{1/p}$. Il se trouve que sur $E = \mathbb{R}^n$, l'inégalité de transport de Talagrand $\mathrm{W}_2(\mu,\nu) \le \sqrt{c\mathrm{H}(\nu \mid \mu)}$ est tensorisable. Par analogie avec la preuve de l'inégalité de log-Sobolev du théorème 2.3.3, il est possible de démontrer une inégalité de Talagrand W_2 sur l'espace à deux points $\{-1,1\}$, puis d'en déduire, par tensorisation et TLC l'inégalité de Talagrand W_2 pour γ^1 puis pour γ^n avec sa constante optimale $c = 2$. Cette approche ne fonctionne pas pour W_1. Il se trouve que l'inégalité de Talagrand W_2 est équivalente à la concentration (sous-gaussienne pour les fonctions Lipschitz) tensorisable, et implique l'inégalité de Talagrand W_1, cf. [2, Ch. 8].

Démonstration. Tout d'abord dans (i) on peut supposer sans perte que $\theta > 0$ en remplaçant f par $-f$ et que $\int f \mathrm{d}\mu = 0$ et $\|f\|_{\mathrm{Lip}} = 1$ par translation et dilatation. À présent, (i) se réécrit, pour une telle fonction $f : E \to \mathbb{R}$,

$$\int \mathrm{e}^g \mathrm{d}\mu \le 1 \quad \text{où} \quad g := \theta f - \theta^2 \tfrac{c}{4},$$

et la formule variationnelle pour l'entropie relative du lemme 2.3.1 donne, pour tout $h : E \to \mathbb{R}_+$, $h \in L^1(\mu)$,

$$\int \Big(\theta f - \theta^2 \frac{c}{4}\Big) h \mathrm{d}\mu \le \mathrm{Ent}_\mu(h).$$

Réciproquement, cette inégalité redonne (i) en prenant $h = \mathrm{e}^{\theta f - \theta^2 \frac{c}{4}}$ car (i) découle de [13]

$$\Big(\int \mathrm{e}^{\theta f - \theta^2 \frac{c}{4}} \mathrm{d}\mu\Big) \log \int \mathrm{e}^{\theta f - \theta^2 \frac{c}{4}} \mathrm{d}\mu \le 0.$$

À présent, par homogénéité, l'analyse peut se réduire au cas où h est une densité de probabilité par rapport à μ, et comme f est de moyenne nulle pour μ, la propriété entropique précédente se reformule en

$$\int (fh - f) \mathrm{d}\mu \le \frac{c}{4}\theta + \frac{1}{\theta} \int h \log(h) \mathrm{d}\mu$$

En prenant l'infimum sur $\theta > 0$ on obtient

$$\int (fh - f) \mathrm{d}\mu \le \sqrt{c \int h \log(h) \mathrm{d}\mu}.$$

En considérant $\mathrm{d}\nu := h\mathrm{d}\mu$ puis le supremum sur f, on obtient, grâce à la dualité de Kantorovich–Rubinstein du théorème 2.5.3, l'inégalité (ii). Il ne reste plus qu'à observer que l'argument est réversible. □

Remarque 2.5.6. Convexité, transformée de Legendre, entropie relative, log-Laplace.

Considérons la formule variationnelle de l'entropie relative du lemme 2.3.1, pour une fonction $f \ge 0$ vérifiant $\int f \mathrm{d}\mu = 1$. En remplaçant g tel que $\int \mathrm{e}^g \mathrm{d}\mu = 1$ par $g - \log \int \mathrm{e}^g \mathrm{d}\mu$ pour un g arbitraire on voit que Ent_μ est la transformée de Legendre de la log-Laplace (voir la section 3.2) au sens où

$$\mathrm{Ent}_\mu(f) = \sup_g \Big\{ \int fg \mathrm{d}\mu - \log \int \mathrm{e}^g \mathrm{d}\mu \Big\}.$$

Réciproquement (dualité convexe) la log-Laplace est la transformée de Legendre de l'entropie relative :

$$\sup_{f \ge 0, \int f \mathrm{d}\mu = 1} \Big\{ \int fg \mathrm{d}\mu - \mathrm{Ent}_\mu(f) \Big\} = \log \int \mathrm{e}^g \mathrm{d}\mu.$$

En fait, la formule variationnelle de l'entropie relative du lemme 2.3.1 est équivalente à la convexité de la fonctionnelle $f \ge 0 \mapsto \mathrm{Ent}_\mu(f)$ via sa représentation comme enveloppe de ses tangentes affines :

$$\begin{aligned}\mathrm{Ent}_\mu(f) &= \sup_{g \ge 0} \Big\{ \int \Big(\log(g) - \log\Big(\int g \mathrm{d}\mu \Big) \Big)(f - g) \mathrm{d}\mu + \mathrm{Ent}_\mu(g) \Big\} \\ &= \sup_{g \ge 0} \Big\{ \int \Big(\log(g) - \log\Big(\int g \mathrm{d}\mu \Big) \Big) f \mathrm{d}\mu \Big\} = \sup_{g \ge 0, \int g \mathrm{d}\mu = 1} \Big\{ \int f \log(g) \mathrm{d}\mu \Big\} = \sup_{\int \mathrm{e}^g \mathrm{d}\mu = 1} \Big\{ \int fg \mathrm{d}\mu \Big\}.\end{aligned}$$

Pour aller plus loin, on peut consulter par exemple [52, 53, 16].

13. Pour tout $u \ge 0$, $u \log(u) \le 0$ ssi $u \in [0,1]$.

Chapitre 3

Principe de grandes déviations de Cramér

Me 19/02

Abordé en cours :
— Concept de grandes déviations par rapport au comportement typique
— Analyse asymptotique exacte du cas gaussien (cas Bernoulli non abordé)
— Transfo de Legendre (sans preuve) et lemme sur transfo de Cramér (avec l'essentiel de la preuve)
— Exemples de transformée de Cramér (Bernoulli et Gaussienne)
— Inégalité de Cramér–Chernoff et l'essentiel de la preuve sous moment exponentiel.

3.1 Concept de grandes déviations, cas Bernoulli et gaussien

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées, intégrables de moyenne m. Soit $S_n := X_1 + \cdots + X_n$. La LGN (faible) indique que

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{\text{loi}} m$$

c'est-à-dire que la loi μ_n de $\frac{S_n}{n}$ converge étroitement vers δ_m quand $n \to \infty$. Pour tout borélien $B \subset \mathbb{R}$:

— si $m \in \mathring{B}$, alors il existe [1] $\varphi : \mathbb{R} \to [0,1]$ continue telle que $\mathbb{1}_B \geq \varphi$ et $\varphi(m) = 1$, et donc

$$\mathbb{P}\Big(\frac{S_n}{n} \in B\Big) \geq \int \varphi \mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \varphi(m) = 1,$$

— si $m \notin \overline{B}$, alors il existe $\varphi : \mathbb{R} \to [0,1]$ continue telle que $\mathbb{1}_B \leq \varphi$ et $\varphi(m) = 0$, et donc

$$\mathbb{P}\Big(\frac{S_n}{n} \in B\Big) \leq \int \varphi \mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \varphi(m) = 0.$$

Pour la loi de $\frac{S_n}{n}$, l'événement B constitue une grande déviation par rapport à la valeur typique m. Alternativement, on retrouve la topologie du problème avec la LGN forte qui donne $\lim_{n\to\infty} \mathbb{1}_{\frac{S_n}{n}\in B} = \mathbb{1}_{m\in B}$ presque sûrement lorsque m est un point de continuité de $\mathbb{1}_B$, d'où, par convergence dominée,

$$\lim_{n\to\infty} \mathbb{P}(\tfrac{S_n}{n} \in B) = \lim_{n\to\infty} \mathbb{E}(\mathbb{1}_{\frac{S_n}{n}\in B}) = \mathbb{E}(\lim_{n\to\infty} \mathbb{1}_{\frac{S_n}{n}\in B}) = \mathbb{1}_{m\in B}.$$

Lorsque $B = [x, +\infty)$, le principe de grandes déviations (PGD) de Cramér du théorème 3.4.1 précise le comportement asymptotique de $\mathbb{P}(\frac{S_n}{n} \in B) = \mathbb{P}(\frac{S_n}{n} \geq x) = \mathbb{P}(\frac{S_n}{n} \geq m + r)$ en fournissant un contrôle du type

$$\frac{1}{n}\log\Big(\frac{S_n}{n} \geq x\Big) \underset{n\to\infty}{\approx} -I(x) \quad \text{c'est-à-dire} \quad \mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) \underset{n\to\infty}{\approx} c_n(x)\mathrm{e}^{-nI(x)},$$

où $I(x) \in [0, +\infty]$, I atteint son minimum 0 en $x = m$, et où $c_n(x)$ est écrasée : $\lim_{n\to\infty} \frac{1}{n}\log c_n(x) = 0$. Ce type d'analyse asymptotique à l'échelle exponentielle ou logarithmique peut différer de celle issue du TLC :

$$\mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) \approx \mathbb{P}(Z \geq \sqrt{n}(x-m)) = \mathrm{erf}\Big(\sqrt{\frac{n}{2}}(x-m)\Big), \quad Z \sim \mathcal{N}(0,1),$$

et chacune des deux approximations peut avoir son régime d'optimalité, y compris quand $X_1 \sim$ Bernoulli.

Une approche pour établir le PGD est d'obtenir d'abord une majoration de $\mu_n([x,\infty)) = \mathbb{P}(\frac{S_n}{n} \geq x)$ via une inégalité de Markov exponentielle, ce qui fait intervenir la transformée de Laplace

$$L(\lambda) := \mathbb{E}(\mathrm{e}^{\lambda X_1}).$$

1. Comme $m \notin \overline{B}$ est équivalent à $m \in$ intérieur(B^c), et $\mathbb{1}_{B^c} = 1 - \mathbb{1}_B$, ce cas est équivalent au cas précédent, en remplaçant φ par $1 - \varphi$.

Pour obtenir une minoration du même ordre, on considère la loi de densité $y \mapsto e^{\lambda y}/L(\lambda)$ par rapport à la loi de X_1, de moyenne x pour λ bien choisi, et on utilise la LGN (faible) pour cette loi. La valeur x, qui est une déviation par rapport à m pour la loi de X_1, devient une valeur typique de la loi déformée exponentiellement.

Historiquement, le premier principe de grandes déviations a été obtenu vers 1938 par Harald Cramér, mathématicien, statisticien, et actuaire[2]. Il était inspiré notamment à la fois par l'inégalité de Markov exponentielle pour la majoration et, pour la minoration, par la technique de la transformation exponentielle introduite par l'actuaire Frederik Esscher pour l'obtention d'une estimée plus fine que celle produite par le TLC dans certains régimes. Les techniques pour obtenir et exploiter les principes de grandes déviations ont été conceptualisées et développées pendant la seconde moitié du vingtième siècle, par de nombreux mathématiciens, notamment par Srinivasa Varadhan, dont les travaux ont été récompensés par le prix Abel en 2007. Les PGD jouent aujourd'hui un rôle important en probabilités, mécanique statistique, et physique statistique.

Dans les cas Bernoulli et gaussien, la loi de S_n est explicite, et on peut procéder par estimation directe.

Théorème 3.1.1. Grandes déviations par rapport à la moyenne dans le cas Bernoulli symétrique.

Si $(X_n)_{n\geq 1}$ sont i.i.d. Bernoulli $\mathbb{P}(X_1=0)=\mathbb{P}(X_1=1)=1/2$, $S_n := X_1+\cdots+X_n$, alors pour tout $x \geq 1/2$,

$$\lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}\Big(\frac{S_n}{n}\geq x\Big) = -I(x) \quad \text{où} \quad I(x) := \begin{cases} \log(2)+x\log(x)+(1-x)\log(1-x) & \text{si } x\in[0,1] \\ +\infty & \text{sinon} \end{cases},$$

et par symétrie, pour tout $r\geq 0$, $\lim_{n\to\infty} \frac{1}{n}\log\mathbb{P}\Big(\Big|\frac{S_n}{n}-\frac{1}{2}\Big|\geq r\Big) = -I\Big(\frac{1}{2}+r\Big)$.

— Idem pour $\mathbb{P}\big(\frac{S_n}{n}\leq x\big)$ pour $x\leq 1/2$, en notant que $I(x)=I(1-x)$.
— La fonction I est symétrique par rapport à l'axe vertical d'abscisse $1/2$, infinie sur $[0,1]^c$, prend la valeur $\log(2)$ en 0 et 1, strictement convexe sur $[0,1]$, avec un unique minimum global 0 atteint en $1/2$.
— On en déduit que pour $\varepsilon>0$ petit, $\sum_n \mathbb{P}(|\frac{1}{n}S_n-\frac{1}{2}|>\varepsilon)<\infty$, d'où la LGN forte $\lim_{n\to\infty}\frac{1}{n}S_n = 1/2$ p.s. Le point critique de I en $1/2$ correspond à la LGN. La dérivée seconde de I en $1/2$ correspond au TLC.

Démonstration. Soit $x\in(1/2,1]$ (sinon il n'y a rien à démontrer). On a alors $\mathbb{P}(S_n\geq nx) = 2^{-n}\sum_{k\geq nx}\binom{n}{k}$, d'où

$$2^{-n}Q_n(x) \leq \mathbb{P}(S_n\geq nx) \leq (n+1)2^{-n}Q_n(x) \quad \text{où} \quad Q_n(x) := \max_{nx\leq k\leq n}\binom{n}{k}.$$

Ce maximum est atteint pour $k=\lceil nx\rceil$, le plus petit entier $\geq nx$. À présent, la formule de Stirling dans sa version semi-quantitative $n! = \sqrt{2\pi n}n^n e^{-n}(1+O(1/n))$ donne alors $\frac{1}{n}\log Q_n(x) \sim -x\log(x)-(1-x)\log(1-x)$. Enfin les bornes inférieures et supérieures de l'encadrement se confondent à l'échelle exponentielle quand $n\to\infty$. On retrouve ici l'entropie comme analyse asymptotique de la combinatoire (remarque 1.2.8 avec $r=2$).

Enfin, nous avons $\mathbb{P}(|\frac{S_n}{n}-\frac{1}{2}|\geq r) = \mathbb{P}(\frac{S_n}{n}\geq\frac{1}{2}+r)+\mathbb{P}(\frac{S_n}{n}\leq\frac{1}{2}-r) = 2\mathbb{P}(\frac{S_n}{n}\geq\frac{1}{2}+r)$ car les deux événements sont disjoints, et de même probabilité par symétrie de la loi de $\frac{S_n}{n}-\frac{1}{2}$. Le préfacteur 2 est écrasé par $\frac{1}{n}\log$. □

Théorème 3.1.2. Grandes déviations par rapport à la moyenne dans le cas gaussien.

Si $(X_n)_{n\geq 1}$ est une suite de v.a.r. i.i.d. de loi $\mathcal{N}(m,\sigma^2)$, $m\in\mathbb{R}$, $\sigma>0$, alors pour tout $r>0$,

$$\mathbb{P}\Big(\Big|\frac{S_n}{n}-m\Big|\geq r\Big) \underset{n\to\infty}{\sim} \frac{\sqrt{2\sigma^2}}{\sqrt{\pi n}r}e^{-\frac{nr^2}{2\sigma^2}}, \quad \text{d'où} \quad \lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\frac{S_n}{n}\in(m-r,m+r)^c\Big) = -\frac{r^2}{2\sigma^2},$$

et par symétrie, pour tout $x\geq 0$, $\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\frac{S_n}{n}\geq x\Big) = -\frac{(x-m)^2}{2\sigma^2}$.

— La première expression est plus précise en raison du préfacteur devant l'exponentielle. Dans la seconde, le $\frac{1}{n}\log$ écrase le préfacteur et ne retient que le terme d'ordre n à l'échelle exponentielle.
— La valeur typique de $\frac{S_n}{n}$ est m, avec des fluctuations de l'ordre de $\frac{\sigma}{\sqrt{n}}$, tandis qu'une valeur déviante de m de r se produit avec une probabilité très petite de l'ordre de $\exp(-\frac{r^2}{2\sigma^2}n+o(n))$.
— Dans ce cas gaussien, l'approximation coïncide avec celle fournie par le TLC.

2. Professionnel spécialiste de l'application du calcul des probabilités et de la statistique aux questions d'assurances, de prévention, de comptabilité et analyse financière associée, et de prévoyance sociale. Fixer le prix de polices d'assurance est un art quantitatif de l'aléatoire.

Démonstration. Comme $\frac{S_n}{n} - m \sim \mathcal{N}(0, \frac{\sigma^2}{n})$, on a, en posant $u = n(x-r)$ pour la dernière égalité,

$$\mathbb{P}\Big(\Big|\frac{S_n}{n} - m\Big| \geq r\Big) = 2\mathbb{P}\Big(\frac{S_n}{n} - m \geq r\Big) = 2\frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}}\int_r^\infty \mathrm{e}^{-\frac{nx^2}{2\sigma^2}}\,\mathrm{d}x = \frac{2\mathrm{e}^{-\frac{nr^2}{2\sigma^2}}}{\sqrt{2\pi n\sigma^2}}\int_0^\infty \mathrm{e}^{-\frac{u^2}{2n\sigma^2}-\frac{ru}{\sigma^2}}\,\mathrm{d}u.$$

Or par convergence dominée, $\lim_{n\to\infty}\int_0^\infty \mathrm{e}^{-\frac{u^2}{2n\sigma^2}-\frac{ru}{\sigma^2}}\,\mathrm{d}u = \int_0^\infty \mathrm{e}^{-\frac{ru}{\sigma^2}}\,\mathrm{d}u = \frac{\sigma^2}{r}$, d'où le résultat.

□

3.2 Transformée de Cramér : transformée de Legendre de la log-Laplace

Examinons le cas où X_1 a des moments exponentiels de tout ordre, c'est-à-dire que $L(\lambda) := \mathbb{E}(\mathrm{e}^{\lambda X_1}) < \infty$ pour tout $\lambda \in \mathbb{R}$, et où $B = [x, +\infty)$, $x \in \mathbb{R}$. On a alors, pour tout $\lambda > 0$, par l'inégalité de Markov,

$$\mathbb{P}\Big(\frac{S_n}{n} \in B\Big) = \mathbb{P}(S_n \geq nx) = \mathbb{P}(\mathrm{e}^{\lambda S_n} \geq \mathrm{e}^{n\lambda x}) \leq \mathrm{e}^{-n\lambda x + \log\mathbb{E}(\mathrm{e}^{\lambda S_n})} = \mathrm{e}^{-n\lambda x + n\log\mathbb{E}(\mathrm{e}^{\lambda X_1})} \leq \mathrm{e}^{-n\sup_{\lambda>0}(\lambda x - \log\mathbb{E}(\mathrm{e}^{\lambda X_1}))}.$$

Nous voyons poindre une transformée de la log-Laplace qui ressemble à la transformée de Legendre. Rappelons que si $\Phi : \mathbb{R} \to (-\infty, +\infty]$, $\Phi \not\equiv +\infty$, alors sa transformée de Legendre[3] $\Phi^* : \mathbb{R} \to (-\infty, +\infty]$ est définie par

$$\Phi^*(x) := \sup_{\lambda\in\mathbb{R}}(\lambda x - \Phi(\lambda)) = \sup_{\Phi(\lambda)<+\infty}(\lambda x - \Phi(\lambda)).$$

$\boldsymbol{\Phi(\lambda)}$	$\{\boldsymbol{\Phi} < +\infty\}$	$\boldsymbol{\Phi^*(x)}$	$\{\boldsymbol{\Phi^*} < +\infty\}$
$\frac{1}{p}\lvert\lambda\rvert^p,\ 1 < p < \infty$	$\mathbb{R}$	$\frac{1}{q}\lvert x\rvert^q, \frac{1}{p}+\frac{1}{q}=1$	$\mathbb{R}$
$\lvert\lambda\rvert$	$\mathbb{R}$	0	$[-1,1]$
$\Phi(\lambda) = \mathrm{e}^\lambda$	$\mathbb{R}$	$x\log(x) - x$	$[0,+\infty)$

TABLE 3.1 – Quelques exemples de transformées de Legendre de fonctions convexes.

La transformée de Legendre est fondamentale en analyse non-linéaire, en raison de son affinité avec la convexité. Elle est tout aussi importante en physique, notamment en thermodynamique : l'énergie libre de Helmholtz est une transformée de Legendre. Au passage, la remarque 2.5.6 incite à interpréter la transformée de Legendre de la log-Laplace comme une entropie relative, idem dans le chapitre suivant (théorème de Sanov).

Lemme 3.2.1. Quelques propriétés de la transformée de Legendre.

(i) La transformée de Legendre Φ^* est convexe et semi-continue inférieurement (s.c.i.).

(ii) Si Φ est convexe, dérivable deux fois sur un intervalle ouvert $I \subset \{\Phi < +\infty\}$ avec Φ' injective, alors Φ^* est dérivable sur $\Phi'(I)$ et $(\Phi^*)' = (\Phi')^{-1}$.

(iii) Inégalité de Young ou de Fenchel : $\Phi(\lambda) + \Phi^*(x) \geq \lambda x$ pour tous λ et x.

Dessin !

— Le (ii) indique que le graphe de $(\Phi^*)'$ s'obtient en retournant la feuille sur laquelle est tracé celui de Φ'.
— Dans (ii), sur I, la fonction Φ' est croissante car Φ est convexe, et comme elle est supposée injective, il vient qu'elle est strictement croissante, donc Φ est strictement convexe[4] sur I.
— Si $\Phi^* \not\equiv +\infty$, alors $\Phi^{**} : \mathbb{R} \to (-\infty, +\infty]$ et l'inégalité de Young donne $\Phi \geq \Phi^{**}$. La fonction Φ^{**} est une convexification de Φ. Comme le suggère le (ii), lorsque Φ est convexe, alors $\Phi = \Phi^{**}$, et Φ et Φ^* sont transformées de Legendre l'une de l'autre : fonctions convexes conjuguées et dualité convexe.

Démonstration.

3. Ou transformée de Fenchel–Legendre. Au-delà de $\mathbb{R}$, si $\Phi : H \to (-\infty, +\infty]$ avec H Hilbert réel alors $\Phi^*(x) := \sup_{\lambda\in H}(\langle\lambda, x\rangle - \Phi(\lambda))$, voire si $\Phi : E \to (-\infty, +\infty]$ avec E espace vectoriel topologique réel alors $\Phi^*(x) := \sup_{\lambda\in E'}(\lambda(x) - \Phi(\lambda))$. Dualité pour la dualité.

4. La convexité stricte n'implique pas la dérivabilité, comme le montre par exemple $x \mapsto x^2 + |x|$ en $x = 0$.

(i) Φ^* est convexe car enveloppe d'une famille de fonctions affines : pour tous $x_1, x_2 \in \mathbb{R}$ et tout $\theta \in (0,1)$,

$$\begin{aligned}\theta\Phi^*(x_1) + (1-\theta)\Phi^*(x_2) &= \sup_{\lambda\in\mathbb{R}}(\theta\lambda x_1 - \theta\Phi(\lambda)) + \sup_{\lambda\in\mathbb{R}}((1-\theta)\lambda x_2 - (1-\theta)\Phi(\lambda)) \\ &\geq \sup_{\lambda\in\mathbb{R}}((\theta x_1 + (1-\theta)x_2)\lambda - \Phi(\lambda)) = \Phi^*(\theta x_1 + (1-\theta)x_2).\end{aligned}$$

(ii) Si $x_n \to x$ dans $\mathbb{R}$ alors pour tout $\lambda \in \mathbb{R}$,

$$\underline{\lim}_{n\to\infty} \Phi^*(x_n) \geq \underline{\lim}_{n\to\infty} (\lambda x_n - \Phi(\lambda)) = \lambda x - \Phi(\lambda), \quad \text{d'où} \quad \underline{\lim}_{n\to\infty} \Phi^*(x_n) \geq \sup_{\lambda\in\mathbb{R}}(\lambda x - \Phi(\lambda)) = \Phi^*(x).$$

(iii) Les hypothèses assurent que $\lambda \mapsto \lambda x - \Phi(\lambda)$ est concave et atteint son maximum [5] en $\lambda = (\Phi')^{-1}(x)$, ce qui donne la formule de Legendre $\Phi^*(x) = x(\Phi')^{-1}(x) - \Phi((\Phi')^{-1}(x))$. Il en découle que Φ^* est dérivable et

$$(\Phi^*)'(x) = (\Phi')^{-1}(x) + x((\Phi')^{-1})'(x) - \Phi'((\Phi')^{-1}(x))((\Phi')^{-1})'(x) = (\Phi')^{-1}(x).$$

(iv) On peut supposer sans perte de généralité que $\Phi(\lambda) < +\infty$. Or pour tout $x \in \mathbb{R}$ et tout λ tel que $\Phi(\lambda) < +\infty$, la définition de $\Phi^*(x)$ entraîne que $\Phi(\lambda) + \Phi^*(x) \geq \Phi(\lambda) + \lambda x - \Phi(\lambda) = \lambda x$.

□

La transformée de Cramér Λ^* d'une variable aléaroire réelle X est la transformée de Legendre du logarithme de sa transformée de Laplace (appelée log-Laplace pour faire court). Elle ne dépend que de la loi de X.

$$L(\lambda) := \mathbb{E}(e^{\lambda X}) \in (0, +\infty], \quad \Lambda(\lambda) := \log L(\lambda) \in (-\infty, +\infty], \quad \Lambda^*(x) := \sup_{\lambda\in\mathbb{R}}(\lambda x - \Lambda(\lambda)) \in (-\infty, +\infty].$$

Notons que $L(0) = 1$ donc $\Lambda(0) = 0$ et donc $\Lambda \not\equiv +\infty$, ce qui permet de définir Λ^*.

Lemme 3.2.2. Transformée de Cramér : transformée de Legendre de la log-Laplace.

Soit X une variable aléatoire réelle, $\Lambda = \log L$ sa log-Laplace, et Λ^* sa transformée de Cramér. Alors :

(i) Λ et Λ^* sont convexes, Λ^* est semi-continue inférieurement, et $\Lambda^* \geq 0$.

(ii) Si $\{\Lambda < +\infty\} = \{0\}$ (c'est-à-dire que X n'a aucun moment exponentiel) alors $\Lambda^* \equiv 0$.
Si $\Lambda(\lambda) < \infty$ pour un $\lambda > 0$, alors $\mathbb{E}(X) \in [-\infty, +\infty)$, et pour tout $x \geq \mathbb{E}(X)$,

$$\Lambda^*(x) = \sup_{\lambda\geq 0}(\lambda x - \Lambda(\lambda)), \quad \text{et cette fonction est croissante pour } x > \mathbb{E}(X).$$

Si $\Lambda(\lambda) < \infty$ pour un $\lambda < 0$, alors $\mathbb{E}(X) \in (-\infty, +\infty]$, et pour tout $x \leq \mathbb{E}(X)$,

$$\Lambda^*(x) = \sup_{\lambda\leq 0}(\lambda x - \Lambda(\lambda)), \quad \text{et cette fonction est décroissante pour } x < \mathbb{E}(X).$$

Dans chacun de ces deux cas, si X est intégrable, alors $\Lambda^*(\mathbb{E}(X)) = 0$.
De plus nous avons toujours $\inf_{\mathbb{R}} \Lambda^* = 0$.

(iii) Λ est deux fois dérivable sur l'intérieur de $\{\Lambda < +\infty\}$ et pour tout λ dans cet ensemble on a

$$\Lambda'(\lambda) = \frac{\mathbb{E}(Xe^{\lambda X})}{L(\lambda)} \quad \text{et} \quad \Lambda''(\lambda) = \frac{\mathbb{E}(X^2e^{\lambda X}) - \mathbb{E}(Xe^{\lambda X})^2}{L(\lambda)^2}.$$

En particulier, si $\{\Lambda < +\infty\}$ contient un voisinage de 0 alors

$$\Lambda'(0) = \mathbb{E}(X) \quad \text{et} \quad \Lambda''(0) = \mathrm{Var}(X).$$

(iv) Comparaison au cas gaussien : si $\{\Lambda < +\infty\}$ contient un voisinage de 0 sur lequel $\Lambda'' > 0$, alors

$$\Lambda^*(\mathbb{E}(X)) = 0, \quad (\Lambda^*)'(\mathbb{E}(X)) = 0, \quad (\Lambda^*)''(\mathbb{E}(X)) = \frac{1}{\mathrm{Var}(X)}.$$

Dessin !

— Λ^* est la transformée de Legendre de la log-Laplace Λ. On dit que c'est la transformée de Cramér de X.
— La transformée de Laplace L est la fonction génératrice des moments : $L^{(n)}(0) = \mathbb{E}(X^n)$, $n \geq 0$.

5. On utilise ici le fait qu'une fonction convexe dérivable en un point avec une dérivée nulle admet ce point comme minimum global.

— La log-Laplace $\Lambda = \log L$ est la fonction génératrice des cumulants, notamment les formules de (iii).
— On a toujours $0 \in \{\Lambda < +\infty\}$ car $L(0) = 1$. Dire que $\{\Lambda < +\infty\}$ contient un voisinage de 0 signifie que X possède des moments exponentiels des deux signes sur λ, et dans ce cas $X \in \cap_{p\geq 1} L^p$.

Démonstration.

(i) La convexité de Λ découle de l'inégalité de Hölder et de la croissance du logarithme : pour tout $\theta \in [0,1]$,

$$\Lambda(\theta\lambda_1 + (1-\theta)\lambda_2) = \log\mathbb{E}((\mathrm{e}^{\lambda_1 X})^\theta (\mathrm{e}^{\lambda_2 X})^{1-\theta}) \leq \log(\mathbb{E}(\mathrm{e}^{\lambda_1 X})^\theta \mathbb{E}(\mathrm{e}^{\lambda_2 X})^{1-\theta}) = \theta\Lambda(\lambda_1) + (1-\theta)\Lambda(\lambda_2).$$

En particulier $\{\Lambda < +\infty\}$ est convexe. La convexité et la semi-continuité inférieure de Λ^* viennent du fait que c'est une transformée de Legendre (lemme 3.2.1). Enfin $\Lambda(0) = \log\mathbb{E}(1) = 0$, d'où $\Lambda^*(x) \geq 0x - \Lambda(0) = 0$.

(ii) Si $\{\Lambda < +\infty\} = \{0\}$ alors $\Lambda^*(x) = \sup_{\lambda\in\mathbb{R}}(\lambda x - \Lambda(\lambda)) = -\Lambda(0) = 0$ pour tout $x \in \mathbb{R}$.
Si $\Lambda(\lambda) = \log L(\lambda) < \infty$ pour un $\lambda > 0$, alors l'inégalité $\lambda \max(x,0) \leq \mathrm{e}^{\lambda x}$ donne

$$\mathbb{E}(X_+) \leq \frac{L(\lambda)}{\lambda} < \infty, \quad \text{donc } \mathbb{E}(X) = \mathbb{E}(X_+) - \mathbb{E}(X_-) \text{ fait sens dans } [-\infty, +\infty).$$

De plus, pour tout $\lambda \in \mathbb{R}$, si $\mathbb{E}(X)$ est fini, alors l'inégalité de Jensen donne

$$\Lambda(\lambda) = \log\mathbb{E}(\mathrm{e}^{\lambda X}) \geq \mathbb{E}(\log \mathrm{e}^{\lambda X}) = \lambda\mathbb{E}(X)$$

donc $\Lambda^*(\mathbb{E}(X)) \leq 0$ donc $= 0$ car $\Lambda^* \geq 0$, et de plus, pour tout $x \geq \mathbb{E}(X)$ et tout $\lambda < 0$,

$$\lambda x - \Lambda(\lambda) \leq \lambda\mathbb{E}(X) - \Lambda(\lambda) \leq \Lambda^*(\mathbb{E}(X)) = 0,$$

d'où $\Lambda^*(x) = \sup_{\lambda\geq 0}(\lambda x - \Lambda(\lambda))$. Cette formule est immédiate quand $\mathbb{E}(X) = -\infty$ car dans ce cas l'inégalité de Jensen précédente donne $\Lambda(\lambda) = +\infty$ pour tout $\lambda < 0$. Cette formule implique également la croissance de Λ^* sur $(\mathbb{E}(X), +\infty)$ comme supremum de fonctions croissantes.
Si $\Lambda(\lambda) < \infty$ pour un $\lambda < 0$, on raisonne symétriquement avec $-X$.
Il reste à établir que $\inf_{\mathbb{R}} \Lambda^* = 0$. Nous le savons déjà si $\{\Lambda < +\infty\} = \{0\}$, ainsi que si $\mathbb{E}(X)$ est fini avec un moment exponentiel fini, et dans ce cas, comme nous venons de le voir, $\Lambda^*(\mathbb{E}(X)) = 0$. Considérons le cas où $\mathbb{E}(X) = -\infty$ et $\Lambda(\lambda) < \infty$ pour un $\lambda > 0$. Alors par l'inégalité de Markov et un résultat précédent,

$$\log\mathbb{P}(X \geq x) \leq \inf_{\lambda\geq 0} \log\mathbb{E}(\mathrm{e}^{\lambda(X-x)}) = -\sup_{\lambda\geq 0}(\lambda x - \Lambda(\lambda)) = -\Lambda^*(x),$$

donc

$$\lim_{x\to-\infty} \Lambda^*(x) \leq \lim_{x\to-\infty} (-\log\mathbb{P}(X \geq x)) = 0,$$

d'où $\inf_{\mathbb{R}} \Lambda^* = 0$. Le cas où $\mathbb{E}(X) = +\infty$ et $\Lambda(\lambda) < \infty$ pour un $\lambda < 0$ se traite symétriquement avec $-X$.

(iii) La formule pour $\Lambda'(\lambda)$ découle d'une dérivation sous le signe intégral, licite par convergence dominée car

$$f_\varepsilon(x) := \frac{\mathrm{e}^{(\lambda+\varepsilon)x} - \mathrm{e}^{\lambda x}}{\varepsilon} \xrightarrow[\varepsilon\to 0]{} x\mathrm{e}^{\lambda x} \quad \text{et} \quad |f_\varepsilon(x)| \leq \mathrm{e}^{\lambda x} \frac{\mathrm{e}^{\delta|x|} - 1}{\delta} =: h_\delta(x) \quad \text{pour tout } \varepsilon \in (-\delta, \delta),$$

tandis que $\mathbb{E}(|h(X)|) < \infty$ pour $\delta > 0$ assez petit car λ est dans l'intérieur de $\{\Lambda < +\infty\}$. Idem pour $\Lambda''(\lambda)$.

(iv) De (ii) $\Lambda^*(\mathbb{E}(X)) = 0$. Le (iii) donne $\Lambda'(0) = \mathbb{E}(X)$ et $\Lambda''(0) = \mathrm{Var}(X) > 0$. Comme $\Lambda'' > 0$, le lemme 3.2.1 (ii) donne $(\Lambda^*)' = (\Lambda')^{-1}$, d'où $(\Lambda')^{-1}(\mathbb{E}(X)) = 0$, $(\Lambda^*)'' = ((\Lambda')^{-1})' = \frac{1}{(\Lambda''((\Lambda')^{-1})}$, et $(\Lambda^*)''(\mathbb{E}(X)) = \frac{1}{\Lambda''(0)} = \frac{1}{\mathrm{Var}(X)}$.

□

3.3 Inégalité de Cramér–Chernoff

Théorème 3.3.1. Inégalité de Cramér–Chernoff.

Si $S_n := X_1 + \cdots + X_n$ avec $X_1, \ldots, X_n$ v.a.r i.i.d. de transformée de Cramér Λ^*, alors pour tout fermé $F \subset \mathbb{R}$,

$$\mathbb{P}\Big(\frac{S_n}{n} \in F\Big) \leq 2\exp\Big(-n\inf_F \Lambda^*\Big) \quad \text{en particulier} \quad \overline{\lim_{n\to\infty}} \frac{1}{n}\log\mathbb{P}\Big(\frac{S_n}{n} \in F\Big) \leq -\inf_F \Lambda^*.$$

— Autrement dit, en notant μ_n la loi de $\frac{S_n}{n}$, pour tout fermé $F \subset \mathbb{R}$, $\mu_n(F) \leq 2\mathrm{e}^{-n\inf_F \Lambda^*}$.

Loi	$\Lambda(\lambda)$	$\Lambda^*(x)$
$(1-p)\delta_0 + p\delta_1,\ 0<p<1$ Bernoulli	$\log(pe^\lambda + (1-p))$	$\begin{cases} x\log\frac{x}{p} + (1-x)\log\frac{1-x}{1-p} & \text{si } x\in[0,1] \\ +\infty & \text{sinon} \end{cases}$
$e^{-\theta}\sum_{n=0}^{\infty}\frac{\theta^n}{n!}\delta_n,\ \theta>0$ Poisson	$\theta(e^\lambda - 1)$	$\begin{cases} \theta - x + x\log\frac{x}{\theta} & \text{si } x\geq 0 \\ +\infty & \text{sinon} \end{cases}$
$\theta e^{-\theta x}\mathbb{1}_{x\geq 0}dx,\ \theta>0$ Exponentielle	$\begin{cases} \log\frac{\theta}{\theta-\lambda} & \text{si } \lambda<\theta \\ +\infty & \text{sinon} \end{cases}$	$\begin{cases} \theta x - 1 - \log(\theta x) & \text{si } x>0 \\ +\infty & \text{sinon} \end{cases}$
$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}}dx,\ m\in\mathbb{R},\ \sigma>0$ Gaussienne	$\lambda m + \frac{\sigma^2}{2}\lambda^2$	$\frac{(x-m)^2}{2\sigma^2}$

TABLE 3.2 – Quelques exemples de transformées de Cramér. La fonction Λ peut prendre la valeur $+\infty$ pour des lois non exotiques comme la loi exponentielle par exemple, ce qui explique l'intérêt d'inclure cette possibilité. Par ailleurs, la loi de densité $f(x) = c(|x|^{-\alpha}\mathbf{1}_{x<0} + e^{-x^\beta}\mathbf{1}_{x\geq 0})$, avec $1<\alpha\leq 2$, $\beta>1$, et $c>0$ constante de normalisation, vérifie $\mathbb{E}(X_+)<\infty$, $\mathbb{E}(X)=-\infty$, $\Lambda(\lambda)<+\infty$ pour tout $\lambda\geq 0$, et $\Lambda(\lambda)=+\infty$ pour tout $\lambda<0$. Cet exemple donne sens au degré de généralité considéré dans le lemme 3.2.2. Enfin la loi de Bernoulli est un exemple de loi dont la fonction Λ est finie partout tandis que la fonction Λ^* peut prendre la valeur $+\infty$.

— On ne suppose pas que les variables sont exponentiellement intégrables ni intégrables. Si les variables ne possèdent aucun moment exponentiel, autrement dit si $\{\Lambda<+\infty\}=\{0\}$, alors par le lemme 3.2.2 (ii) on a $\Lambda^*\equiv 0$, et la borne $\mathbb{P}\leq 2$ fournie par le théorème est vraie (mais sans intérêt).

— Si les variables possèdent une moyenne $m\in\mathbb{R}$, et si pour tout $\varepsilon>0$, en posant $F := B(m,\varepsilon)^c$, on a $\inf_F \Lambda^* > 0$, alors par le lemme 1.1.1 de Borel–Cantelli (première partie), il vient $\sum_n \mathbb{P}(\frac{S_n}{n}\in F)<\infty$, d'où $\lim_{n\to\infty}\frac{S_n}{n} = m$ presque sûrement, et il s'agit d'une convergence complète au sens du théorème 1.1.2. Ainsi l'inégalité de Cramér–Chernoff peut entraîner une LGN forte.

— La preuve révèle que le préfacteur 2 peut être remplacé par 1 si $F=(-\infty, m-r]$ ou $F=[m+r,+\infty)$, $r\geq 0$.

— Sous les hypothèses du lemme 3.2.2 (iv) on a $\Lambda^*(m+x) = \frac{x^2}{2\sigma^2} + o_{x\to 0}(x^2)$, d'où un lien entre PGD et TLC :

$$\mathbb{P}\Big(\sqrt{n}\Big(\frac{S_n}{n}-m\Big)\geq r\Big) = \mathbb{P}\Big(\frac{S_n}{n}\geq m+\frac{r}{\sqrt{n}}\Big) \leq e^{-n\Lambda^*\left(m+\frac{r}{\sqrt{n}}\right)} = e^{-\frac{r^2}{2\sigma^2}+o_{n\to\infty}(1)}.$$

En ce sens, PGD et TLC sont équivalents pour les petites déviations par rapport à la moyenne.
En revanche, pour les grandes déviations, l'estimée par PGD peut être meilleure que celle issue du TLC.
La figure 3.1 permet de comparer à l'échelle exponentielle l'approximation par PGD et par TLC.

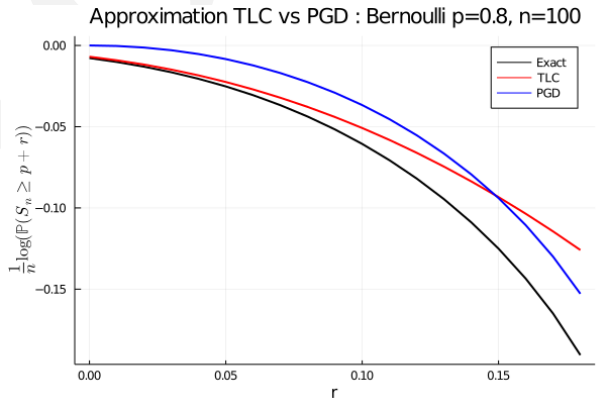


FIGURE 3.1 – Approximation par PGD versus TLC dans le cas Bernoulli.

```
using LaTeXStrings, Printf, Plots, Distributions # Julia
n = 100 ; p = .8 ; q = 1-p ; dx = .01 ; r = collect(0:dx:1-p-dx) ; x = r.+p
exa = Distributions.ccdf.(Binomial(n,p),n*x)
tlc = Distributions.ccdf.(Normal(n*p,sqrt(n*p*q)),n*x)
pgd = exp.(-n.*(x.*log.(x./p)+(1 .-x).*log.((1 .-x)./q)))
plot(r,log.(exa)./n,color=[:black],lw=2,label="Exact")
plot!(r,log.(tlc)./n,color=[:red],lw=2,label="TLC")
plot!(r,log.(pgd)./n,color=[:blue],lw=2,label="PGD")
title!(@sprintf("Approximation TLC vs PGD : Bernoulli p=%2.1f, n=%d",p,n))
xlabel!("r") ; ylabel!(L"\frac{1}{n}\log(\mathbb{P}(S_n\geq p+r))") ; savefig("pgd-tlc-ber.png") ; gui()
```

Dessin !

Démonstration. Soit $F \subset \mathbb{R}$ un fermé tel que $\inf_F \Lambda^* > 0$ (sinon il n'y a rien à démontrer). Par le lemme 3.2.2 (ii), $m := \mathbb{E}(X_1)$ a un sens (éventuellement infini). Pour tous x et $\lambda \geq 0$, par l'inégalité de Markov et l'indépendance,

$$\mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) = \mathbb{P}\Big(\lambda S_n \geq \lambda x n\Big) = \mathbb{P}\Big(\mathrm{e}^{\lambda S_n} \geq \mathrm{e}^{n\lambda x}\Big) \leq \mathrm{e}^{-n\lambda x}\mathbb{E}(\mathrm{e}^{\lambda S_n}) = \mathrm{e}^{-n(\lambda x - \Lambda(\lambda))}.$$

Si $m < +\infty$ alors par le lemme 3.2.2 (ii), pour tout $x > m$,

$$\mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) \leq \mathrm{e}^{-n\Lambda^*(x)}.$$

Similairement, si $m > -\infty$ alors pour tout $x < m$, en notant que $\Lambda_{-X_1}(\lambda) = \Lambda(-\lambda)$ et $\Lambda^*_{-X_1}(-x) = \Lambda^*(x)$,

$$\mathbb{P}\Big(\frac{S_n}{n} \leq x\Big) = \mathbb{P}\Big(\frac{-S_n}{n} \geq -x\Big) \leq \mathrm{e}^{-n\Lambda^*_{-X_1}(-x)} = \mathrm{e}^{-n\Lambda^*(x)}.$$

Considérons le cas où m est fini. Alors $\Lambda^*(m) = 0$ par le lemme 3.2.2, et comme $\inf_F \Lambda^* > 0$, il vient que m est dans l'ouvert $F^c = \mathbb{R} \setminus F$. Soit (m_-, m_+) le plus grand intervalle ouvert (obtenu par réunion) contenant m et inclus dans F^c. On a forcément $m_- < m_+$, et l'une de ces deux bornes doit être finie car F est non-vide. Si m_- est fini alors $m_- \in F$ car F est fermé, et donc $\Lambda^*(m_-) \geq \inf_F \Lambda^*$. De même si m_+ est fini alors $m_+ \in F$, et donc $\Lambda^*(m_+) \geq \inf_F \Lambda^*$. En utilisant les bornes exponentielles précédentes avec $x = m_-$ et $x = m_+$ on obtient

$$\mathbb{P}\Big(\frac{S_n}{n} \in F\Big) \leq \mathbb{P}\Big(\frac{S_n}{n} \leq m_-\Big) + \mathbb{P}\Big(\frac{S_n}{n} \geq m_+\Big) \leq 2\mathrm{e}^{-n\inf_F \Lambda^*}.$$

Considérons le cas $m = -\infty$. Comme la fonction Λ^* est croissante et $\inf_{\mathbb{R}} \Lambda^* = 0$ par le lemme 3.2.2 (ii), on a $\lim_{x \to -\infty} \Lambda^*(x) = 0$, et donc m_+ est fini sinon on aurait $\inf_F \Lambda^* = 0$. Comme F est fermé, on a $m_+ \in F$ et donc $\Lambda^*(m_+) \geq \inf_F \Lambda^*$. De plus $F \subset [m_+, \infty)$, d'où, cette fois-ci avec $x = m_+$,

$$\mathbb{P}\Big(\frac{S_n}{n} \in F\Big) \leq \mathbb{P}\Big(\frac{S_n}{n} \geq m_+\Big) \leq \mathrm{e}^{-n\Lambda^*(m_+)} \leq \mathrm{e}^{-n\inf_F \Lambda^*}.$$

Le cas $m = +\infty$ se traite symétriquement de la même manière. □

3.4 Théorème de Cramér dans $\mathbb{R}$

Je 20/02

Abordé en cours :
— Théorème de Cramér et sa preuve dans le cas réduit.
— Théorème sur les propriétés de la fonction de taux.
— Théorème de Laplace–Varadhan et sa preuve (lemme d'analyse sans sa preuve)
— Notion générale de PGD de Varadhan.

Théorème 3.4.1. de Cramér dans $\mathbb{R}$.

Soient $(X_n)_{n\geq 1}$ des v.a.r. i.i.d. de transformée de Cramér Λ^*, et $S_n := X_1 + \cdots + X_n$. Alors :

(i) Pour tout ouvert $O \subset \mathbb{R}$, on a $\underline{\lim}_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \in O\Big) \geq -\inf_O \Lambda^*$.

(ii) Pour tout fermé $F \subset \mathbb{R}$, on a $\overline{\lim}_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \in F\Big) \leq -\inf_F \Lambda^*$.

Autrement dit, en notant μ_n la loi de $\frac{S_n}{n}$, pour tout borélien $B \subset \mathbb{R}$,

$$-\inf_{\mathring{B}} \Lambda^* \leq \underline{\lim}_{n\to\infty} \frac{1}{n} \log \mu_n(B) \leq \overline{\lim}_{n\to\infty} \frac{1}{n} \log \mu_n(B) \leq -\inf_{\overline{B}} \Lambda^*.$$

— Analyse asymptotique à l'échelle exponentielle a_n de la convergence $\mu_n \xrightarrow[n\to\infty]{\mathscr{C}_b} \delta_m$ (LGN). On dit que $(\mu_n)_{n\geq 1}$ satisfait à un principe de grandes déviations (PGD) de vitesse $a_n = n$ et de fonction de taux Λ^*.
— On ne suppose pas que les variables sont exponentiellement intégrables ou intégrables. En revanche, et comme déjà mentionné, le résultat est vide si les variables n'ont aucun moment exponentiel.
— De même que pour l'inégalité de Cramér–Chernoff du théorème 3.3.1, la borne supérieure du PGD de Cramér peut entraîner une LGN forte via le lemme de Borel–Cantelli, cf. TD.

Démonstration.

(i) Découle de l'inégalité de Cramér–Chernoff du théorème 3.3.1 : le préfacteur 2 est écrasé par $\frac{1}{n}\log$.

(ii) Supposons dans un premier temps que la loi μ des X_i vérifie $\mu((-\infty,0))>0$, $\mu((0,+\infty))>0$, et que μ est à support borné. Il en découle que $\lim_{|\lambda|\to\infty}\Lambda(\lambda)=\infty$, et que $\Lambda(\lambda)$ est fini pour tout $\lambda\in\mathbb{R}$. Par le lemme 3.2.2 (iii), Λ est donc dérivable, et il existe η tel que

$$\Lambda(\eta)=\inf_{\lambda\in\mathbb{R}}\Lambda(\lambda)\quad\text{et}\quad\Lambda'(\eta)=0.$$

Soit $\widetilde{\mu}$ la mesure de probabilité sur $\mathbb{R}$ définie par

$$\widetilde{\mu}(\mathrm{d}x)=\frac{\mathrm{e}^{\eta x}}{L(\eta)}\mu(\mathrm{d}x)=\mathrm{e}^{\eta x-\Lambda(\eta)}\mu(\mathrm{d}x).$$

On parle de transformation d'Esscher, de famille exponentielle, de tilt exponentiel. Par le lemme 3.2.2 (iii),

$$\int x\widetilde{\mu}(\mathrm{d}x)=\frac{1}{L(\eta)}\int x\mathrm{e}^{\eta x}\mu(\mathrm{d}x)=\Lambda'(\eta)=0.$$

Soit à présent μ_n la loi de $\frac{S_n}{n}$. Alors pour tout $\varepsilon>0$,

$$\begin{aligned}\mu_n((-\varepsilon,\varepsilon))&=\int\mathbb{1}_{|x_1+\cdots+x_n|<n\varepsilon}\mu(\mathrm{d}x_1)\cdots\mu(\mathrm{d}x_n)\\&\geq\mathrm{e}^{-n\varepsilon|\eta|}\int\mathbb{1}_{|x_1+\cdots+x_n|<n\varepsilon}\mathrm{e}^{\eta(x_1+\cdots+x_n)}\mu(\mathrm{d}x_1)\cdots\mu(\mathrm{d}x_n)\\&=\mathrm{e}^{-n\varepsilon|\eta|}\mathrm{e}^{n\Lambda(\eta)}\widetilde{\mu}_n((-\varepsilon,\varepsilon)),\end{aligned}$$

où $\widetilde{\mu}_n$ est la loi de la moyenne empirique de n v.a.r. i.i.d. de loi $\widetilde{\mu}$.
L'intervalle $(-\varepsilon,\varepsilon)$, pas forcément typique pour μ_n, est typique pour $\widetilde{\mu}_n$ qui est de moyenne nulle.
La LGN (faible) donne $\lim_{n\to\infty}\widetilde{\mu}((-\varepsilon,\varepsilon))=1$, d'où, pour tous $0<\varepsilon<\delta$,

$$\varliminf_{n\to\infty}\frac{1}{n}\log\mu_n((-\delta,\delta))\geq\varliminf_{n\to\infty}\frac{1}{n}\log\mu_n((-\varepsilon,\varepsilon))\geq\Lambda(\eta)-\varepsilon|\eta|.$$

En prenant la limite quand $\varepsilon\to0$, il vient

$$\varliminf_{n\to\infty}\frac{1}{n}\log\mu_n((-\delta,\delta))\geq\Lambda(\eta)\geq\inf_{\lambda\in\mathbb{R}}\Lambda(\lambda)=-\Lambda^*(0).$$

Comme $\Lambda_{X-x}(\lambda)=\Lambda_X(\lambda)-\lambda x$ et $\Lambda^*_{X-x}(y)=\Lambda^*(y+x)$, il vient par translation que pour tout x et tout $\delta>0$,

$$\varliminf_{n\to\infty}\frac{1}{n}\log\mu_n((x-\delta,x+\delta))\geq-\Lambda^*(x),$$

ce qui conduit, par emboîtement, à la borne inférieure souhaitée pour tout ouvert $O\subset\mathbb{R}$. Notons que ces deux derniers arguments de translation et d'emboîtement ne font pas appel aux hypothèses faites sur μ.
Supposons à présent que μ vérifie $\mu((-\infty,0))>0$ et $\mu((0,+\infty))>0$, mais que μ n'est pas à support borné. Procédons par approximation par troncature. Soit M assez grand pour que $\mu([-M,0))>0$ et $\mu((0,M])>0$. Soit ν la mesure de probabilité sur $\mathbb{R}$ définie par $\nu(B)=\mu(B\cap[-M,M])/\mu([-M,M])$ pour tout borélien $B\subset\mathbb{R}$. Soit ν_n la loi de la moyenne empirique de n variables i.i.d. de loi ν. Pour tout n et tout $\delta>0$,

$$\mu_n((-\delta,\delta))\geq\nu_n((-\delta,\delta))\mu([-M,M])^n.$$

Comme la log-Laplace de ν est $\Lambda_M-\log\mu([-M,M])$ où $\Lambda_M(\lambda):=\log\int_{-M}^{M}\mathrm{e}^{\lambda x}\mu(\mathrm{d}x)$, le résultat pour ν donne

$$\varliminf_{n\to\infty}\frac{1}{n}\log\mu_n((-\delta,\delta))\geq\log\mu([-M,M])+\varliminf_{n\to\infty}\frac{1}{n}\log\nu_n((-\delta,\delta))\geq\inf_{\lambda\in\mathbb{R}}\Lambda_M(\lambda).$$

En notant $I_M:=-\inf_{\mathbb{R}}\Lambda_M$ et $I^*:=\limsup_{M\to\infty}I_M$, il vient

$$\varliminf_{n\to\infty}\frac{1}{n}\log\mu_n((-\delta,\delta))\geq-I^*.$$

Comme Λ_M est croissante en M, il en va de même de $-I_M$. De plus $-I_M \le \Lambda_M(0) \le \Lambda(0) = 0$, et donc $-I^* \le 0$. À présent, comme $-I_M$ est bornée pour M assez grand, on a $-I^* > -\infty$, donc les ensembles de sous-niveau $K_M := \{\lambda : \Lambda_M(\lambda) \le -I^*\}$ sont non-vides. Comme ils sont compacts et emboîtés, leur intersection $\cap_M K_M$ est non-vide, et contient donc au moins un point λ_*. Par convergence monotone,

$$\Lambda(\lambda_*) = \lim_{M\to\infty} \Lambda_M(\lambda_*) \le -I^*,$$

et ceci conduit au résultat pour $O = (-\delta, \delta)$, et ensuite pour tout ouvert O par translation et emboîtement. Supposons enfin que μ vérifie $\mu((-\infty, 0)) = 0$ ou $\mu((0, +\infty)) = 0$. Dans ce cas Λ est une fonction monotone et $\inf_{\mathbb{R}} \Lambda = \log \mu(\{0\})$, et le résultat provient alors du fait que $\mu_n((-\delta, \delta)) \ge \mu_n(\{0\}) = \mu(\{0\})^n$.

□

Théorème 3.4.2. Variante raffinée du théorème de Cramér dans $\mathbb{R}$.

Sous les hypothèses et avec les notations du théorème 3.4.1 de Cramér dans $\mathbb{R}$, pour tout $y \in \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \ge y\Big) = -\inf_{x\ge y} \Lambda^*(x).$$

De plus si $\{\Lambda < +\infty\}$ contient un voisinage de 0 alors $\inf_{x\ge y} \Lambda^*(x) = \Lambda^*(y)$.

Démonstration. Soit μ_n la loi de $\frac{S_n}{n}$. Comme $[x, x+\delta) \subset [y, +\infty)$ pour tout $x \ge y$ et tout $\delta > 0$, il vient

$$\underline{\lim}_{n\to\infty} \frac{1}{n} \log \mu_n([y, +\infty)) \ge \sup_{x\ge y} \underline{\lim}_{n\to\infty} \frac{1}{n} \log \mu_n([x, x+\delta)).$$

Le résultat découle alors du renforcement suivant d'une inégalité dans la preuve du théorème 3.4.1 de Cramér :

$$\underline{\lim}_{n\to\infty} \frac{1}{n} \log \mu_n([x, x+\delta)) \ge -\Lambda^*(x).$$

Comme dans la preuve du théorème 3.4.1 de Cramér, on se ramène par translation au cas $x = 0$, et on procède avec $[0, \delta)$ et $[0, \varepsilon)$ en lieu et place de $(-\delta, \delta)$ et $(-\varepsilon, \varepsilon)$ dans toutes les étapes de la preuve de l'inégalité dans la preuve du théorème 3.4.1. Le point crucial est le remplacement de la LGN faible par le TLC, qui donne

$$\lim_{n\to\infty} \widetilde{\mu}_n([0, \varepsilon)) = \frac{1}{2}.$$

Enfin la propriété sous moments exponentiels vient de la monotonie fournie par le lemme 3.2.2 (ii). □

Théorème 3.4.3. Propriétés de la fonction de taux.

(i) Si $\{\Lambda < +\infty\}$ contient un voisinage de 0 (existence de moments exponentiels avec $\lambda < 0$ et $\lambda > 0$) alors les ensembles de sous-niveau de Λ^* sont compacts.

(ii) Si $\{\Lambda < +\infty\} = \mathbb{R}$ alors $\lim_{|x|\to\infty} \dfrac{\Lambda^*(x)}{|x|} = \infty$.

— Le (i) fait qu'on qualifie Λ^* de bonne fonction de taux pour le PGD de Cramér du théorème 3.4.1.
— Sous certaines hypothèses, la fonction de taux est $\mathscr{C}^\infty$ et strictement convexe, cf. TD.

Démonstration.

(i) Comme il existe $\lambda_- < 0$ et $\lambda_+ > 0$ dans $\{\Lambda < +\infty\}$, et comme la définition de $\Lambda^*(x)$ donne, pour tout $\lambda \in \mathbb{R}$,

$$\frac{\Lambda^*(x)}{|x|} \ge \lambda \operatorname{signe}(x) - \frac{\Lambda(\lambda)}{|x|}, \quad \text{on obtient} \quad \underline{\lim}_{|x|\to\infty} \frac{\Lambda^*(x)}{|x|} \ge \min(\lambda_+, -\lambda_-) > 0.$$

En particulier $\lim_{|x|\to\infty} \Lambda^*(x) = \infty$, les ensembles de sous-niveau $\{x \in \mathbb{R} : \Lambda^*(x) \le r\}$, $r \ge \inf_{\mathbb{R}} \Lambda^*$, de Λ^* sont bornés. Comme ils sont fermés car Λ^* est s.c.i. d'après le lemme 3.2.2 (i), ils sont compacts.

(ii) Comme $\{\Lambda < +\infty\} = \mathbb{R}$, on peut prendre $\lambda_+ = -\lambda_- \to \infty$ dans la preuve du (i) d'où le résultat.

□

3.5 Méthode de Laplace et lemme de Varadhan

La méthode de Laplace, vue en TD, affirme que si $\varphi : \mathbb{R} \to \mathbb{R}$ est continue et bornée supérieurement, et si μ est une mesure de probabilité sur $\mathbb{R}$ chargeant tout intervalle non vide, alors

$$\lim_{n\to\infty} \frac{1}{n} \log \int \mathrm{e}^{n\varphi(x)} \mu(\mathrm{d}x) = \sup_{\mathbb{R}} \varphi.$$

En effet, d'une part $\int \mathrm{e}^{n\varphi(x)}\mu(\mathrm{d}x) \le \mathrm{e}^{n \sup_{\mathbb{R}} \varphi}$, et d'autre part, comme φ est continue, pour tous $x \in \mathbb{R}$ et $\varepsilon > 0$, il existe un voisinage V_x de x tel que $\varphi(y) \ge \varphi(x) - \varepsilon$ pour tout $y \in V_x$, d'où $\int \mathrm{e}^{n\varphi(x)}\mu(\mathrm{d}x) \ge \mathrm{e}^{n(\varphi(x)-\varepsilon)}\mu(V_x) > 0$.

Théorème 3.5.1. ou lemme de Laplace–Varadhan.

Soit $(Z_n)_{n\ge1}$ une suite de variables réelles, et $I : \mathbb{R} \to [0, +\infty]$ à ensembles de sous-niveau compacts, telles que pour tout borélien $B \subset \mathbb{R}$,

$$-\inf_{\mathring{B}} I \le \underline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}(Z_n \in B) \le \overline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{P}(Z_n \in B) \le -\inf_{\overline{B}} I.$$

Alors pour toute fonction $\varphi : \mathbb{R} \to \mathbb{R}$ continue et bornée supérieurement,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}(\mathrm{e}^{n\varphi(Z_n)}) = \sup_{\mathbb{R}} (\varphi - I).$$

— Dans le cas où les Z_n sont de même loi μ, la propriété sur les Z_n a lieu avec $I \equiv 0$ pour les boréliens B vérifiant $\mu(B) > 0$, et le résultat coïncide avec la méthode de Laplace mentionnée plus haut dans le sens où $\mathbb{E}(\mathrm{e}^{n\varphi(Z_n)}) = \int \mathrm{e}^{n\varphi(x)}\mu(\mathrm{d}x)$. La coïncidence est formelle car I n'est pas à ensembles de sous-niveau compacts. D'autre part, ce n'est pas une instance de Cramér : Z_n n'est pas une moyenne empirique.

— Si $(X_n)_{n\ge1}$ sont des v.a.r. i.i.d. de log-Laplace Λ et de transformée de Cramér Λ^* et si $\{\Lambda < +\infty\}$ contient un voisinage de 0, alors, en combinant le théorème 3.4.1 de Cramér et le théorème 3.4.3 sur Λ^*, les hypothèses du théorème de Laplace–Varadhan sont vérifiées par $Z_n = \frac{S_n}{n}$ et $I = \Lambda^*$, et on obtient

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}(\mathrm{e}^{n\varphi(\frac{S_n}{n})}) = \sup_{\mathbb{R}} (\varphi - \Lambda^*).$$

— Un résultat analogue s'obtient en remplaçant partout la vitesse n par β_n.

Démonstration. On procède en établissant la minoration et la majoration suivantes :

$$\underline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{E}(\mathrm{e}^{n\varphi(Z_n)}) \ge \sup_{\mathbb{R}} (\varphi - I) \quad \text{et} \quad \overline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{E}(\mathrm{e}^{n\varphi(Z_n)}) \le \sup_{\mathbb{R}} (\varphi - I).$$

Preuve de la minoration. On procède par localisation par restriction et monotonie. Pour tout $x \in \mathbb{R}$, comme φ est continue en x, pour tout $\varepsilon > 0$ il existe un voisinage ouvert V_x de x tel que $\inf_{V_x} \varphi \ge \varphi(x) - \varepsilon$, d'où

$$\mathbb{E}(\mathrm{e}^{n\varphi(Z_n)}) \ge \mathbb{E}(\mathrm{e}^{n\varphi(Z_n)} \mathbb{1}_{Z_n \in V_x}) \ge \mathrm{e}^{n(\varphi(x)-\varepsilon)} \mathbb{P}(Z_n \in V_x),$$

et en utilisant la borne inférieure de la propriété sur les Z_n, il vient alors

$$\underline{\lim_{n\to\infty}} \frac{1}{n} \log \mathbb{E}(\mathrm{e}^{n\varphi(Z_n)}) \ge \varphi(x) - \varepsilon - \inf_{V_x} I \ge \varphi(x) - \varepsilon - I(x).$$

Ceci étant vrai pour tout $x \in \mathbb{R}$ et tout $\varepsilon > 0$, on obtient la minoration souhaitée.

Preuve de la majoration. Comme φ est bornée supérieurement : $m := \sup_{\mathbb{R}} \varphi < +\infty$. D'autre part, comme les ensembles de sous-niveau de I sont compacts, ils sont fermés et I est semi-continue inférieurement (s.c.i.).

Comme φ est continue et I est s.c.i., pour tous $x \in \mathbb{R}$ et $\varepsilon > 0$, il existe un voisinage ouvert V_x de x tel que

$$\sup_{V_x} \varphi \le \varphi(x) + \varepsilon \quad \text{et} \quad \inf_{\overline{V_x}} I \ge I(x) - \varepsilon.$$

Localisons par compacité. Pour tout $r > \inf I$, l'ensemble de sous-niveau $K_r := \{x : I(x) \le r\}$ est compact par hypothèse. On peut donc extraire de $K_r \subset \cup_{x\in K} V_x$ un sous-recouvrement fini $K \subset V := \cup_{i=1}^N V_{x_i}$. On a alors

$$\mathbb{E}(\mathrm{e}^{n\varphi(Z_n)}) = \mathbb{E}(\mathrm{e}^{n\varphi(Z_n)} \mathbb{1}_{Z_n \notin V}) + \mathbb{E}(\mathrm{e}^{n\varphi(Z_n)} \mathbb{1}_{Z_n \in V}) \le \mathbb{E}(\mathrm{e}^{n\varphi(Z_n)} \mathbb{1}_{Z_n \notin V}) + \sum_{i=1}^N \mathrm{e}^{n(\varphi(x_i)+\varepsilon)} \mathbb{P}(Z_n \in V_{x_i}).$$

En utilisant le lemme d'analyse ci-dessous puis la borne supérieure de la propriété sur les Z_n, on obtient

$$\begin{aligned}\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{E}(e^{n\varphi(Z_n)}) &\le \max\Big(m+\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{P}(Z_n\in V^c),\max_{1\le i\le N}(\varphi(x_i)+\varepsilon+\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{P}(Z_n\in V_{x_i}))\Big)\\ &\le \max\Big(m-\inf_{V^c}I,\max_{1\le i\le N}(\varphi(x_i)+\varepsilon-\inf_{\overline{V_{x_i}}}I)\Big).\end{aligned}$$

Or $\inf_{V^c} I \ge \inf_{K_r^c} I \ge r$ par définition de V et K_r, tandis que $\inf_{\overline{V_{x_i}}} I \ge I(x_i)-\varepsilon$ par définition de V_{x_i}, d'où

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{E}(e^{n\varphi(Z_n)}) \le \max\Big(m-r,\max_{1\le i\le N}(\varphi(x_i)-I(x_i)+2\varepsilon)\Big) \le \max\Big(m-r,\sup_{\mathbb{R}}(\varphi-I)+2\varepsilon\Big).$$

La majoration souhaitée s'obtient en faisant $r\to+\infty$ et $\varepsilon\to 0$. Note : les x_i et N dépendent de ε et r.

Lemme d'analyse. Si (a_n) et (b_n) sont des suites de réels strictement positifs, alors

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\log(a_n+b_n) = \max\Big(\overline{\lim_{n\to\infty}}\frac{1}{n}\log(a_n),\overline{\lim_{n\to\infty}}\frac{1}{n}\log(b_n)\Big).$$

En effet, la croissante de log et la positivité de a_n et b_n donne

$$\max\Big(\overline{\lim_{n\to\infty}}\frac{1}{n}\log(a_n),\overline{\lim_{n\to\infty}}\frac{1}{n}\log(b_n)\Big) \le \overline{\lim_{n\to\infty}}\frac{1}{n}\log(a_n+b_n),$$

tandis que d'autre part,

$$\frac{1}{n}\log(a_n+b_n) \le \max\Big(\frac{1}{n}\log(2a_n),\frac{1}{n}\log(2b_n)\Big) \le \frac{\log(2)}{n}+\max\Big(\frac{1}{n}\log(a_n),\frac{1}{n}\log(b_n)\Big),$$

d'où

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\log(a_n+b_n) \le \max\Big(\overline{\lim_{n\to\infty}}\frac{1}{n}\log(a_n),\overline{\lim_{n\to\infty}}\frac{1}{n}\log(b_n)\Big).$$

□

— Dans le théorème 3.5.1 de Laplace–Varadhan, on peut remplacer l'hypothèse de bornitude sur φ par

$$\lim_{r\to\infty}\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{E}(e^{n\varphi(Z_n)}\mathbb{1}_{\varphi(Z_n)\ge r}) = -\infty.$$

En effet, en introduisant la troncature $\varphi_r := \min(\varphi, r)$, on obtient

$$\mathbb{E}(e^{n\varphi(Z_n)}) = \mathbb{E}(e^{n\varphi(Z_n)}\mathbb{1}_{\varphi(Z_n)<r}) + \mathbb{E}(e^{n\varphi(Z_n)}\mathbb{1}_{\varphi(Z_n)\ge r})$$

Par le lemme d'analyse utilisé dans la preuve précédente, on obtient

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{E}(e^{n\varphi(Z_n)}) = \max\Big(\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{E}(e^{n\varphi(Z_n)}\mathbb{1}_{\varphi(Z_n)<r}),\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{E}(e^{n\varphi(Z_n)}\mathbb{1}_{\varphi(Z_n)\ge r})\Big).$$

Or en utilisant le théorème pour la fonction bornée supérieurement φ_r, il vient

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{E}(e^{n\varphi(Z_n)}\mathbb{1}_{\varphi(Z_n)<r}) \le \overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{E}(e^{n\varphi_r(Z_n)}) \le \sup_{\mathbb{R}}(\varphi_r-I) \le \sup_{\mathbb{R}}(\varphi-I),$$

et comme $\lim_{r\to\infty}\overline{\lim}_{n\to\infty}\frac{1}{n}\log\mathbb{E}(e^{n\varphi(Z_n)}\mathbb{1}_{\varphi(Z_n)\ge r}) = -\infty$, on obtient $\overline{\lim_{n\to\infty}}\frac{1}{n}\log\mathbb{E}(e^{n\varphi(Z_n)}) \le \sup_{\mathbb{R}}(\varphi-I)$.

3.6 À propos des principes de grandes déviations

Soit $(\mu_n)_{n\ge1}$ une famille de mesures de probabilités sur un espace topologique E muni de sa tribu borélienne. On dit qu'elle satisfait à un principe de grandes déviations (PGD) de vitesse $(\beta_n)_{n\ge1}$ et de fonction de taux $I : E\to[0,\infty]$ lorsque les trois conditions suivantes sont réalisées :

(i) $\beta_n \nearrow +\infty$ quand $n\to\infty$

(ii) Les ensembles de sous-niveau de I sont compacts (donc fermés c'est-à-dire que I est s.c.i.)

(iii) Pour tout borélien $B\subset E$, $-\inf_{\mathring{B}} I \le \underline{\lim}_{n\to\infty}\frac{1}{\beta_n}\log\mu_n(B) \le \overline{\lim}_{n\to\infty}\frac{1}{\beta_n}\log\mu_n(B) \le -\inf_{\overline{B}} I$.

La fonction de taux est unique et atteint son minimum global, égal à 0.

— Si $\mathring{B}$ contient les minimiseurs de I alors $\lim_{n\to\infty} \mu_n(B) = 1$.

— Si $\overline{B}$ ne continent pas de minimiseur de I alors $\lim_{n\to\infty} \mu_n(B) = 0$.

Cela permet de concevoir approximativement μ_n comme une mesure de Boltzmann – Gibbs : $\mu_n(B) \approx \mathrm{e}^{-n \inf_B I}$. En pratique, dans les modèles, n est le nombre de particules, ou la dimension du système. À l'opposé, pour la mesure de Boltzmann – Gibbs d'un modèle, un PGD est une approximation énergétique en grande dimension.

Ce concept général de PGD a été introduit par Varadhan vers 1966. Le lemme de Laplace–Varadhan date de la même époque. Il affirme la convergence faible de μ_n quand $n \to \infty$ vers une loi portée par les minimiseurs de I, cette loi étant une Dirac lorsque I a un unique minimiseur comme dans le théorème de Cramér. Varadhan a développé ensuite amplement cette thématique notamment avec Donsker aux États-Unis, tandis que Freidlin et Wentzell suivaient un chemin parallèle mais indépendant en Union soviétique.

L'argument pour la borne supérieure utilisé dans la preuve du théorème 3.4.1 de Cramér sur $\mathbb{R}$ est unidimensionnel. Il est possible de le remplacer par un argument de réduction aux boules via compacité associé à une tension exponentielle. Ceci conduit notamment au théorème de Cramér sur $\mathbb{R}^d$, avec

$$\Lambda^*(x) := \sup_{\lambda\in\mathbb{R}^d} (\langle \lambda, x\rangle - \Lambda(\lambda)) \quad \text{et} \quad \Lambda(x) := \log L(\lambda) := \log \mathbb{E}(\mathrm{e}^{\langle \lambda, X_1\rangle}),$$

cf. [30, th. 2.2.30]. Le théorème de Cramér a été obtenu par Cramér vers 1938 sous des hypothèses restrictives, levées par Chernoff en 1952. Entre 1965 et 1975, Ruelle puis Lanford ont proposé une preuve des théorèmes de Cramér basée sur la sous-additivité, sans faire appel à une déformation exponentielle, on la trouvera dans [30, ch. 6] et [21, ch. 7]. Une preuve élémentaire du théorème de Cramér dans $\mathbb{R}$ basée sur la sous-additivité se trouve dans [22]. Donsker et Varadhan ont généralisé le théorème de Cramér à un Banach en 1976 et Bahadur et Zabell l'ont ensuite généralisé en 1979 à des espaces vectoriels topologiques localement convexes séparés. L'étude de théorèmes de Cramér de dimension infinie est revisitée dans [21] et [61].

Le théorème de Cramér possède également une généralisation à des variables aléatoires dépendantes, qui porte le nom de théorème de Gärtner–Ellis (1977, 1984). Des principes de grande déviation pour des mesures de Boltzmann–Gibbs sont étudiés dans [66], et pour des processus de Markov en lien avec l'inégalité de Sobolev logarithmique dans [32]. Un principe de grandes déviations est disponible pour le mouvement Brownien en temps petit (théorème de Schilder, 1962), pour la formule de Feynmann–Kac associée à un opérateur de diffusion et son équation de Schrödinger réelle (théorème de Donsker–Varadhan, 1983), pour des systèmes dynamiques perturbés aléatoirement (théorème de Freidlin–Wentzell, 1970).

Un exposé sur les PGD à la fois accessible et axé sur les modèles se trouve dans [31].

On trouvera dans [42] et [17], voir aussi [35], un PGD pour le modèle de Curie–Weiss pour lequel une transition a lieu entre un régime où le minimiseur de la fonction de taux n'est pas unique et un régime où la fonction de taux est strictement convexe, en rapport avec un TLC non-standard à la valeur critique.

Le point de vue d'un physicien sur les PGD pour la mécanique statistique est exposé dans [75].

Le développement asymptotique précis du cas gaussien fourni par le théorème 3.1.2 possède une généralisation au-delà du cas gaussien (théorème de Bahadur–Rao, 1960), un raffinement du théorème de Cramér.

Il est possible de déduire un TLC à partir d'un PGD, la variance asymptotique étant donnée par la dérivée seconde de la fonction de taux en le minimiseur qui est la moyenne, cf. [19].

Il est possible de déduire une LGN forte à partir du PGD de Cramér via le lemme de Borel–Cantelli, cf. TD.

Le PGD de Cramér est relié au PGD de Sanov qui fait l'objet du chapitre suivant.

Ce chapitre est localement à la fois librement et fortement inspiré de [30], [76], et [31].

Chapitre 4

Principe de grandes déviations de Sanov

Me 05/03

Abordé en cours :
— Lemme et théorème de Sanov discrets, lien avec analyse asymptotique du coefficient multinomial
— Cramér discret et sa preuve par contraction
— Énoncé du principe de contraction

Le théorème de Sanov a été obtenu par Ivan Nikolaevich Sanov vers 1957, mais il semble que beaucoup n'y ont pas cru. Nous étudions en détail la version pour des variables aléatoires discrètes finies, ce qui correspond au jeu de dé, et fait émerger très naturellement l'entropie. Nous en déduisons le théorème de Cramér pour des variables aléatoires du même type par une instance du principe de contraction. Nous présentons ensuite brièvement la version générale, explorée par Donsker et Varadhan et Bahadur et Zabell dans les années 1970.

4.1 Théorème de Sanov pour le jeu de dé

Soit $A = \{a_1, \dots, a_r\}$ un ensemble fini de cardinal r, interprétable comme un alphabet, muni de la topologie et de la tribu discrètes. L'ensemble $\mathscr{P}(A)$ des mesures de probabilités sur A est identifiable au simplexe

$$\{(p_1, \dots, p_r) \in \mathbb{R}_+^r : p_1 + \cdots + p_r = 1\},$$

et on note $\mu_i := \mu(\{a_i\})$ si $\mu \in \mathscr{P}(A)$. Toute fonction $f : A \to \mathbb{R}$ est continue et bornée. En identifiant aussi bien les fonctions que les mesures à des vecteurs de $\mathbb{R}^r$, on a $\langle f, \mu \rangle = \sum_{i=1}^r f_i \mu_i = \int f \mathrm{d}\mu$. La topologie induite sur $\mathscr{P}(A)$ par celle de $\mathbb{R}^r$ coïncide avec la topologie de la convergence étroite. Pour tout $\mu \in \mathscr{P}(A)$ on note $\mathrm{supp}(\mu) := \{a_i : \mu_i > 0\}$. Si $(X_n)_{n\geq 1}$ sont des variables aléatoires à valeurs dans A et si $\mu \in \mathscr{P}(A)$ alors

$$X_n \xrightarrow[n\to\infty]{\text{loi}} \mu \quad \text{ssi} \quad \lim_{n\to\infty} \mathbb{P}(X_n = x) = \mu(x) \quad \text{pour tout } x \in A.$$

À toute suite [1] $(x_1, \dots, x_n) \in A^n$ est associée une mesure de probabilité $L_n(x_1, \dots, x_n) \in \mathscr{P}(A)$ définie par

$$L_n(x_1, \dots, x_n) := \Big(\frac{1}{n} \sum_{k=1}^n \mathbb{1}_{x_k = a_1}, \dots, \frac{1}{n} \sum_{k=1}^n \mathbb{1}_{x_k = a_r} \Big).$$

Les éléments de $\mathscr{P}(A)$ réalisables de la sorte sont $L_n(A^n) := \{L_n(x_1, \dots, x_n) : (x_1, \dots, x_n) \in A^n\} \subset \mathscr{P}(A)$.

Si $X_1, \dots, X_n$ sont des variables aléatoires i.i.d. à valeur dans A modélisant, par exemple, les n lancers d'un dé à r faces, alors $L_n(X_1, \dots, X_n)$ est une mesure de probabilité aléatoire sur A, v.a. à valeurs dans $\mathscr{P}(A)$, appelée mesure empirique, constituée du vecteur des fréquences empiriques de chacune des faces $a_1, \dots, a_r$ dans l'échantillon $X_1, \dots, X_n$. En notant $\mu \in \mathscr{P}(A)$ la loi des X_i, la LGN donne (presque sûrement et dans L^1)

$$L_n(X_1, \dots, X_n) \xrightarrow[n\to\infty]{} (\mu_1, \dots, \mu_r).$$

Le lemme suivant souligne notamment que de simples considérations combinatoires, volumiques, font surgir l'entropie, en liaison avec la mesure empirique d'un échantillon, de manière non-asymptotique.

Lemme 4.1.1. Combinatoire du jeu de dé et entropie.

(i) D'une part $|L_n(A^n)| \leq (n+1)^{r-1}$, et d'autre part $\inf_{\nu' \in L_n(A^n)} \|\nu - \nu'\|_\infty \leq \frac{r-1}{n}$ pour tout $\nu \in \mathscr{P}(A)$ et pour n assez grand, et cela reste vrai si on impose de plus que $\mathrm{supp}(\nu') \subset \mathrm{supp}(\nu)$.

1. Du point de vue de la théorie de l'information, ces x_i sont les lettres successives d'un message écrit dans l'alphabet A.

(ii) Si $X_1,\dots,X_n$ sont des v.a. i.i.d. de loi $\mu\in\mathscr{P}(A)$, alors pour tous $\nu\in L_n(A^n)$ et $(x_1,\dots,x_n)\in L_n^{-1}(\nu)$,

$$\mathbb{P}((X_1,\dots,X_n)=(x_1,\dots,x_n))=\mathrm{e}^{-n(\mathrm{S}(\nu)+\mathrm{H}(\nu|\mu))},$$

où $\mathrm{S}(\nu)$ et $\mathrm{H}(\nu\mid\mu)$ sont respectivement l'entropie de Boltzmann–Shannon de ν et l'entropie relative ou information de Kullback–Leibler de ν par rapport à μ définies par

$$\mathrm{S}(\nu):=-\sum_{i=1}^r \nu_i\log\nu_i \quad\text{et}\quad \mathrm{H}(\nu\mid\mu):=\sum_{i=1}^r \nu_i\log\frac{\nu_i}{\mu_i}.$$

(iii) Pour tout $\nu\in L_n(A^n)$, on a $(n+1)^{-(r-1)}\mathrm{e}^{n\mathrm{S}(\nu)}\le|L_n^{-1}(\nu)|=\dfrac{n!}{(n\nu_1)!\cdots(n\nu_r)!}\le\mathrm{e}^{n\mathrm{S}(\nu)}$.

— On adopte la convention $0\log(0)=0$ et $\mathrm{H}(\nu\mid\mu)=+\infty$ si $\mathrm{supp}(\nu)\not\subset\mathrm{supp}(\mu)$.
— On a $\mathrm{H}(\nu\mid\mu)=\infty$ ssi $\mathrm{supp}(\nu_i)\not\subset\mathrm{supp}(\mu)$ c'est-à-dire qu'il existe $1\le i\le r$ tel que $\nu_i>0$ et $\mu_i=0$.
— L'entropie discrète S prend ses valeurs dans l'intervalle $[0,\log(r)]$ et atteint son maximum $\log(r)$ pour la loi uniforme $(1/r,\dots,1/r)$, et son minimum 0 pour les masses de Dirac δ_i, $1\le i\le r$. En effet, $u\mapsto u\log(u)$ est ≤ 0 sur $[0,1]$, nulle en 0 et 1, et S est continue et strictement concave. On rappelle au passage que l'entropie continue ou différentielle $-\int f(x)\log(f(x))\mathrm{d}x$, elle, prend ses valeurs dans tout $\mathbb{R}$.
— Pour tous $\mu,\nu\in\mathscr{P}(A)$, l'entropie relative vérifie $\mathrm{H}(\nu\mid\mu)\ge 0$ avec égalité ssi $\mu=\nu$, et ceci découle de l'inégalité de Jensen et de son cas d'égalité et de la convexité stricte de $u\mapsto u\log(u)$.
— La formule de Stirling $n!\sim\sqrt{2\pi n}(\frac{n}{\mathrm{e}})^n$ donne

$$\frac{1}{n}\log|L_n^{-1}(\nu)|=\frac{1}{n}\log\frac{n!}{(n\nu_1)!\cdots(n\nu_r)!}\xrightarrow[n\to\infty]{}-\sum_{i=1}^r\nu_i\log(\nu_i)=\mathrm{S}(\nu),$$

ce qui fait du (iii) une version quantitative de cette analyse asymptotique. Cette analyse asymptotique combinatoire a été menée par Boltzmann au milieu du dix-neuvième siècle en physique statistique, et redécouverte et exploitée par Shannon soixante ans plus tard en théorie de l'information [2].

Démonstration.

(i) On a $|L_n(A^n)|\le|\{0/n,\dots,n/n\}^r|=(n+1)^{r-1}$, le $r-1$ venant de la contrainte de somme à 1.
Soit $\nu\in\mathscr{P}(A)$. Montrons qu'il existe $\nu'\in L_n(A^n)$ tel que $\|\nu'-\nu\|_\infty\le(r-1)/n$ pour n assez grand. Quitte à permuter A, on peut supposer que $\nu_r>0$. Comme $L_n(A^n)\subset\{0/n,\dots,n/n\}^r$, pour tout $1\le i\le r-1$, il existe $\nu'_i\in\{0/n,\dots,n/n\}$ tel que $\max_{1\le i\le r-1}|\nu_i-\nu'_i|\le 1/n$. Soit $\nu'_r:=1-\sum_{i=1}^{r-1}\nu'_i$. On a $\nu'_r\ge\nu_r-(r-1)/n$, qui est ≥ 0 dès que $(r-1)/n\le\nu_r$, ce qui a lieu pour n assez grand. Enfin, dans ce cas, $|\nu'_r-\nu_r|\le(r-1)/n$.

(ii) On écrit, en notant que $\mathrm{S}(\nu)+\mathrm{H}(\nu\mid\mu)=-\sum_{i=1}^r\nu_i\log\mu_i$,

$$\mathbb{P}((X_1,\dots,X_n)=(x_1,\dots,x_n))=\prod_{k=1}^n\mathbb{P}(X_k=x_k)=\prod_{i=1}^r\mu_i^{\sum_{k=1}^n\mathbb{1}_{x_k=a_i}}=\prod_{i=1}^r\mu_i^{n\nu_i}=\mathrm{e}^{-n(\mathrm{S}(\nu)+\mathrm{H}(\nu|\mu))}.$$

(iii) Soit $\nu\in L_n(A^n)$ et $Y=(Y_1,\dots,Y_n)$ avec $Y_1,\dots,Y_n$ i.i.d. de loi ν. Le (ii) donne $\mathbb{P}(Y=y)=\mathrm{e}^{-n\mathrm{S}(\nu)}$ pour tout $y\in L_n^{-1}(\nu)$ car $\mathrm{H}(\nu\mid\nu)=0$, donc $1\ge\mathbb{P}(L_n(Y)=\nu)=|L_n^{-1}(\nu)|\mathrm{e}^{-n\mathrm{S}(\nu)}$, d'où la majoration $|L_n^{-1}(\mu)|\le\mathrm{e}^{n\mathrm{S}(\mu)}$.
Pour la minoration [3], comme $|L_n^{-1}(\nu)|=\frac{n!}{(n\nu_1)!\cdots(n\nu_r)!}$, on a, pour tout ν' tel que $\mathbb{P}(L_n(Y)=\nu')>0$,

$$\frac{\mathbb{P}(L_n(Y)=\nu)}{\mathbb{P}(L_n(Y)=\nu')}=\frac{|L_n^{-1}(\nu)|\prod_{i=1}^r\nu_i^{n\nu_i}}{|L_n^{-1}(\nu')|\prod_{i=1}^r\nu_i^{n\nu'_i}}=\prod_{i=1}^r\frac{(n\nu'_i)!}{(n\nu_i)!}\nu_i^{n(\nu_i-\nu'_i)},$$

produit d'expressions du type $\frac{a!}{b!}(\frac{b}{n})^{b-a}$. En distinguant les cas $a\ge b$ et $a<b$ il vient que $\frac{a!}{b!}\ge b^{a-b}$, d'où

$$\frac{\mathbb{P}(L_n(Y)=\nu)}{\mathbb{P}(L_n(Y)=\nu')}\ge\prod_{i=1}^r n^{n(\nu'_i-\nu_i)}=n^{n\sum_{i=1}^r(\nu'_i-\nu_i)}=1,$$

ce qui donne

$$1=\sum_{\nu'}\mathbb{P}(L_n(Y)=\nu')\le|L_n(A^n)|\mathbb{P}(L_n(Y)=\nu)=|L_n(A^n)||L_n^{-1}(\nu)|\mathrm{e}^{-n\mathrm{S}(\nu)}.$$

d'où enfin $(n+1)^{-(r-1)}\mathrm{e}^{n\mathrm{S}(\nu)}\le|L_n^{-1}(\nu)|$ par (i).

2. En utilisant un logarithme en base 2, la quantité $\mathrm{S}(\nu)$ est la longueur moyenne en bits des symboles d'un codage conservatif optimal de fréquence ν, dans lequel la longueur des codes est inversement proportionnelle à leur fréquence, cf. [27].
3. L'idée vient du fait que la LGN suggère que $\mathbb{P}(L_n(Y)=\nu')$ atteint son maximum pour $\nu'=\nu$ (maximum de vraisemblance!).

□

Théorème 4.1.2. Principe de grandes déviations de Sanov pour le jeu de dé.

Soient $(X_n)_{n\geq 1}$ des v.a. i.i.d. de loi $\mu \in \mathscr{P}(A)$. Alors pour tout borélien $B \subset \mathscr{P}(A)$,

$$-\inf_{\nu\in\mathring{B}} \mathrm{H}(\nu \mid \mu) \leq \varliminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}(L_n(X_1,\dots,X_n)\in B) \leq \varlimsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}(L_n(X_1,\dots,X_n)\in B) \leq -\inf_{\nu\in\overline{B}} \mathrm{H}(\nu \mid \mu).$$

— Lorsque B est convexe d'intérieur non-vide, la stricte convexité de $\mathrm{H}(\cdot \mid \mu)$ permet d'établir qu'il existe un unique ν_B dans l'adhérence de l'intérieur de B tel que

$$\lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}(L_n(X_1,\dots,X_n)\in B) = -\inf_{\nu\in B} \mathrm{H}(\nu \mid \mu) = -\mathrm{H}(\nu_B \mid \mu).$$

À l'opposé, la limite n'existe pas lorsque $B = \{\nu\}$ avec $\nu \in L_n(A^n)$ pour un n et $\mathrm{supp}(\nu) \subset \mathrm{supp}(\mu)$.

Démonstration. Pour tout $\nu \in L_n(A^n)$, grâce au lemme 4.1.1 (ii), en notant $X := (X_1,\dots,X_n)$ et $x := (x_1,\dots,x_n)$,

$$\mathbb{P}(L_n(X)=\nu) = \sum_{x\in L_n^{-1}(\nu)} \mathbb{P}(X=x) = |L_n^{-1}(\nu)|\mathrm{e}^{-n(\mathrm{S}(\nu)+\mathrm{H}(\nu|\mu))},$$

et en combinant avec le lemme 4.1.1 (iii) il vient, pour tout $\nu \in L_n(A^n)$,

$$(n+1)^{-(r-1)}\mathrm{e}^{-n\mathrm{H}(\nu|\mu)} \leq \mathbb{P}(L_n(X)=\nu) \leq \mathrm{e}^{-n\mathrm{H}(\nu|\mu)}.$$

À présent, pour tout borélien $B \subset \mathscr{P}(A)$, on a

$$\mathbb{P}(L_n(X)\in B) = \sum_{\nu\in B\cap L_n(A^n)} \mathbb{P}(L_n(X)=\nu) \leq |B\cap L_n(A^n)|\mathrm{e}^{-n\inf_{\nu\in B\cap L_n(A^n)}\mathrm{H}(\nu|\mu)} \leq (n+1)^{r-1}\mathrm{e}^{-n\inf_{\nu\in B\cap L_n(A^n)}\mathrm{H}(\nu|\mu)}$$

et

$$\mathbb{P}(L_n(X)\in B) = \sum_{\nu\in B\cap L_n(A^n)} \mathbb{P}(L_n(X)=\nu) \geq \sum_{\nu\in B\cap L_n(A^n)} (n+1)^{-(r-1)}\mathrm{e}^{-n\mathrm{H}(\nu|\mu)} \geq (n+1)^{-(r-1)}\mathrm{e}^{-n\inf_{\nu\in B\cap L_n(A^n)}\mathrm{H}(\nu|\mu)}.$$

Comme $\lim_{n\to\infty}\frac{1}{n}\log((n+1)^{r-1}) = 0$, il vient

$$\varlimsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}(L_n(X)\in B) = -\varliminf_{n\to\infty} \inf_{\nu\in B\cap L_n(A^n)} \mathrm{H}(\nu \mid \mu) \quad \text{et} \quad \varliminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}(L_n(X)\in B) = -\varlimsup_{n\to\infty} \inf_{\nu\in B\cap L_n(A^n)} \mathrm{H}(\nu \mid \mu)$$

À ce stade, la borne supérieure s'obtient en observant que $B\cap L_n(A^n) \subset B \subset \overline{B}$ (on démontre donc mieux!).

Établissons la borne inférieure. Cette fois-ci $\mathring{B} \not\subset B \cap L_n(A^n)$. Soit $\nu \in \mathring{B}$ tel que $\mathrm{supp}(\nu) \subset \mathrm{supp}(\mu)$ (sinon $\mathrm{H}(\nu \mid \mu) = +\infty$ et il n'y a rien à démontrer). Par le lemme 4.1.1 (i), il existe une suite (ν_n) dans $L_n(A^n)$ telle que $\nu_n \to \nu$ et $\mathrm{supp}(\nu_n) \subset \mathrm{supp}(\mu)$. Par ailleurs, la topologie étant celle de $\mathbb{R}^r$, comme $\nu \in \mathring{B}$, il existe $\varepsilon > 0$ tel que $\{\nu' : |\nu-\nu'| < \varepsilon\} \subset B$, et on a donc $\nu_n \in B \cap L_n(A^n)$ à partir d'un certain rang sur n. Cela donne

$$-\varlimsup_{n\to\infty} \inf_{\nu'\in B\cap L_n(A^n)} \mathrm{H}(\nu' \mid \mu) \geq -\lim_{n\to\infty} \mathrm{H}(\nu_n \mid \mu) = -\mathrm{H}(\nu \mid \mu),$$

d'où enfin

$$-\varlimsup_{n\to\infty} \inf_{\nu\in B\cap L_n(A^n)} \mathrm{H}(\nu \mid \mu) \geq -\inf_{\nu\in\mathring{B}} \mathrm{H}(\nu \mid \mu).$$

□

4.2 Entropie relative en probabilité, statistique, et physique statistique

Entropie relative comme énergie de déviation

La preuve du théorème 4.1.2 indique que si $X = (X_1,\dots,X_n)$ avec $X_1,\dots,X_n$ i.i.d. de loi $\mu \in \mathscr{P}(A)$, alors pour tout $\nu \in \mathscr{P}(A)$, quand $n \gg 1$, l'énergie $\mathrm{H}(\nu \mid \mu)$ est associée à l'événement $\{L_n(X) = \nu\}$ au sens où

$$\mathbb{P}(L_n(X)=\nu) \approx \mathrm{e}^{-n\mathrm{H}(\nu|\mu)}.$$

Autrement dit, la loi de la mesure empirique aléatoire $L_n(X)$ est $\approx$ une mesure de Boltzmann–Gibbs de température inverse n et d'énergie $\mathrm{H}(\cdot \mid \mu)$, minimisée par μ. Autrement dit, pour tous $\nu, \nu' \in \mathscr{P}(A)$, quand $n \gg 1$,

$$\frac{\mathbb{P}(L_n(X) = \nu)}{\mathbb{P}(L_n(X) = \nu')} \approx \mathrm{e}^{-n(\mathrm{H}(\nu|\mu) - \mathrm{H}(\nu'|\mu))}.$$

Comme pour toute mesure de Boltzmann–Gibbs, les rapports de probabilités se transforment en différences d'énergie. On peut concevoir $\mathrm{H}(\nu \mid \mu)$ comme l'énergie associée à la déviation ν par rapport au comportement typique μ. Voilà un aspect fondamental pour les praticiens des grandes déviations et de la mécanique/physique statistique, rompus aux techniques pour peser les énergies des événements et leurs contributions relatives.

Entropie relative comme contraste asymptotique associé à la log-vraisemblance

Soit $(X_n)_{n\geq 1}$ des v.a. i.i.d. de loi $\mu_* \in \mathscr{P}(A)$ inconnue. Lorsque $n \gg r$, il est naturel d'estimer μ_* avec la mesure empirique $L_n(X_1, \dots, X_n)$ qui converge vers μ_* quand $n \to \infty$ grâce à la LGN. En revanche, si $n \ll r$, alors il est possible de réduire la complexité ou la dimension avec un modèle paramétrique $(\mu^{(\theta)})_{\theta\in\Theta} \subset \mathscr{P}(A)$ de sorte que $\dim(\Theta) \ll n$ et $\mu_* = \mu_{\theta_*}$. Le but devient alors de construire, à partir de $X_1, \dots, X_n$, un estimateur de θ_*.

Pour des données $(x_1, \dots, x_n) \in A^n$ et un paramètre $\theta \in \Theta$, la vraisemblance (likelihood en anglais) est

$$\ell_{x_1,\dots,x_n}(\theta) := \mathbb{P}(Y_1 = x_1, \dots, Y_n = x_n) = \prod_{i=1}^{n} \mu_{x_i}^{(\theta)} \quad \text{où } Y_1, \dots, Y_n \text{ sont des v.a. i.i.d. de loi } \mu^{(\theta)}.$$

La vraisemblance est d'autant plus grande que les données $(x_1, \dots, x_n)$ sont localisées dans les zones les plus probables pour $\mu^{(\theta)}$, ou que le paramètre θ est localisé dans des zones qui rendent très probables les données $(x_1, \dots, x_n)$. Il s'agit à la fois de la vraisemblance de x pour θ et de la vraisemblance de θ pour x. L'estimateur de maximum de vraisemblance de θ^* au vu des données observées $X_1, \dots, X_n$ est

$$\widehat{\theta}_n := \arg\max_{\theta\in\Theta} \ell_{X_1,\dots,X_n}(\theta) = \arg\max_{\theta\in\Theta} \Big(\frac{1}{n} \log \ell_{X_1,\dots,X_n}(\theta) \Big).$$

La log-vraisemblance normalisée est une moyenne empirique, ce qui donne par LGN

$$\frac{1}{n} \log \ell_{X_1,\dots,X_n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \mu_{X_i}^{(\theta)} \xrightarrow[n\to\infty]{\text{p.s.}} \sum_{k=1}^{r} \mu_k^{(\theta_*)} \log \mu_k^{(\theta)} = \underbrace{-\mathrm{S}(\mu^{(\theta_*)})}_{\text{constante}} - \mathrm{H}(\mu^{(\theta_*)} \mid \mu^{(\theta)}).$$

Ainsi, maximiser la vraisemblance est équivalent asymptotiquement à minimiser l'entropie relative de la loi inconnue. On parle d'estimateur de maximum de contraste, de contraste asymptotique $-\mathrm{H}(\mu^{(\theta_*)} \mid \cdot)$. La même trame s'applique au cadre continu avec l'entropie continue, en remplaçant les sommes par des intégrales, la mesure de comptage par la mesure de Lebesgue. Il s'agit d'un lien entre statistique mathématique et physique statistique. On voit poindre là deux usages actuels du mot statistique, qui correspondent à des cultures et à des communautés scientifiques qui se recoupent de fait assez peu. Il se trouve qu'un lien entre les deux est assuré notamment par une certaine informatique de l'apprentissage automatique.

Maximum d'entropie, minimum d'énergie libre

Revenons au concept boltzmannien de maximum d'entropie S à énergie moyenne fixée, en utilisant les notations du lemme 4.1.1 et du chapitre 1. Si $A = \{1, \dots, r\}$ et $E : A \to (-\infty, +\infty]$ autrement dit $E \in (-\infty, +\infty]^r$, avec $E \not\equiv +\infty$, si $\beta \in \mathbb{R}$, et si $\mu^\beta \in \mathscr{P}(A)$ est définie par

$$\mu_i^\beta := \frac{\mathrm{e}^{-\beta E_i}}{Z_\beta}, \quad \text{avec } Z_\beta := \langle \mathrm{e}^{-\beta V}, 1 \rangle = \sum_{i=1}^{r} \mathrm{e}^{-\beta E_i},$$

alors pour tout $\mu \in \mathscr{P}(A)$ tel que $\langle V, \mu \rangle = \langle V, \mu^\beta \rangle$, on a $\mathrm{S}(\mu^\beta) - \mathrm{S}(\mu) = \mathrm{H}(\mu \mid \mu^\beta)$, d'où

$$\mathrm{S}(\mu) \leq S(\mu^\beta) \quad \text{avec égalité ssi } \mu = \mu^\beta.$$

En introduisant l'énergie libre de Helmholtz $\mathrm{F}(\mu) := \langle V, \mu \rangle - \frac{1}{\beta}\mathrm{S}(\mu)$, il vient, pour tous $\mu, \nu \in \mathscr{P}(A)$,

$$\mathrm{F}(\mu^\beta) - \mathrm{F}(\mu) = \mathrm{H}(\mu \mid \mu_\beta) \geq 0 \quad \text{avec égalité ssi } \mu = \mu^\beta, \quad \text{et} \quad \mathrm{F}(\mu^\beta) = -\frac{1}{\beta} \log Z_\beta.$$

Notons que les températures négatives sont permises par la finitude de l'espace A.

Statistiques de Maxwell–Boltzmann, Bose–Einstein, et Fermi–Dirac

Considérons un système formé de n particules distinguables chacune pouvant être dans l'un des r niveaux d'énergie $E_1,\dots,E_r$. La description microscopique du système est $(e_1,\dots,e_n)$ où $e_k \in \{1,\dots,r\}$ est le niveau d'énergie de la particule k, tandis que la description macroscopique du système est donnée par $(n_1,\dots,n_r)$ où n_i est le nombre de particules au niveau E_i. On a $n = n_1 + \cdots + n_r$ et l'énergie moyenne de la configuration macroscopique $(n_1,\dots,n_r)$ est $E = n_1E_1 + \cdots + n_rE_r$. Il y a $\binom{n}{n_1,\dots,n_r} = n!\prod_{i=1}^r \frac{1}{n_i!}$ configurations microscopiques compatibles avec la configuration macroscopique $(n_1,\dots,n_r)$. Ce coefficient multinomial revient à répartir n boules distinguables dans r urnes. Afin d'obtenir une mesure additive des degrés de liberté du système, on peut considérer le logarithme. Cette quantité par particule est alors $\frac{1}{n}\log\binom{n}{n_1,\dots,n_r}$. Comme dans la remarque 1.2.8, si $n\to\infty$ avec $n_i/n \to \mu_i$ pour tout $1 \le i \le n$ alors $\mu_1 + \cdots + \mu_r = 1$, et par la formule de Stirling $n! \sim \sqrt{2\pi n}n^n\mathrm{e}^{-n}$, le nombre asymptotique de degrés de libertés « additifs » par particule est

$$\mathrm{S}(\mu) := -\sum_{i=1}^r \mu_i \log(\mu_i),$$

Maximiser cette entropie S sous contrainte d'énergie moyenne fixée $E = \sum_{i=1}^r \mu_i E_i$ donne

$$\mu_i = \frac{1}{Z}\mathrm{e}^{-\beta E_i}, \quad 1 \le i \le r,$$

où le multiplicateur de Lagrange $\beta > 0$ peut être interprété comme une température inverse, et où $Z = \sum_{i=1}^r e^{-\beta E_i}$ est la constante de normalisation. On parle de distribution de Boltzmann voir de Boltzmann–Gibbs. Mais comme observé par Gibbs, le raisonnement ci-dessus souffre d'un problème important. Plus précisément, si on considère un système de $n+n'$ particules, alors le logarithme des degrés de liberté n'est pas additif en réalité, la combinatoire n'est pas la bonne, ce qui viole le second principe de la thermodynamique. Une manière de dépasser ce paradoxe consiste à supposer que les particules sont indistinguables. Ceci rend la combinatoire triviale car pour toute configuration macroscopique $(n_1,\dots,n_r)$, il existe une unique configuration microscopique compatible, c'est-à-dire une unique façon de répartir n boules indistinguables dans r urnes avec n_i boules dans l'urne i pour tout i. Pour échapper à ce problème, on peut supposer que chaque niveau d'énergie E_i possède s_i états, en d'autre termes que chaque urne i possède s_i sous-urnes. Le nombre de configuration microscopiques compatibles avec la configuration macroscopique $(n_1,\dots,n_r)$ est alors donné par $\prod_{i=1}^r \frac{(n_i+s_i-1)!}{n_i!(s_i-1)!}$. Si $s_1 = \cdots = s_r = 1$ alors ce nombre est 1. Si $n_i \ll s_i$ alors grâce à la formule de Stirling, le nombre de configurations microscopiques compatibles avec la configuration macroscopique $(n_1,\dots,n_r)$ est $\approx \prod_{i=1}^r \frac{s_i^{n_i}}{n_i!}$. En utilisant l'approche variationnelle utilisée dans le cas distinguable, on obtient une nouvelle distribution pour les niveaux d'énergie, appelée distribution ou statistique de Maxwell–Boltzmann :

$$\mu_i \approx \frac{1}{Z}s_i e^{-\beta E_i}.$$

Lorsque $s_1 = \cdots = s_r$ on retrouve la distribution de Boltzmann précédente du cas distinguable.

En suivant la même méthode sans l'hypothèse $n_i \ll s_i$, on obtient la statistique de Bose–Einstein :

$$\mu_i \approx \frac{1}{Z}\frac{s_i}{e^{\beta E_i-\alpha}-1}.$$

Par ailleurs, si on suppose qu'au plus une particule occupe un sous-niveau d'énergie, on obtient que le nombre de configurations microscopiques compatibles avec l'état macroscopique $(n_1,\dots,n_r)$ est $\prod_{i=1}^m \frac{s_i!}{n_i!(s_i-n_i)!}$, ce qui donne cette fois la statistique de Fermi–Dirac :

$$\mu_i = \frac{1}{Z}\frac{s_i}{e^{\beta E_i-\alpha}+1}.$$

Ici le terme de statistique est celui historiquement utilisé pour parler de distribution de probabilité.

4.3 Modèle de Curie–Weiss

Il s'agit de la mesure de Boltzmann–Gibbs « champ moyen » suivante sur $\{-1,+1\}^n$:

$$\mu_n(x) = \frac{1}{Z_n}\mathrm{e}^{-\beta H_n(x)} \quad \text{où} \quad H_n(x) := \sum_{i=1}^n x_i \frac{1}{n}\sum_{j=1}^n x_j + \sum_{i=1}^n hx_i \quad \text{et} \quad Z_n := \sum_{x\in\{-1,1\}^n} \mathrm{e}^{-\beta H_n(x)}.$$

La loi μ_n n'est pas produit et on ne peut donc pas utiliser le théorème de Cramér pour obtenir un PGD pour la moyenne empirique $m_n := \frac{1}{n}\sum_{i=1}^n x_i$. Cependant, l'énergie $H_n(x)$, et donc la probabilité $\mu_n(x)$, ne dépend que de la moyenne empirique $m_n(x)$ au sens où $H_n(x) = n(m_n(x)^2 + h m_n(x))$. Aussi, en écrivant

$$\mu_n(x : m_n(x) = m) = \frac{1}{Z_n} \mathrm{e}^{-n(m^2+hm)} |\{x : m_n(x) = m\}|$$

et en estimant Z_n et $|\{x : m_n(x) = m\}|$, dans le même esprit que dans la preuve du théorème 4.1.2 de Sanov pour le jeu de dé, on peut établir que sous la loi μ_n, la moyenne empirique m_n vérifie un PGD, et en particulier la distance de m_n à l'ensemble des minimiseurs de la fonction de taux converge vers zéro quand $n \to \infty$. De plus un phénomène de seuil se produit : la fonction de taux a un ou deux minimiseurs en fonction de β et h.

4.4 Théorème de Cramér discret

Corollaire 4.4.1. Théorème de Cramér pour le jeu de dé.

Soient $(X_n)_{n\geq 1}$ des v.a.r. i.i.d. de loi $\mu \in \mathscr{P}(A)$ portée par une partie finie $A = \{a_i : 1 \leq i \leq r\} \subset \mathbb{R}$, et $S_n := X_1 + \cdots + X_n$. Alors pour tout $B \subset \mathbb{R}$, avec $I(x) := \inf_{\nu \in \mathscr{P}(A) : \langle a, \nu\rangle = x} \mathrm{H}(\nu \mid \mu)$,

$$-\inf_{\mathring{B}} I \leq \varliminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \in B\Big) \leq \varlimsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \in B\Big) \leq -\inf_{\overline{B}} I.$$

De plus, pour tout $x \in K := [\min_{1\leq i\leq r}(a_i), \max_{1\leq i\leq r}(a_i)]$, I est continue en x et

$$I(x) := \Lambda^*(x) := \sup_{\lambda\in\mathbb{R}} (\lambda x - \Lambda(\lambda)) \quad \text{où} \quad \Lambda(\lambda) := \log \mathbb{E}(\mathrm{e}^{\lambda X_1}) = \log \sum_{i=1}^r \mu_i \mathrm{e}^{\lambda a_i}.$$

— Comme la fonction de taux I est continue on obtient, lorsque $B = \overline{\mathring{B}} \subset K$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \in B\Big) = -\inf_B I.$$

Si μ n'est pas une masse de Dirac, alors Λ^* est strictement convexe, et son unique minimum est la moyenne $m = \langle a, \mu\rangle$ de μ. Comme $\inf_{(m-\varepsilon, m+\varepsilon)} I \in (0, \infty)$ pour tout $\varepsilon > 0$, la borne supérieure du PGD pour $B = (m-\varepsilon, m+\varepsilon)^c$ et le lemme de Borel–Cantelli donnent la LGN forte (donc PGD $\Rightarrow$ LGN forte)

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{\text{p.s.}} m.$$

— Le corollaire 4.4.1 se généralise à $\mathbb{R}^d$, et K devient dans ce cas l'enveloppe convexe des a_i.

Démonstration. L'observation cruciale est la suivante : $\frac{S_n}{n} = \langle a, L_n(X_1, \ldots, X_n)\rangle$. Ainsi, pour tout $B \subset \mathbb{R}$,

$$\frac{S_n}{n} \in B \quad \text{ssi} \quad L_n(X_1, \ldots, X_n) \in \{\nu \in \mathscr{P}(A) : \langle a, \nu\rangle \in B\}.$$

Lorsque B est ouvert, l'ensemble $\{\nu \in \mathscr{P}(A) : \langle a, \nu\rangle \in B\}$ est ouvert, et les bornes inférieure et supérieure de PGD du théorème découlent de celles du théorème 4.1.2 de Sanov pour le jeu de dé.

Pour tout $\nu \in \mathscr{P}(A)$ et tout $\lambda \in \mathbb{R}$, l'inégalité de Jensen donne [4]

$$\Lambda(\lambda) = \log \sum_{i=1}^r \mu_i \mathrm{e}^{\lambda a_i} \geq \sum_{i=1}^r \nu_i \log \frac{\mu_i \mathrm{e}^{\lambda a_i}}{\nu_i} = \lambda\langle a, \nu\rangle - \mathrm{H}(\nu \mid \mu) \quad \text{avec égalité ssi} \quad \nu = \nu_\lambda := \mathrm{e}^{\lambda a - \Lambda(\lambda)} \mu.$$

Donc pour tous λ et x,

$$\lambda x - \Lambda(\lambda) \leq \inf_{\nu\in\mathscr{P}(A):\langle a,\nu\rangle = x} \mathrm{H}(\nu \mid \mu) = I(x), \quad \text{avec égalité si } \langle a, \nu_\lambda\rangle = x.$$

Comme Λ est dérivable et $\Lambda'(\lambda) = \langle a, \nu_\lambda\rangle$, on obtient la formule pour I lorsque $x \in \{\Lambda'(\lambda) : \lambda \in \mathbb{R}\}$. Or Λ' est croissante car Λ est convexe, $\min_i a_i = \inf_\lambda \Lambda'(\lambda)$, $\max_i a_i = \sup_\lambda \Lambda'(\lambda)$, d'où la formule pour I pour tout $x \in \mathring{K}$. Considérons enfin un point du bord de K, disons $x = a_*$. Alors en notant $\nu_* := \delta_{a_*}$, on a $\langle a, \nu_*\rangle = x$ et

$$-\log \mu(a_*) = \mathrm{H}(\nu_* \mid \mu) \geq I(x) \geq \sup_\lambda (\lambda x - \Lambda(\lambda)) \geq \lim_{\lambda\to-\infty} (\lambda x - \Lambda(\lambda)) = -\log \mu(a_*).$$

Enfin la continuité de I sur K est une conséquence de celle de $\mathrm{H}(\cdot \mid \mu)$. □

4. Ce n'est rien d'autre que l'inégalité de Young pour la transformée de Legendre d'une log-Laplace, qui est une entropie relative !

La preuve du corollaire 4.4.1 à partir du théorème 4.1.2 est une instance d'une méthode générale appelée principe de contraction, qui permet de déduire un PGD d'un PGD en composant avec une fonction.

Théorème 4.4.2. Principe de contraction.

Soit $f : E \to F$ continue entre deux espaces topologiques E et F avec F séparé.

(i) Si $I : E \to [0, +\infty]$ a des ensembles de sous-niveau compacts alors $J := \inf(I \circ f^{-1}) : F \to [0, +\infty]$ a aussi des ensembles de sous-niveau compacts, avec la convention naturelle $\inf \varnothing = +\infty$.

(ii) Si E et F sont munis de la tribu borélienne, et si $(\mu_n)_{n\geq 1}$ est une suite de lois sur E vérifiant un PGD de vitesse $(\beta_n)_{n\geq 1}$ et de fonction de taux $I : E \to [0, +\infty]$: pour tout borélien $B \subset E$,

$$-\inf_{\mathring{B}} I \leq \underline{\lim}_{n\to\infty} \frac{1}{\beta_n} \log \mu_n(B) \leq \overline{\lim}_{n\to\infty} \frac{1}{\beta_n} \log \mu_n(B) \leq -\inf_{\overline{B}} I,$$

alors $(\nu_n)_{n\geq 1} := (\mu_n \circ f^{-1})_{n\geq 1}$ vérifie un PGD de même vitesse et de fonction de taux $J = \inf(I \circ f^{-1})$.

— La fonction f contracte l'espace E en F.

— La formule $J = \inf(I \circ f^{-1})$ signifie que pour tout $y \in F$, on a

$$J(y) = \inf_{x\in E : f(x)=y} I(x).$$

Comme f n'est pas forcément bijective, $f^{-1}(y)$ est un ensemble, éventuellement vide, et il en va de même pour $(I \circ f^{-1})(y)$, de sorte que $J(y) = \inf(I \circ f^{-1})(y) \in (-\infty, +\infty]$ peut être égal à $+\infty$ parce que I prend cette valeur sur l'ensemble $(I \circ f^{-1})(y)$ ou bien parce que $(I \circ f^{-1})(y)$ est vide.

— Ce principe permet à partir d'un PGD d'obtenir une grande variété de PGD. Cependant la fonction de taux contractée n'est pas toujours très intuitive. À ce sujet, à la question « appréciez-vous les grandes déviations ? » certains sont tentés de répondre « *montrez-moi votre fonction de taux et je vous le dirai !* ».

— Le PGD pour $(\mu_n)_{n\geq 1}$ est une manière de concevoir μ_n quand $n \gg 1$ comme la mesure de Boltzmann–Gibbs $\mu_n(B) \approx \mathrm{e}^{-\beta_n \inf_B I}$. Les PGD sont ainsi des Boltzmann–Gibbsations de grande dimension !

Démonstration.

(i) Pour tout $r \in \mathbb{R}$, l'ensemble de sous-niveau

$$\{y \in F : J(y) \leq r\} = \{y \in F : \forall \epsilon > 0, \exists x_\epsilon \in E, f(x_\epsilon) = y, I(x_\epsilon) \leq r + \varepsilon\} = \bigcap_{\varepsilon>0} f(\{x \in E : I(x) \leq r + \varepsilon\}).$$

Or pour tout $\varepsilon > 0$, l'ensemble $f(\{x \in E : I(x) \leq r + \varepsilon\})$ est compact dans F (éventuellement vide) car image par f continue du compact $\{x \in E : I(x) \leq r + \varepsilon\}$ de E. En particulier $\{y \in F : J(y) \leq r\}$ est fermé comme intersection de fermés. Comme l'intersection est décroissante quand $\varepsilon \searrow 0$, elle est incluse dans un compact de l'espace séparé F. Elle est donc compacte car fermée dans un compact d'un espace séparé.

(ii) Pour tout $B' \subset F$, on a $\inf_{B'} J = \inf_{f^{-1}(B')} I$, et comme f est continue, l'ensemble $f^{-1}(B')$ est fermé si B est fermé et ouvert si B est ouvert, d'où le PGD pour $(\nu_n)_{n\geq 1}$ à partir de celui pour $(\mu_n)_{n\geq 1}$.

□

4.5 Théorème de Sanov général

Soient $(X_n)_{n\geq 1}$ des v.a. i.i.d. à valeurs dans un espace polonais E muni de sa tribu borélienne. Soit $\mathscr{P}(E)$ l'ensemble des mesures de probabilités sur E, muni de la topologie de la convergence pour les fonctions tests continues et bornées $\mathscr{C}_b(E \to \mathbb{R})$. On munit $\mathscr{P}(E)$ de sa tribu borélienne. La mesure empirique

$$L_n(X_1, \ldots, X_n) := \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k}$$

est une variable aléatoire à valeurs dans $\mathscr{P}(E)$. Pour tout $f \in \mathscr{C}_b(E \to \mathbb{R})$, la LGN appliquée à $(f(X_n))_{n\geq 1}$ donne

$$\langle L_n(X_1, \ldots, X_n), f\rangle = \frac{1}{n} \sum_{k=1}^{n} f(X_k) \xrightarrow[n\to\infty]{} \langle \mu, f\rangle \quad \text{presque sûrement,}$$

où $\langle \mu, f \rangle := \int f \mathrm{d}\mu$ (crochet de dualité). Comme E est séparable, la convergence faible est caractérisable avec une classe dénombrable de fonctions test, d'où la LGN sur les mesures empiriques : $L_n(X_1,\dots,X_n) \xrightarrow[n\to\infty]{\mathscr{C}_b} \mu$ p.s. Le théorème de Sanov général suivant constitue le PGD naturellement associé à cette LGN [5].

Théorème 4.5.1. de Sanov.

Soient $(X_n)_{n\geq 1}$ des v.a. i.i.d. à valeurs dans un espace polonais E muni de sa tribu borélienne. Soit $\mathscr{P}(E)$ l'ensemble des mesures de probabilités sur E, muni de la topologie de la convergence pour les fonctions tests continues et bornées $\mathscr{C}_b(E \to \mathbb{R})$. On munit $\mathscr{P}(E)$ de sa tribu borélienne. La mesure empirique $L_n(X_1,\dots,X_n) := \frac{1}{n}\sum_{k=1}^n \delta_{X_k}$ est une variable aléatoire à valeurs dans $\mathscr{P}(E)$. Pour tout borélien $B \subset \mathscr{P}(E)$,

$$-\inf_{\mathring{B}} \Lambda^* \leq \underline{\lim}_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(L_n(X_1,\dots,X_n) \in B\Big) \leq \overline{\lim}_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(L_n(X_1,\dots,X_n) \in B\Big) \leq -\inf_{\overline{B}} \Lambda^*,$$

avec

$$\Lambda^*(\nu) := \sup_{f\in\mathscr{C}_b(E)} (\langle f, \nu\rangle - \Lambda(f)) = \mathrm{H}(\nu \mid \mu) \quad \text{où} \quad \Lambda(f) := \log \int \mathrm{e}^f \mathrm{d}\mu.$$

— Le théorème 4.5.1 de Sanov pour le jeu de dé est le cas particulier $E = A = \{a_1,\dots,a_r\}$.
— Le théorème 4.5.1 de Sanov est disponible en particulier pour $E = \mathbb{R}^d$, $d \geq 1$.
— La LGN exprimée sur les mesures empiriques reste valable pour la topologie plus forte relative aux fonctions tests mesurables et bornées $\mathscr{M}_b(E \to \mathbb{R})$, appelée topologie τ :

$$L_n(X_1,\dots,X_n) \xrightarrow[n\to\infty]{\mathscr{M}_b} \mu \quad \text{presque sûrement.}$$

Il est possible d'établir le théorème de Sanov pour la topologie τ, cf. par exemple [6] pour une déduction par discrétisation ou approximation finie à partir du théorème 4.1.2 de Sanov pour le jeu de dé, via

$$\mathrm{H}(\nu \mid \mu) = \sup_{A\in\Pi} \mathrm{H}(\nu_A \mid \mu_A) \quad \text{avec} \quad \mathrm{H}(\nu_A \mid \mu_A) = \sum_{i=1}^{r(A)} \nu(A_i) \log \frac{\nu(A_i)}{\mu(A_i)},$$

où Π est l'ensemble des partitions finies de E et $r(A)$ le nombre de parties de la partition $A \in \Pi$.
— La dernière égalité dans le théorème 4.5.1 fait écho à la remarque 2.5.6 sur l'entropie relative.
— Dans le théorème de Sanov, l'usage de la mesure empirique libère E de la structure vectorielle. Cela fait du théorème de Sanov une sorte de théorème de Cramér pour les variables aléatoires i.i.d. $(\delta_{X_n})_{n\geq 1}$ à valeurs dans la partie convexe $\mathscr{P}(E)$ de l'espace $\mathscr{M}(E)$ des mesures signées sur E équipées de la même topologie faible. Il se trouve que $\mathscr{M}(E)$ est un espace vectoriel topologique localement convexe séparé dont le dual est identifiable à $\mathscr{C}_b(E)$. Or si $(Z_n)_{n\geq 1}$ sont des v.a. i.i.d. de loi m à valeurs dans un espace vectoriel topologique M localement convexe séparé muni de sa tribu borélienne, alors un théorème de Cramér général est disponible, la log-Laplace est définie sur le dual topologique M^* par

$$\Lambda(\lambda) := \log \int \mathrm{e}^{\langle \lambda, x\rangle} m(\mathrm{d}x), \quad \lambda \in M^*,$$

et la transformée de Cramér est la transformée de Legendre [6]

$$\Lambda^*(x) = \sup_{\lambda\in M^*} (\langle \lambda, x\rangle - \Lambda(\lambda)), \quad x \in M.$$

Pour retrouver le théorème de Sanov on observe que si $M = \mathscr{M}(E)$ et $Z_n = \delta_{X_n}$, alors

$$\Lambda(f) = \log \mathbb{E}\Big(\mathrm{e}^{\langle f, \delta_{X_1}\rangle}\Big) = \log \int \mathrm{e}^f \mathrm{d}\mu, \quad f \in M^* = \mathscr{C}_b(E).$$

Ces développements et bien d'autres se trouvent dans [32], [30, ch. 6], [21] par exemple.
— Nathanaël Gozlan a montré dans [45] que le théorème de Sanov permet de relier concentration sous-gaussienne tensorisable pour les fonctions Lipschitz et inégalité de transport de Talagrand W_2.

Ce chapitre est localement librement et fortement inspiré de [30] et [76].

5. Formuler la LGN en terme de mesure empirique permet de se passer de structure vectorielle sur l'espace E, qui peut alors être très général : graphe fini ou infini, variété, ensemble de parties d'un ensemble, partie non-linéaire d'un espace de fonctions ou de mesures, etc. La mesure empirique est une moyenne empirique sur les mesures. Elle s'obtient par plongement de l'espace E, qui n'a pas forcément de structure vectorielle, dans l'espace des mesures sur E, qui a toujours une structure vectorielle. Cette technique de plongement dans un espace plus gros et plus linéaire est classique, on la retrouve notamment dans les support vector machine en apprentissage automatique.
6. Dans ce contexte de dimension infinie, on parle de transformée de Fenchel–Legendre.

Chapitre 5

Queues lourdes, lois stables, universalité

Je 06/03

Vu en cours :
— Principe de contraction et sa preuve du chapitre précédent
— Commentaires sur $H(\nu \mid \mu)$ liés à Sanov et à la statistique du chapitre précédent
— Sanov général et lien avec Cramér général du chapitre précédent
— Introduction de ce chapitre, Pareto, théorème sur les queues lourdes avec 2/3 de preuve
— Théorème sur queue en loi de puissance et invariance d'échelle, idée de la preuve.

Ce chapitre est une exploration du comportement des variables aléatoires peu intégrables.

Soient $(X_n)_{n\geq 1}$ des v.a.r. i.i.d. Si $\mathbb{E}(X_1^2) < \infty$ et $\sigma^2 := \mathbb{E}(X_1^2) - m^2 > 0$ où $m := \mathbb{E}(X_1)$, alors par le TLC,

$$\frac{X_1 + \cdots + X_n - nm}{\sqrt{n\sigma^2}} \xrightarrow[n\to\infty]{\text{loi}} \mathcal{N}(0,1),$$

et il y a de plus égalité en loi pour tout n dans le cas gaussien : $\frac{X_1 - m}{\sigma} \sim \mathcal{N}(0,1)$ et

$$X_1 \sim \mathcal{N}(0,1) \quad \Rightarrow \quad \frac{X_1 + \cdots + X_n}{\sqrt{n}} \overset{\text{d}}{=} X_1.$$

Que se passe-t-il si $\mathbb{E}(X_1^2) = \infty$? Le cas où X_1 suit la loi Cauchy(x_0, γ), $x_0 \in \mathbb{R}$, $\gamma > 0$, de densité [1]

$$x \in \mathbb{R} \mapsto \frac{1}{\gamma\pi\left(1 + \left(\frac{x - x_0}{\gamma}\right)^2\right)} \sim_{|x|\to\infty} \frac{\gamma}{\pi}\frac{1}{x^2}$$

est instructif : dans ce cas, $\mathbb{E}(|X_1|) = \infty$ et $\mathbb{E}(X_1^2) = \infty$, $\frac{X_1 - x_0}{\gamma} \sim$ Cauchy$(0,1)$, et

$$X_1 \sim \text{Cauchy}(0,1) \quad \Rightarrow \quad \frac{X_1 + \cdots + X_n}{n} \overset{\text{loi}}{=} X_1,$$

qui suggère un analogue du TLC avec une normalisation de n au lieu de $\sqrt{n}$ et une loi limite Cauchy au lieu de gaussienne. En terme de fonctions caractéristiques, l'égalité en loi ci-dessus est liée au fait que pour tout $t \in \mathbb{R}$,

$$\varphi_{X_1}(t) := \mathbb{E}(e^{itX_1}) = \int_{\mathbb{R}} \frac{e^{itx}}{\pi(1+x^2)}\mathrm{d}x = \int_{\mathbb{R}} \frac{\cos(tx)}{\pi(1+x^2)}\mathrm{d}x = e^{-|t|},$$

qui donne

$$\varphi_{\frac{X_1+\cdots+X_n}{n}}(t) = \mathbb{E}(e^{it\frac{X_1+\cdots+X_n}{n}}) = \left(\varphi_{X_1}\left(\frac{t}{n}\right)\right)^n = e^{n i \frac{|t|}{n}} = \varphi_{X_1}(t).$$

La loi gaussienne et la loi de Cauchy sont stables par convolution, à translation et dilatation près.

Cela conduit à clarifier les deux points suivants :
— Lois stables par convolution, à translation et dilatation près.
— Universalité de la convergence vers ces lois pour les sommes de v.a.r. i.i.d., à translation et dilatation près, sous condition de queue de distribution $\mathbb{P}(X > x) = 1 - F(x)$ sur les v.a.r.

5.1 Queues lourdes, invariance d'échelle, variation régulière, variation lente

1. On dit que x_0 est un paramètre de position tandis que γ est un paramètre d'échelle.

Théorème 5.1.1. Queue lourde.

Soit X une v.a.r. positive de loi μ et de fonction de répartition F. Ces propriétés sont équivalentes :

(i) X a une queue plus lourde que toute exponentielle : pour tout $\lambda > 0$, $\overline{\lim}_{x\to+\infty} \frac{1-F(x)}{e^{-\lambda x}} = +\infty$.

(ii) X ne possède aucun moment exponentiel : $\mathbb{E}(e^{\lambda X}) = \infty$ pour tout $\lambda > 0$.

(iii) X a une queue sous-log-linéaire : $\underline{\lim}_{x\to+\infty} -\frac{\log(1-F(x))}{x} = 0$.

— Lorsque les conditions du théorème sont vérifiées, on dit que X ou μ est à queue lourde ou épaisse.
— Pour les variables sur $\mathbb{R}$, on peut distinguer queue à droite et queue à gauche.
— La loi de Cauchy est à queue lourde. Les lois exponentielle et normale ne sont pas à queue lourde.
— Pour cette notion de queue lourde, la loi exponentielle est le cas critique, exclu, en quelque sorte.

Démonstration. (i)⇒(ii). Par (i), pour tout $\lambda > 0$, il existe une suite (x_n) t.q. $\lim_{n\to\infty} e^{\lambda x_n}(1-F(x_n)) = +\infty$, d'où

$$\mathbb{E}(e^{\lambda X}) = \int_0^\infty e^{\lambda x} d\mu(x) \geq e^{\lambda x_n} \int_{x_n}^\infty d\mu(x) = e^{\lambda x_n}(1-F(x_n)) \to +\infty.$$

(ii)⇒(iii). Démontrons la contraposée ou supposons par l'absurde que $\underline{\lim}_{x\to+\infty} -\frac{\log(1-F(x))}{x} > 0$. Il existe donc $\lambda_0 > 0$ et $x_0 > 0$ tels que $-\frac{\log(1-F(x))}{x} \geq \lambda_0$ pour tout $x \geq x_0$, c'est-à-dire $1-F(x) \leq e^{-\lambda_0 x}$, pour tout $x \geq x_0$. À présent, pour tout $\lambda \in (0, \lambda_0)$, cela permet de contredire (ii) :

$$\mathbb{E}(e^{\lambda X}) = \lambda \int_0^\infty e^{\lambda x}(1-F(x))dx \leq \lambda e^{\lambda x_0} + \lambda \int_{x_0}^\infty e^{\lambda x}(1-F(x))dx < +\infty.$$

(iii)⇒(i). De (iii) on tire une suite (x_n) telle que à la fois $\lim_{n\to\infty} x_n = \infty$ et $\lim_{n\to\infty} -\log(1-F(x_n))/x_n = 0$. Pour tout $\lambda > 0$ et $\lambda_0 \in (0, \lambda)$, il existe donc un rang n_0 tel que $-\log(1-F(x_n))/x_n < \lambda_0$ pour tout $n \geq n_0$, c'est-à-dire que $(1-F(x_n)) > e^{-\lambda_0 x_n}$ pour tout $n \geq n_0$, ce qui implique que $\lim_{n\to\infty}(1-F(x_n))/e^{-\lambda x_n} = \infty$, ce qui implique que $\overline{\lim}_{n\to\infty}(1-F(x))/e^{-\lambda x} = \infty$, et comme cela est vrai pour tout $\lambda > 0$, on obtient bien (i). □

Exemple 5.1.2. Loi de Pareto.

Les distributions dont la queue est en loi de puissance forment une classe naturelle de distributions à queue lourde. Un exemple important est la loi de Pareto, dont la densité est [a]

$$x \mapsto \frac{\alpha x_{\min}^\alpha}{x^{\alpha+1}} \mathbb{1}_{x\geq x_{\min}}, \quad x_{\min} > 0,\ \alpha > 0.$$

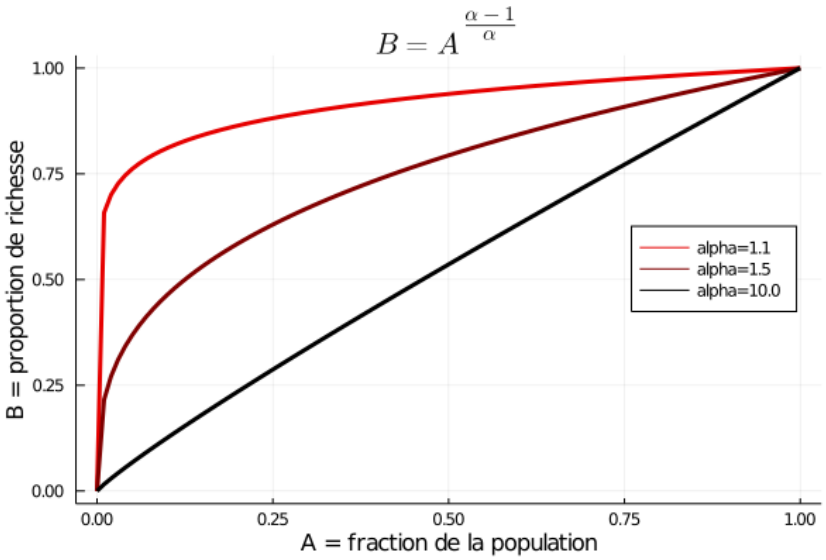
Si $X \sim \text{Pareto}(x_{\min}, \alpha)$ alors $\mathbb{E}(X^\beta) < \infty$ ssi $\beta < \alpha$, en particulier $\mathbb{E}(X^\alpha) = +\infty$. La loi de Pareto a été utilisée par le sociologue et économiste Vilfredo Pareto pour modéliser la distribution des revenus. En notant f la densité ci-dessus, pour tout $x \geq x_{\min}$, la fraction de la population dont le revenu dépasse x est [b]

$$A(x) = 1 - F(x) = \int_x^\infty f(y)dy = \left(\frac{x}{x_{\min}}\right)^{-\alpha}$$

tandis que la proportion de richesse de cette fraction par rapport à la richesse totale est, si $\alpha > 1$,

$$B(x) = \frac{\int_x^\infty y f(y)dy}{\int_{x_{\min}}^\infty y f(y)dy} = \left(\frac{x}{x_{\min}}\right)^{-\alpha+1}, \quad \text{d'où} \quad B = A^{\frac{\alpha-1}{\alpha}}.$$

Cette relation entre A et B mène au principe de Pareto lorsque α est proche de 1 par excès : une petite fraction de la population détient une grande part de la richesse totale [c]. À l'opposé $B = A$ lorsque $\alpha \to \infty$.



Notons enfin que la loi de Pareto pourrait être appelée loi log-exponentielle au sens où

$$Y \sim \mathrm{Exp}(\alpha) \text{ ssi } x_{\min} e^{Y} \sim \mathrm{Pareto}(x_{\min}, \alpha) \quad \text{et réciproquement} \quad X \sim \mathrm{Pareto}(x_{\min}, \alpha) \text{ ssi } \log \frac{X}{x_{\min}} \sim \mathrm{Exp}(\alpha).$$

Remplacer la loi exponentielle par la loi normale conduit à la loi log-normale, très utilisée aussi.

a. On dit que m est un paramètre de position tandis que α est un paramètre de forme.
b. La queue de distribution de la loi de Pareto est en loi de puissance $x^{-\alpha}$, par construction.
c. Dans le carré $[0,1]^2$, l'aire entre la première bissectrice et la courbe de B en fonction de A, divisée par l'aire du triangle sous la première bissectrice qui représente la maximalité, donne l'indice de Gini, réel entre 0 et 1, utilisé en économie pour quantifier les inégalités. L'égalité parfaite correspond à la première bissectrice ($B = A$) et à un indice de Gini égal à 0.

Exemple 5.1.3. Distributions à queue lourde voire en loi de puissance.

Les distributions à queue lourde et en particulier en lois de puissance sont très répandues.

- Si $X \sim \mathrm{Cauchy}(0,1)$ alors $\mathbb{P}(X > x) = \frac{1}{2} - \frac{1}{\pi}\arctan(x)$, et comme $\arctan(x) + \arctan(1/x) = \pi/2$, on a $\mathbb{P}(X > x) \sim_{x\to+\infty} 1/(\pi x)$, qui est asymptotiquement en loi de puissance, donc à queue lourde.
- En statistique, la studentisation, qui consiste à normaliser la moyenne empirique par l'écart-type empirique, conduit à la loi de Student, qui est à queue lourde polynomiale avec une puissance dépendant de n. Dans le même esprit, si X et Y sont indépendantes et de loi gaussienne $\mathcal{N}(0,1)$ alors X/Y suit la loi de Cauchy$(0,1)$.
- En géométrie, l'image de la loi uniforme sur le cercle par projection stéréographique est une loi de Cauchy. Plus généralement, en dimension n, on obtient une loi de type Student radialement.
- Si $S_n = X_1 + \cdots + X_n$ est une marche aléatoire sur $\mathbb{Z}$ d'incréments $(X_n)_{n\geq 1}$ i.i.d. de loi $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, issue de l'origine $S_0 := 0$, alors le premier temps de retour à l'origine $T = \inf\{n \geq 1 : S_n = 0\}$ vérifie $\mathbb{P}(T > t) \sim_{t\to\infty} c/\sqrt{t}$ pour tout $t > 0$, avec $c := \sqrt{2/\pi}$, cf. TD.
- La queue de distribution des degrés dans certains modèles de graphes aléatoires à attachement préférentiel est en loi de puissance, cf. TD.

Théorème 5.1.4. Queues en loi de puissance et invariance d'échelle.

Soit μ une loi sur $\mathbb{R}$ de fonction de répartition F. Le deux propriétés suivantes sont équivalentes :

(i) μ a une queue en loi de puissance : il existe $x_0 > 0$, $c \geq 0$, $\alpha > 0$, tels que

$$1 - F(x) = cx^{-\alpha} \quad \text{pour tout } x \geq x_0.$$

(ii) μ a une invariance d'échelle : il existe $x_0 > 0$ et $g : \mathbb{R} \to (0, +\infty)$ continue tels que

$$1 - F(\lambda x) = g(\lambda)(1 - F(x)) \quad \text{pour tous } x, \lambda \text{ avec } \lambda x \geq x_0.$$

— L'équivalence de (i) et (ii) indique que si (ii) a lieu alors forcément $g(\lambda) = \lambda^{-\alpha}$ pour un $\alpha > 0$.
— La loi Pareto(x_0, α) a une queue en loi de puissance α
— Cauchy(x_0, γ) a une queue asymptotiquement en loi de puissance 1, cf. variation régulière plus loin.
— Si $X \sim$ Cauchy, alors e^X (log-Cauchy) a une queue logarithmique, plus lourde que toute loi de puissance!

Démonstration. L'implication (i)⇒(ii) est immédiate, avec $g(\lambda) = \lambda^{-\alpha}$. Démontrons que (ii)⇒(i). On peut supposer sans perte de généralité que $1 - F(x) > 0$ pour tout $x \geq x_0$ car toute annulation se propage sur la droite par (ii), ce qui donnerait (i) avec $c = 0$. Fixons à présent $x, y > 0$, et considérons z assez grand pour que $z, zx, zxy \geq x_0$. Le (ii) donne, avec $\lambda = xy$, $1-F(xyz) = g(xy)(1-F(z))$, mais aussi $1-F(xyz) = g(x)g(y)(1-F(z))$ avec $\lambda = x$ puis $\lambda = y$. Comme $1 - F(z) > 0$, il vient que

$$g(xy) = g(x)g(y) \quad \text{pour tous } x, y > 0,$$

qui peut s'écrire $f(x+y) = f(x) + f(y)$ avec le changement de fonction $f = \log(g)$. Les seules solutions continues, positives, et non-identiquement nulles de cette équation fonctionnelle sont de la forme $f(x) = \alpha x$ avec $\alpha \in \mathbb{R}$ c'est-à-dire $g(x) = x^{-\alpha}$ avec $\alpha \in \mathbb{R}$. Comme $1 - F$ décroît et tend vers 0 en $+\infty$ on obtient $\alpha > 0$, et donc $1 - F(x) = cx^{-\alpha}$ pour $x \geq x_0$ et des constantes $c, \alpha > 0$, qui est bien (i). □

Me 12/03

Vu en cours :
— Pré-écrire d'abord au tableau les formules Fourier des théorèmes 5.1.9 et 5.2.2
— L'intégralité du chapitre à partir d'ici.
— Début du chapitre sur les matrices aléatoires (matrices de covariance empirique).

Théorème 5.1.5. Invariance d'échelle asymptotique et variation régulière.

Soit μ une loi sur $\mathbb{R}$ de fonction de répartition F. Le deux propriétés suivantes sont équivalentes :

(i) μ a une queue $1 - F$ à variation régulière : il existe $\rho \leq 0$ tel que pour tout $y > 0$,

$$\lim_{x \to +\infty} \frac{(1-F)(xy)}{(1-F)(x)} = y^\rho.$$

(ii) μ a une invariance d'échelle asymptotique : il existe $g : (0, +\infty) \to (0, +\infty)$ continue telle que

$$\lim_{x \to +\infty} \frac{(1-F)(\lambda x)}{(1-F)(x)} = g(\lambda) \quad \text{pour tout } \lambda > 0.$$

— On dit alors que μ est une distribution à variation régulière d'indice ρ.
— La loi Pareto$(x_{\min}, \alpha)$ est à variation régulière d'indice $-\alpha$.
— La loi Cauchy(x_0, γ) est à variation régulière d'indice -1.
— Pour les queues de distributions $1 - F$, le concept de variation régulière fait sens pour les indices $\rho \leq 0$, tandis que pour les fonctions plus générales, il fait sens plus généralement pour tout $\rho \in \mathbb{R}$.

Démonstration. (i)⇒(ii) Immédiat! (ii)⇒(i) Fixons $x, y > 0$. Le (ii) donne

$$\lim_{z \to \infty} \frac{1-F(xyz)}{1-F(z)} = g(xy) \quad \text{et} \quad \frac{1-F(xyz)}{1-F(z)} = \frac{1-F(xyz)}{1-F(xz)} \frac{1-F(xz)}{1-F(z)} \xrightarrow[x \to \infty]{} g(y)g(x),$$

d'où $g(xy) = g(x)g(y)$ pour tous $x, y > 0$, et comme g est continue, on obtient $g(x) = x^\rho$, $\rho \in \mathbb{R}$, qui est le (i). □

Théorème 5.1.6. Variation régulière et loi de puissance asymptotique.

Pour toute fonction $G : \mathbb{R}_+ \to \mathbb{R}_+$ et tout $\rho \in \mathbb{R}$, ces propriétés sont équivalentes :

(i) G est à variation régulière d'indice ρ : $\lim_{x \to +\infty} \frac{G(xy)}{G(x)} = y^\rho$ pour tout $y > 0$.

(ii) $G(x) = x^\rho L(x)$ où $L : (0, +\infty) \to (0, +\infty)$ est à variation lente : $\lim_{x \to +\infty} \frac{L(xy)}{L(x)} = 1$ pour tout $y > 0$.

— Les fonctions à variation lente sont les fonctions à variation régulière d'indice 0. Exemples : $L(x) = a > 0$, $L(x) = (\log(1+x))^b$, $a \in \mathbb{R}$, $L(x) = \log(\log(e+x))$, $L(x) = \exp((\log(1+x))^b))$, $0 < b < 1$.

Démonstration. (i)⇒(ii) Soit G à variation régulière d'indice $\rho \in \mathbb{R}$. La fonction $L(x) := G(x)/x^\rho$ est à variation lente car en utilisant le fait que G est à variation régulière,

$$\lim_{x\to+\infty} \frac{L(xy)}{L(x)} = \lim_{x\to+\infty} \frac{G(xy)}{G(x)} \frac{x^\rho}{(xy)^\rho} = 1.$$

(ii)⇒(i). Soit L à variation lente et $\rho \in \mathbb{R}$. La fonction $G(x) := x^\rho L(x)$ est à variation régulière car pour tout $y > 0$,

$$\lim_{x\to+\infty} \frac{G(xy)}{G(x)} = \lim_{x\to+\infty} \frac{(xy)^\rho}{x^\rho} \frac{L(xy)}{L(x)} = y^\rho.$$

□

Lemme 5.1.7. Variation lente et puissances.

Si $L : (0,+\infty) \to (0,+\infty)$ est à variation lente alors $\lim_{x\to+\infty} x^\rho L(x) = 0$ si $\rho < 0$ et $= +\infty$ si $\rho > 0$.

Démonstration. Soient $\rho > 0$, $\lambda > 1$, $\varepsilon > 0$. Il existe $x_{\lambda,\varepsilon}$ tel que $L(\lambda x) \geq (1-\varepsilon)L(x)$ pour tout $x \geq x_{\lambda,\varepsilon}$, donc

$$(\lambda x)^\rho L(\lambda x) \geq \lambda^\rho x^\rho (1-\varepsilon) L(x) \geq c x^\rho L(x),$$

pour $\varepsilon = \varepsilon_{\lambda,\rho}$ assez petit, tout $x \geq x_{\lambda,\rho}$, et $c := \lambda^{\rho/2} > 1$, ce qui donne $(\lambda^n x)^\rho L(\lambda^n x) \geq c^n x^\rho L(x)$ pour tout entier $n \geq 1$, d'où $\lim_{x\to+\infty} x^\rho L(x) = +\infty$. Enfin $\lim_{x\to+\infty} x^{-\rho} L(x) = 0$ car $1/L$ est aussi à variation lente. □

Corollaire 5.1.8. Distribution à variation régulière ⇒ queue lourde.

Si une distribution sur $\mathbb{R}_+$ est à variation régulière alors elle est à queue lourde.

Démonstration. Si F est la fonction de répartition d'une distribution à variation régulière, le théorème 5.1.6 donne $\alpha \geq 0$ et une fonction à variation lente L tels que $1 - F(x) = x^{-\alpha} L(x)$. Donc pour tout $\lambda > 0$ et $\beta > \alpha$,

$$\frac{1-F(x)}{\mathrm{e}^{-\lambda x}} = x^{-\beta} \mathrm{e}^{\lambda x} (x^{\beta-\alpha} L(x)) \xrightarrow[x\to+\infty]{} +\infty \quad \text{car } \lim_{x\to+\infty} x^{-\beta} \mathrm{e}^{\lambda x} = +\infty \text{ et } \lim_{x\to+\infty} x^{\beta-\alpha} L(x) = +\infty,$$

où $\lim_{x\to\infty} x^\rho L(x) = +\infty$ avec $\rho = \beta - \alpha > 0$ provient du fait que L est à variation lente (lemme 5.1.7). □

Théorème 5.1.9. taubérien de Pitman.

Si L est à variation lente et $\alpha \in (0,2)$, alors pour toute v.a.r. X, ces deux propositions sont équivalentes :

(i) $\mathbb{P}(|X| > x) \sim_{x\to+\infty} x^{-\alpha} L(x)$.

(ii) $1 - \Re\varphi_X(t) \sim_{t\searrow 0} \dfrac{\pi}{2\Gamma(\alpha)\sin\frac{\pi\alpha}{2}} t^\alpha L\Big(\dfrac{1}{t}\Big) =: t^\alpha \widetilde{L}\Big(\dfrac{1}{t}\Big)$ où $\varphi_X(t) := \mathbb{E}(\mathrm{e}^{\mathrm{i}tX})$.

— Cela caractérise quantitativement la queue de distribution via le comportement de la fonction caractéristique au voisinage de l'origine. La preuve du théorème 5.1.9 est laborieuse, cf. [63] et [12, p. 336]. Plusieurs variantes sont disponibles. D'autres théorèmes taubériens sont présentés dans [60, Sec. 2.3.2].

— Pour une v.a.r. la finitude des moments d'ordre k est équivalente à la dérivabilité k fois en 0 de la fonction caractéristique. Pour les v.a.r. à queue lourde, et notamment à variation régulière d'indice $-\alpha$, $\alpha \in (0,2)$, le moment d'ordre deux est infini, et la fonction caractéristique n'est pas deux fois dérivable en 0, et n'est pas même dérivable en 0 lorsque $\alpha \in (0,1]$. Cependant, le théorème taubérien de Pitman indique que le comportement asymptotique en 0 est toujours relié à la queue de distribution.

— Le cas $\alpha = 2$ est critique car il correspond pour la fonction caractéristique au cas gaussien, qui n'a pas une queue à variation régulière d'indice $-\alpha = -2$: elle n'est pas à queue lourde.

— Les théorèmes abéliens et taubériens, nommés en l'honneur de Niels Henrik Abel et Alfred Tauber, donnent des conditions reliant deux manières de sommer des séries divergentes. Le théorème abélien de base affirme que si une série converge alors sa somme d'Abel converge vers la même limite. Le théorème taubérien de base est une sorte de réciproque qui affirme que si la série d'Abel d'une série converge et si ses cofficients sont assez petits, alors la série converge vers la même limite.

5.2 Lois stables

Théorème 5.2.1. Lois stables.

Soit μ une mesure de probabilité sur $\mathbb{R}$. Ces propriétés sont équivalentes :

(i) Pour tout $n \geq 1$ et $X_1, \dots, X_n$ v.a.r. i.i.d. de loi μ, il existe des constantes $c_n > 0$ et $d_n \in \mathbb{R}$ telles que

$$X_1 + \cdots + X_n \overset{\text{loi}}{=} c_n X_1 + d_n.$$

(ii) Il existe $(X_n)_{n\geq 1}$ v.a.r. i.i.d. et des suites déterministes réelles $(a_n)_{n\geq 1}$, $a_n > 0$, et $(b_n)_{n\geq 1}$, telles que

$$\frac{X_1 + \cdots + X_n - b_n}{a_n} \xrightarrow[n\to\infty]{\text{loi}} \mu.$$

— On dit alors que μ est stable (par convolution, à translation et dilatation près).
— C'est le cas des (translatées et dilatées) de la loi gaussienne et de la loi de Cauchy sur $\mathbb{R}$.

Démonstration. (i)⇒(ii). Soit $(X_n)_{n\geq 1}$ des v.a.r. i.i.d. de loi μ. Pour tout $n \geq 1$, par (i), il existe des constantes $c_n > 0$ et d_n telles que $X_1 + \cdots + X_n \overset{\text{loi}}{=} c_n X_1 + d_n$, c'est-à-dire $\frac{X_1+\cdots+X_n-d_n}{c_n} \overset{\text{loi}}{=} X_1 = \mu$, d'où (ii).

(ii)⇒(i). Lemme de Gnedenko – Kolmogorov[2] : si $Z_m \xrightarrow{\text{loi}} Z$ et $Z'_m := \alpha_m Z_m + \beta_m \xrightarrow{\text{loi}} Z'$ pour des réels $\alpha_m > 0$ et β_m et Z et Z' non-constantes, alors il existe des réels $\alpha > 0$ et β tels que $Z' \overset{\text{loi}}{=} \alpha Z + \beta$.

Par (ii), $Y_m := \frac{X_1+\cdots+X_m-b_m}{a_m} \xrightarrow{\text{loi}} \mu$ quand $m \to \infty$. Fixons $n \geq 1$. Comme la v.a. $Z'_m := \frac{X_1+\cdots+X_{mn}-nb_m}{a_m}$ est la somme de n v.a.r. i.i.d. de même loi que Y_m, il vient que $Z'_m \xrightarrow{\text{loi}} \mu^{*n}$ quand $m \to \infty$. Or Z'_m s'obtient à partir de $Z_m := Y_{mn}$ par translation et dilatation, tandis que $Z_m \xrightarrow{\text{loi}} \mu$. Le lemme de Gnedenko – Kolmogorov entraîne alors que μ^{*n} et l'image de μ par une transformation affine, et cela pour tout $n \geq 1$, ce qui est bien (i). □

Théorème 5.2.2. Fonction caractéristique des lois stables.

Une v.a.r. X non-constante est stable[a] ssi $X \overset{\text{loi}}{=} aZ + b$ où $a > 0$ et b sont des constantes et Z v.a.r. t.q.

$$\varphi_Z(t) := \mathbb{E}(e^{itZ}) = \exp\Big(-|t|^\alpha (1 - i\beta\gamma_{t,\alpha}\operatorname{sign}(t))\Big), \quad t \in \mathbb{R},$$

où $\alpha \in (0,2]$, $\beta \in [-1,1]$, $\gamma_{t,\alpha} := \begin{cases} \tan(\frac{\pi\alpha}{2}) & \text{si } \alpha \neq 1 \\ -\frac{2}{\pi}\log|t| & \text{si } \alpha = 1 \end{cases}$, et $\operatorname{sign}(t) := \begin{cases} -1 & \text{si } t < 0 \\ 0 & \text{si } t = 0 \\ 1 & \text{si } t > 0 \end{cases}$.

a. Au sens du théorème 5.2.1.

— Le cas gaussien correspond à $\alpha = 2$ et $\beta = 0$, tandis que le cas Cauchy correspond à $\alpha = 1$ et $\beta = 0$.
— On parle de loi α-stable.
— En particulier, dans le cas symétrique où $\beta = 0$, le (ii) du théorème 5.2.1 a lieu avec $a_n = n^{1/\alpha}$ et $b_n = 0$.
— La loi est symétrique ssi $\beta = 0$ (φ_Z est réelle), et on dit que β est le paramètre d'asymétrie.
— La loi possède une moyenne ssi $\alpha > 1$, et une variance ssi $\alpha = 2$.

Démonstration. La preuve, relativement élémentaire mais longue, se trouve par exemple dans [38, Ch. 17]. □

Théorème 5.2.3. Queue de distribution des lois stables.

Si X est α-stable, $\alpha \in (0,2)$, alors $|X|$ est à variation régulière d'indice $-\alpha$.

— Le cas $\alpha = 2$ est exclus : la loi gaussienne n'est pas à variation régulière car n'est pas à queue lourde.

Démonstration. Pour simplifier, on ne traite pas le cas $\alpha = 1$. Comme la variation régulière est invariante par translation et dilatation, on peut supposer que $a = 1$ et $b = 0$ dans le théorème 5.2.2. Au vu du théorème 5.1.9 taubérien de Pitman, il suffit d'établir que $1 - \Re\varphi_X(t) \sim_{t\to 0^+} t^\alpha$. Or le théorème 5.2.2 donne, pour $t > 0$,

$$\Re\varphi_X(t) = e^{-t^\alpha}\cos(\eta t^\alpha), \quad \eta := \beta\tan\frac{\pi\alpha}{2}.$$

2. Cf. [44, Section 10]. Une affaire de loi, pas de v.a. En termes plus algébrico-géométriques, on ne sort pas de la classe d'équivalence des images de la loi par translation et dilatation.

Si $\eta = 0$, alors $\Re\varphi_X(t) = \mathrm{e}^{-t^\alpha} = 1 - t^\alpha + o(t^\alpha)$. Si $\eta \neq 0$, alors $\Re\varphi_X(t) = (1 - t^\alpha + o(t^\alpha))\cos(\eta t^\alpha)$, d'où

$$1 - \Re\varphi_X(t) = (1 - \cos(\eta t^\alpha)) + \cos(\eta t^\alpha)\, t^\alpha - \cos(\eta t^\alpha) o(t^\alpha) \sim_{t\to 0^+} t^\alpha.$$

□

5.3 Universalité : TLC stable

Théorème 5.3.1. TLC stable dans le cas symétrique.

Si $(X_n)_{n\geq 1}$ sont des v.a.r. i.i.d. de loi symétrique avec $\mathbb{P}(|X_1| > x) \sim_{x\to+\infty} cx^{-\alpha}$, $c > 0$, $\alpha \in (0,2)$, alors

$$\frac{X_1 + \cdots + X_n}{n^{1/\alpha}} \xrightarrow[n\to\infty]{\text{loi}} Z_\alpha \quad \text{où } Z_\alpha \text{ est de loi } \alpha\text{-stable symétrique.}$$

— Ce théorème exprime un phénomène d'universalité au sens où la loi limite ne dépend de la loi des ingrédients sommés que via l'indice de queue α. Il complète le TLC habituel, qui correspond à $\alpha > 2$.
— Pour une version plus générale, qui n'impose pas la symétrie, et qui contient le cas gaussien, et dont la preuve est plus laborieuse, on renvoie par exemple à [38, Ch. 17], et à la discussion [60, Sec. 5.4].
— En particulier le (i) du théorème 5.2.1 a lieu avec $a_n = n^{1/\alpha}$ et $b_n = 0$.

Démonstration. Comme X_1 est symétrique, sa fonction caractéristique φ_{X_1} est réelle, et comme la partie réelle d'une fonction caractéristique est toujours paire, la fonction φ_{X_1} est donc paire. L'idée est de procéder comme pour le TLC habituel avec un développement en t autour de 0, et la difficulté et que X_1 n'a pas de variance. Cependant le théorème 5.1.9 taubérien de Pitman donne, en observant que $L \equiv 1$,

$$\varphi_{X_1}(t) = \Re\varphi_{X_1}(t) = 1 - (1 - \Re\varphi_{X_1}(t)) = 1 - a|t|^\alpha(1 + o_{t\to 0}(1)) \quad \text{où} \quad a := \frac{\pi}{2\Gamma(\alpha)\sin\frac{\pi\alpha}{2}},$$

et le caractère i.i.d. de $X_1, \ldots, X_n$ donne alors, pour tout $u \in \mathbb{R}$, ce qui précède avec $t = n^{-1/\alpha}u$ donne

$$\varphi_{\frac{X_1+\cdots+X_n}{n^{1/\alpha}}}(u) = \Big(\varphi_{X_1}(n^{-1/\alpha}u)\Big)^n = \Big(1 - a(1 + o_{n\to\infty}(1))\frac{|u|^\alpha}{n}\Big)^n \xrightarrow[n\to\infty]{} \mathrm{e}^{-a|u|^\alpha}.$$

On reconnaît la fonction caractéristique d'une loi α-stable symétrique donnée par le théorème 5.2.2. □

Régimes en fonction de α dans $\mathbb{P}(|X_1| > x) \sim cx^{-\alpha}$:
— $0 < \alpha \leq 1$: pas de moyenne, donc pas de loi des grands nombres (mais TLC stable !)
— $\alpha = 1$: loi de Cauchy, qui est stable
— $1 < \alpha \leq 2$: moyenne finie, donc LGN, mais TLC stable
— $\alpha = 2$: n'est pas le cas gaussien car il s'agit du α de la queue, et non pas de la fonction caractéristique !
— $\alpha > 2$: variance finie, LGN et TLC classique.

5.4 Vols de Lévy, phénomène one-big-jump pour somme et maximum

Les suites de v.a.r. i.i.d. à queue lourde font apparaître de grandes valeurs, parfois qualifiées de catastrophes (les grands riches dans l'approche de Pareto !), qui font que la somme peut être du même ordre que l'un des facteurs : en anglais on parle parfois de one-big-jump ou de one-big-outlier. Plus précisément, si $(X_n)_{n\geq 1}$ sont des v.a.r. i.i.d. $S_n := X_1 + \cdots + X_n$ et $M_n := \max(X_1, \ldots, X_n)$, alors :
— $\frac{M_n}{S_n} \xrightarrow[n\to\infty]{\text{p.s.}} 0$ ssi $\mathbb{E}(|X_1|) < \infty$
— $\frac{M_n}{S_n} \xrightarrow[n\to\infty]{\text{loi}} Y$ non constante ssi $x \mapsto \mathbb{P}(|X_1| > x)$ est à variation régulière d'indice $-\alpha$ avec $\alpha \in (0,1)$.
— $\frac{S_n - n\mathbb{E}(X_1)}{M_n} \xrightarrow[n\to\infty]{\text{loi}} Y$ non constante ssi $x \mapsto \mathbb{P}(|X_1| > x)$ est à variation régulière d'indice $-\alpha$ avec $\alpha \in (1,2)$.
— $\frac{M_n}{S_n} \xrightarrow[n\to\infty]{\mathbb{P}} 1$ ssi $x \mapsto \mathbb{P}(|X_1| > x)$ est à variation régulière d'indice 0.

Une manière d'illustrer le phénomène est de considérer de cas où les X_i sont de loi α-stable, de sorte que

$$\mathbb{P}(S_n \geq x) \approx \mathbb{P}(M_n \geq x)$$

quand $x \gg 1$, car comme $X_1 + \cdots + X_n \overset{\text{loi}}{=} n^{1/\alpha}X_1$ et $1 - F_{X_1}(t) \approx t^{-\alpha}$, il vient

$$\mathbb{P}(S_n \geq x) = 1 - F_{X_1}(n^{-1/\alpha}x) \approx nx^{-\alpha} \quad \text{et} \quad \mathbb{P}(M_n \geq x) = 1 - F_{X_1}(x)^n = 1 - (1 - (1 - F_{X_1}(x)))^n \approx nx^{-\alpha}.$$

Pour en savoir plus : [60, Sec. 3.4.1], [36, Sec. 8.2.4 et Sec. 6.2.6], mais aussi le plus récent [11].

Concernant la somme, les marches aléatoires à incréments à queue lourde sont appelé vols de Lévy. En raison des catastrophes, leur comportement est beaucoup plus erratique que les marches aléatoires dont les incréments ont une variance finie. La limite d'échelle d'un vol de Lévy n'est plus le mouvement brownien mais un processus de Lévy : processus à temps continu et à accroissement indépendants et stationnaires. Cette notion est reliée à la notion de loi infiniment divisible et à la formule de Lévy–Khintchine.

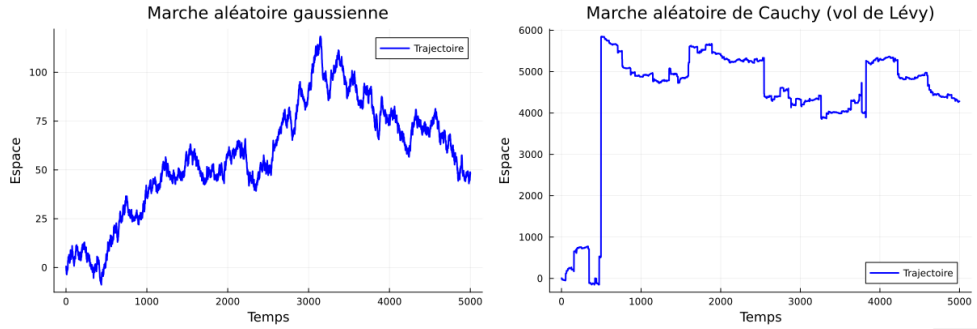


FIGURE 5.1 –
À gauche, une marche aléatoire d'incréments gaussiens, et à droite d'incréments Cauchy (vol de Lévy).

Concernant le maximum, le comportement asymptotique fait apparaître à translation et dilatation près les lois max-stables, analogue des lois stables pour la somme. À translation et dilatation près, il n'y que trois lois max-stables possibles : loi de Gumbel, loi de Fréchet, et loi de Weibull, qui correspondent à la fluctuation asymptotique du maximum de v.a.r. i.i.d. de loi exponentielle, Cauchy, et uniforme. L'analogue pour le maximum du phénomène d'universalité du TLC stable pour la somme est appelé théorème de Fisher–Tippett–Gnedenko. La situation est différente de celle des lois stables des sommes dont le paramétrage est continu.

Une étude du théorème de Cramér en rapport avec les queues lourdes se trouve dans [39].

Les lois à queue lourde jouent un rôle important en modélisation stochastique, mais leur étude peut s'avérer laborieuse. Pour aller plus loin : [38], [36], [12], [68, 67], [49].

Ce chapitre est essentiellement inspiré de [60].

Chapitre 6

Matrices aléatoires, théorème de Wigner, théorème de Marchenko–Pastur

Ce chapitre est consacré à un phénomène d'universalité concernant le comportement asymptotique du spectre de matrices aléatoires de grande dimension. Nous commençons par une motivation statistique, dont le dénouement n'est abordé que dans la section 6.5. Soient $X_1, \dots, X_n$ des vecteurs colonnes aléatoires i.i.d. de $\mathbb{R}^d$ centrés et de matrice de covariance Σ. La matrice Σ est symétrique $d \times d$ et ses valeurs propres sont positives ou nulles. On a $\Sigma_{ij} = \mathbb{E}(X_{ki} X_{kj})$ pour tous $1 \le i, j \le d$ et tout $1 \le k \le n$ ou encore

$$\Sigma = \mathbb{E}(X_1 X_1^\top) = \cdots = \mathbb{E}(X_n X_n^\top).$$

La matrice de covariance empirique $\widehat{\Sigma}_n$ est la matrice aléatoire symétrique $d \times d$ suivante

$$\widehat{\Sigma}_n = \frac{1}{n} \sum_{k=1}^{n} X_k X_k^\top = \frac{1}{n} (X_1 \cdots X_n)(X_1 \cdots X_n)^\top.$$

C'est un estimateur sans biais de Σ : $\mathbb{E}(\widehat{\Sigma}_n) = \Sigma$. Par la LGN appliquée aux $d \times d$ coefficients,

$$\widehat{\Sigma}_n \xrightarrow[n\to\infty]{\text{p.s.}} \Sigma.$$

On souhaite étudier le comportement de la matrice aléatoire $\widehat{\Sigma}_n$ lorsque la dimension des données $d = d_n$ dépend de n et tend vers ∞. Simplifions en posant $\Sigma = I_d$. Cela nous conduit au modèle suivant : on se donne une famille $(Y_{ij})_{i,j\ge 1}$ de v.a.r. i.i.d. de moyenne 0 et de variance 1, et pour tout entier $n \ge 1$, on considère la matrice aléatoire $d_n \times d_n$ symétrique

$$\frac{1}{n} Y Y^\top$$

où $Y = (Y_{ij})_{1\le i\le d_n, 1\le j\le n}$. Si $n \mapsto d_n$ est constante et égale à d alors $\frac{1}{n} Y Y^\top$ converge presque sûrement vers la matrice $\Sigma = I$. Il est naturel de chercher à comprendre le comportement de la matrice aléatoire $\frac{1}{n} Y Y^\top$ lorsque d_n dépend de n, par exemple en étudiant son spectre. L'analyse de la matrice symétrique $Y Y^\top$ est rendue difficile par le fait que ses coefficients sont dépendants. Cela suggère d'analyser en première approche des matrices aléatoires symétriques dont les coefficients sont indépendants dans le triangle supérieur, ce qui conduit au théorème de Wigner. Nous revenons aux matrices de covariance empiriques dans la section 6.5.

6.1 Théorème de Wigner

Je 13/03

Abordé en cours :
- Dessiner au tableau d'abord les exemples de graphes des figures 6.3 et 6.4
- Énoncé du théorème de Wigner et du théorème de Wigner simplifié
- Moments de la loi du demi-cercle, sans la preuve laissée en exercice
- Preuve du théorème de Wigner simplifié par la méthode des moments
- Commentaires sur les généralisations, et introduction de GOE

On considère le tableau aléatoire symétrique infini

$$\begin{pmatrix} M_{1,1} & M_{1,2} & \cdots \\ M_{2,1} & M_{2,2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

où $(M_{ij})_{i\ge j\ge 1}$ sont des v.a.r. et $M_{ji} := M_{ij}$. On se donne un entier $n \ge 1$ et on note $M := (M_{ij})_{1\le i,j\le n}$ la matrice réelle symétrique aléatoire obtenue en extrayant le coin supérieur gauche du tableau. Soient

$$\lambda_{n,1}, \dots, \lambda_{n,n}$$

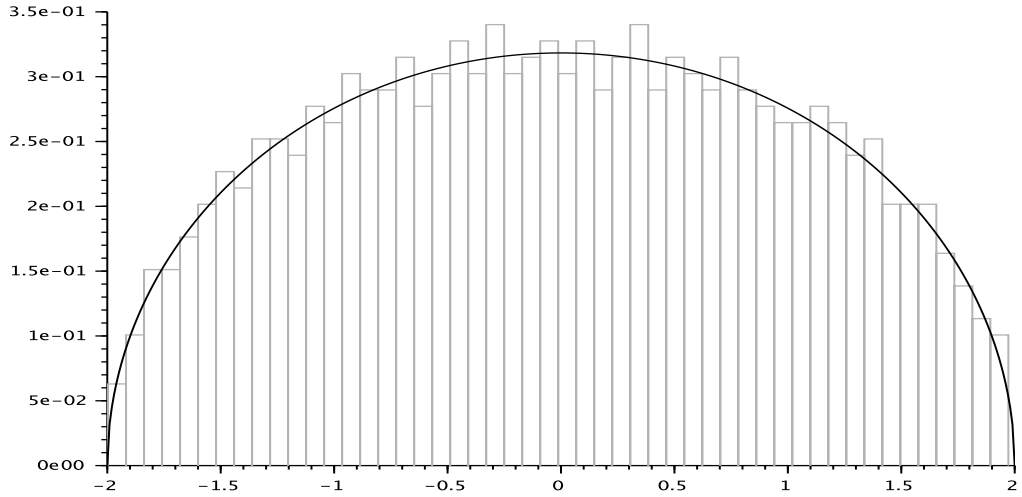


FIGURE 6.1 – Densité de la loi du demi-cercle μ^{DC} et histogramme des valeurs propres d'une matrice symétrique de taille $n = 1000$ dont les coefficients du triangle supérieur, diagonale incluse, sont i.i.d. de loi uniforme sur l'intervalle $[-\sqrt{3/n}, \sqrt{3/n}]$ (la variance vaut donc $1/n$). Les fluctuations de l'histogramme disparaissent quand la dimension $n \to \infty$, pour laisser place à un déterminisme : la loi du demi-cercle de Wigner.

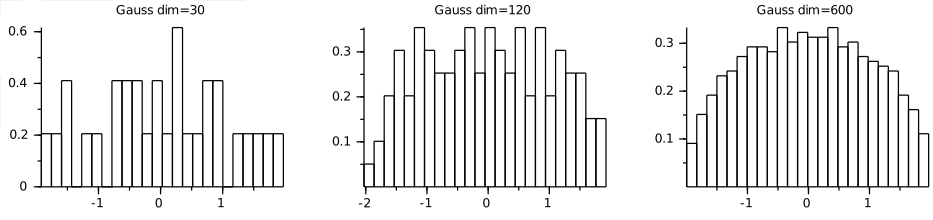


FIGURE 6.2 – Histogramme du spectre d'une matrice du GOE normalisé, de dimension 30, 120, et 600. Les fluctuations disparaissent en grande dimension pour laisser place à un déterminisme : la loi du demi-cercle.

les valeurs propres de la matrice réelle symétrique $\frac{1}{\sqrt{n}}M$, ordonnées de sorte que $\lambda_{n,1} \geq \cdots \geq \lambda_{n,n}$. On s'intéresse à leur mesure de comptage, appelée mesure spectrale empirique, définie par

$$\mu_n := \frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda_{n,k}}.$$

C'est une mesure de probabilité aléatoire, comme la mesure empirique dans le théorème de Glivenko–Cantelli ou de Sanov, mais les atomes $\lambda_{n,1}, \ldots, \lambda_{n,n}$ sont ici des v.a.r. dépendantes [1]. Pour tout borélien $B \subset \mathbb{R}$,

$$\mu_n(B) = \frac{\mathrm{Card}\{1 \leq k \leq n : \lambda_{n,k} \in B\}}{n}$$

est la proportion de valeurs propres appartenant à B. Pour tout $f : \mathbb{R} \to \mathbb{R}$ mesurable, on a

$$\int f \mathrm{d}\mu_n = \frac{1}{n}\sum_{k=1}^{n} f(\lambda_{n,k}).$$

La fonction de répartition F_n de μ_n est donnée pour tout $x \in \mathbb{R}$ par

$$F_n(x) := \mu_n(]-\infty, x]) = \int \mathbb{1}_{]-\infty,x]} \mathrm{d}\mu_n = \frac{\mathrm{Card}\{1 \leq k \leq n : \lambda_{n,k} \leq x\}}{n}.$$

Les quantités $\mu_n(B)$, $\int f \mathrm{d}\mu_n$, et $F_n(x)$ sont des v.a.r.

Théorème 6.1.1. Wigner ou loi du demi-cercle.

Considérons la matrice aléatoire symétrique $M = (M_{ij})_{1 \leq i,j \leq n}$, $M_{ji} := M_{ij}$, où
- Les v.a.r. $(M_{ij})_{j \geq i \geq 1}$ sont indépendantes.
- Les v.a.r. $(M_{ii})_{i \geq 1}$ sont i.i.d. de carré intégrable : $\mathbb{E}(M_{11}^2) < \infty$.
- Les v.a.r. $(M_{ij})_{j > i \geq 1}$ sont i.i.d. centrées et réduites : $\mathbb{E}(M_{12}) = 0$, $\mathbb{E}(M_{12}^2) = 1$.

Soient $\lambda_{n,1} \geq \cdots \geq \lambda_{n,n}$ les valeurs propres de $\frac{1}{\sqrt{n}}M$. Alors presque sûrement,

$$\mu_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda_{n,k}} \xrightarrow[n\to\infty]{\mathscr{C}_b} \mu^{\mathrm{DC}}$$

où μ^{DC} est la loi du demi-cercle sur $[-2,2]$ de densité

$$x \mapsto \frac{\sqrt{4-x^2}}{2\pi}\mathbb{1}_{[-2,2]}(x).$$

En d'autres termes, p.s. pour toute fonction $f : \mathbb{R} \to \mathbb{R}$ continue et bornée,

$$\int f \mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int f \mathrm{d}\mu^{\mathrm{DC}}.$$

- Analogue de la LGN pour un tableau triangulaire de v.a.r. dépendantes : $(\lambda_{n,1}, \ldots, \lambda_{n,n})_{n \geq 1}$.
- Le terme loi du demi-cercle désigne à la fois la loi limite μ^{DC} et le théorème de Wigner, car il constitue l'expression d'un phénomène d'universalité, une loi de la nature, comme la loi des grands nombres.
- Le nom loi du demi-cercle utilisé pour désigner μ^{DC} vient du fait que la densité est l'équation d'un demi-cercle, à une constante de normalisation près égale à la moitié de l'aire du cercle. La loi du demi-cercle μ^{DC} est aussi une loi Beta$(\frac{3}{2}, \frac{3}{2})$ sur $[-2,2]$, car $\sqrt{4-x^2} = (2-x)^{\frac{3}{2}-1}(2+x)^{\frac{3}{2}-1}$, image par l'application affine $t \in [0,1] \mapsto x(t) := 2(2t-1)$ de la loi Beta sur $[0,1]$ de densité proportionnelle à $\sqrt{t(1-t)}$. La densité de μ^{DC} en coordonnées trigonométriques est $\frac{4}{\pi}\sin(\theta)^2 \mathbb{1}_{[0,\pi/2]}(\theta)\mathrm{d}\theta$. La loi μ^{DC} est connue sous le nom de distribution de Sato–Tate en théorie des nombres, et décrit le comportement de courbes elliptiques. Elle y apparaît à travers SU(2), difféomorphe à $\mathbb{S}^3 \subset \mathbb{R}^4$, dont la loi uniforme projetée sur un diamètre donne la loi μ^{DC} (principe d'Archimède!). Enfin la loi μ^{DC} est un cas particulier de profil de Barenblatt, stable par l'équation de la chaleur non-linéaire $\partial_t u = \Delta(u^m)$, $m = 3$ (équation des milieux poreux).
- Comme μ^{DC} n'a pas d'atomes, le théorème de Wigner affirme que p.s. pour tout intervalle $I \subset \mathbb{R}$,

$$\lim_{n\to\infty} \mu_n(I) = \mu^{\mathrm{DC}}(I) = \int_{I \cap [-2,2]} \frac{\sqrt{4-x^2}}{2\pi}\mathrm{d}x.$$

1. Fonctions non-linéaires de M : racines du polynôme caractéristique!

La proportion de valeurs propres de M hors de l'intervalle $[-2\sqrt{n}, 2\sqrt{n}]$ tend vers zéro quand $n \to \infty$.

— Le théorème de Wigner met en lumière un phénomène d'universalité, en ce sens que la loi limite μ^{DC} ne dépend pas de la loi des coefficients de la matrice M. De ce point de vue, le théorème de Wigner, qui est une LGN pour variables dépendantes, fait aussi penser au TLC !

— Le comportement de la moyenne de μ_n peut être compris en observant que par la LGN,

$$\int x \mathrm{d}\mu_n(x) = \frac{1}{n}\sum_{k=1}^{n} \lambda_{n,k} = \frac{1}{n}\mathrm{Tr}\Big(\frac{1}{\sqrt{n}}M\Big) = \frac{1}{n^{3/2}} \sum_{1\le i\le n} M_{ii} \xrightarrow[n\to\infty]{\text{p.s.}} 0 = \int x \mathrm{d}\mu^{\mathrm{DC}}(x).$$

— La normalisation en $1/\sqrt{n}$ de M peut être comprise à son tour en observant qu'à nouveau par la LGN,

$$\int x^2 \mathrm{d}\mu_n(x) = \frac{1}{n}\sum_{k=1}^{n} \lambda_{n,k}^2 = \frac{1}{n}\mathrm{Tr}\Big(\frac{1}{\sqrt{n}}M^2\Big) = \frac{1}{n^2} \sum_{1\le i,j\le n} M_{ij}^2 \xrightarrow[n\to\infty]{\text{p.s.}} 1 = \int x^2 \mathrm{d}\mu^{\mathrm{DC}}(x).$$

Les fonctions tests x et x^2 permettent de passer, via les traces de puissances, de l'objet d'intérêt μ_n fait de variables dépendantes $\lambda_{n,k}$, au modèle M fait de constituants i.i.d. justiciables de la LGN classique. Mais la convergence pour les fonctions test $\mathscr{C}_b$ n'est pas la convergence pour les fonctions tests polynômiales.

— Une autre manière de comprendre la normalisation en $1/\sqrt{n}$ consiste à observer qu'elle permet de stabiliser en grande dimension la norme euclidienne de chaque ligne et de chaque colonne (via LGN !).

6.2 Approche combinatoire : méthode des moments

Cette section est dédiée à une démonstration d'une version simplifiée du théorème de Wigner, en utilisant comme fonctions tests les polynômes (méthode des moments) pour un modèle matriciel simplifié. Des méthodes de réduction à cette version simplifiée sont disponibles, et certaines sont abordées en TD :

— Hypothèses sur M : techniques de troncature et recentrage, d'ablation de la diagonale [2] : inégalité de Hoffman–Wielandt, inégalité de rang découlant des formules de Courant–Fischer.

— Loi $\mathbb{E}\mu_n$ au lieu de μ_n : techniques d'exploitation du phénomène de concentration de la mesure : contrôle de la variance par la méthode des moments, inégalité de Poincaré ou de log-Sobolev si disponible, etc.

— Fonctions tests : techniques pour relier convergence étroite et convergence des moments.

Théorème 6.2.1. de Wigner simplifié.

Soit $M = (M_{ij})_{1\le i,j\le n}$ une matrice réelle symétrique aléatoire, à diagonale nulle $M_{11} = \cdots = M_{nn} = 0$, telle que les v.a.r. $(M_{ij})_{1\le i<j\le n}$ sont i.i.d. de moyenne $\mathbb{E}(M_{ij}) = 0$, variance $\mathbb{E}(M_{ij}^2) = 1$, bornées $|M_{ij}| \le C$. Soit $\mu_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda_{n,k}}$ la mesure empirique des valeurs propres de $\frac{1}{\sqrt{n}}M$. Alors pour tout $r \ge 0$,

$$\lim_{n\to\infty} \mathbb{E}\int x^r \mathrm{d}\mu_n(x) = \int x^r \mathrm{d}\mu^{\mathrm{DC}}(x).$$

Démonstration. Identifions tout d'abord les moments de la limite.

Lemme 6.2.2. Moments de la loi du demi-cercle.

Les moments impairs de μ^{DC} sont nuls tandis que les moments pairs sont les nombres de Catalan :

$$\int x^{2r+1} \mathrm{d}\mu^{\mathrm{DC}}(x) = 0 \quad \text{et} \quad \int x^{2r} \mathrm{d}\mu^{\mathrm{DC}}(x) = \frac{1}{r+1}\binom{2r}{r}, \quad \text{pour tout entier } r \ge 0.$$

En particulier μ^{DC} a pour moyenne 0 et variance 1.

Démonstration. Les moments impairs sont nuls car μ^{DC} est symétrique. Pour calculer les moments pairs, on a

$$\int x^{2r} \mathrm{d}\mu^{\mathrm{DC}}(x) = \frac{1}{\pi}\int_0^2 x^{2r}\sqrt{4-x^2}\mathrm{d}x,$$

2. Le fait que la diagonale ne compte pas pour l'analyse asymptotique de la mesure spectrale peut surprendre un néophyte. Cela s'explique par le fait que ces matrices ne sont pas à diagonale dominante : le spectre n'est pas contrôlé par les valeurs diagonales.

par parité, par le changement de variable $x = 2\cos(u)$ et une intégration par parties,

$$\int_0^2 x^{2r}\sqrt{4-x^2}\mathrm{d}x = 4^{r+1}\int_0^{\frac{\pi}{2}} \cos^{2r}(u)\sin^2(u)\mathrm{d}u = 4^{r+1}(W_{2r} - W_{2r+2}) \quad \text{où} \quad W_n := \int_0^{\frac{\pi}{2}} \cos^n(u)\mathrm{d}u.$$

Or W_n est une classique intégrale de Wallis calculable par récurrence sur n par intégration par parties :

$$W_n = W_{n-2} - \int_0^{\frac{\pi}{2}} \cos^{n-2}(u)\sin^2(u)\mathrm{d}u = W_{n-2} - \frac{1}{n-1}W_n.$$

Cela donne $W_n = \frac{n-1}{n}W_{n-2}$, puis $W_{2r} = \frac{(2r-1)(2r-3)\cdots 1}{2r(2r-2)\cdots 2}W_0 = \frac{(2r)!}{2^{2r}(r!)^2}\frac{\pi}{2}$, d'où le résultat □

Démontrons à présent le théorème 6.2. Les moments de μ^{DC} sont donnés par le lemme 6.2.2. Pour tout entier $r \geq 1$, le moment d'ordre r de $\mathbb{E}\mu_n$ s'écrit (avec $i_{r+1} := i_1$)

$$\mathbb{E}\int x^r \mathrm{d}\mu_n(x) = \frac{1}{n}\mathbb{E}\Big(\sum_{k=1}^n \lambda_{n,k}^r\Big) = \frac{1}{n^{1+r/2}}\mathbb{E}\Big(\mathrm{Tr}(M^r)\Big) = \frac{1}{n^{1+r/2}}\sum_{1\leq i_1,\ldots,i_r\leq n} \mathbb{E}(M_{i_1 i_2}\cdots M_{i_r i_{r+1}}).$$

Comme les coefficients diagonaux de M sont nuls (centrés suffit ici), on a

$$\mathbb{E}\int x\mathrm{d}\mu_n(x) = \frac{1}{n^{3/2}}\sum_{i=1}^n \mathbb{E}M_{ii} = 0 = \int x\mathrm{d}\mu^{\mathrm{DC}}(x),$$

et comme M a des coefficient hors diagonale de variance unité,

$$\mathbb{E}\int x^2\mathrm{d}\mu_n(x) = \frac{1}{n^2}\sum_{1\leq i,j\leq n} \mathbb{E}(M_{ij}^2) = \frac{n^2-n}{n^2} \to 1 = \int x^2\mathrm{d}\mu^{\mathrm{DC}}(x).$$

L'étude du moment d'ordre 3 est un peu plus subtile. On a

$$\mathbb{E}\int x^3\mathrm{d}\mu_n(x) = \frac{1}{n^{1+3/2}}\sum_{1\leq i,j,k\leq n} \mathbb{E}(M_{ij}M_{jk}M_{ki}).$$

Si deux éléments parmi $\{\{i,j\},\{j,k\},\{k,i\}\}$ sont distincts alors on a forcément $\mathbb{E}(M_{ij}M_{jk}M_{ki}) = 0$ par indépendance et centrage. Dans le cas contraire, on a $i = k$ ou $i = j$ ou $k = j$, d'où $\mathbb{E}(M_{ij}M_{jk}M_{ki}) = 0$ car la diagonale de M est nulle. Ainsi, le moment d'ordre 3 de $\mathbb{E}\mu_n$ est égal à zéro, tout comme le moment d'ordre 3 de μ^{DC}.

Pour le moment d'ordre 4, on pourrait procéder de même en utilisant

$$\mathbb{E}\int x^4\mathrm{d}\mu_n(x) = \frac{1}{n^{1+4/2}}\sum_{1\leq i,j,k,l\leq n} \mathbb{E}(M_{ij}M_{jk}M_{kl}M_{li}).$$

Un phénomène nouveau a lieu : le cas $i = k$ et $j = l$ avec $i \neq j$ donne $\mathbb{E}(M_{ij}M_{jk}M_{kl}M_{li}) = \mathbb{E}(M_{ij}^4)$, qui n'est pas fonction des deux premiers moments des coefficients de M. Il n'y a pas de contradiction : ces termes n'ont pas de contribution asymptotique quand $n \to \infty$ car leur nombre, $n(n-1)$, est négligeable devant $n^{1+4/2} = n^3$. Le phénomène d'universalité vient de cette non contribution des moments d'ordre > 2 des coefficients de M.

Plus généralement, pour tout entier $r > 1$, on peut supposer que $i_k \neq i_{k+1}$ pour tout $1 \leq k \leq r$ car M a une diagonale nulle. On associe à chaque r-uplet d'indices de ce type $i_1, \ldots, i_r$ un multi-graphe orienté :

- Voir figures 6.3 et 6.4 pour des exemples.
- Les sommets sont les valeurs distinctes prises par ces indices, et on note t leur nombre.
- Les arêtes sont les liaisons (i_k, i_{k+1}) avec $1 \leq k \leq r$, il y en a r. Elles peuvent avoir une multiplicité et sont orientées, et le multi-graphe orienté est connexe et cyclique : on part de i_1 pour y revenir en r étapes.

On dit qu'il s'agit d'un multi-graphe orienté $G(r,t)$. Deux multi-graphes orientés $G(r,t)$ sont équivalents lorsque qu'on peut passer de l'un à l'autre en permutant les indices $\{1,\ldots,n\}$. Des multi-graphes orientés $G(r,t)$ équivalents donnent la même valeur à $\mathbb{E}(M_{i_1 i_2}\cdots M_{i_r i_{r+1}})$, notée $\mathbb{E}(M_G)$, car la loi de la matrice M est invariante par conjugaison par une matrice de permutation. Il y a

$$n(n-1)\cdots(n-t+1)$$

multi-graphes orientés $G(r,t)$ dans chaque classe d'équivalence (nombre d'arrangements de t objets parmi n). Chaque classe d'équivalence contient un représentant pour lequel les t valeurs distinctes prises par les indices $i_1, \ldots, i_r$ sont successivement $1, \ldots, t$. Afin de calculer les contributions, on distingue trois types de classes de multi-graphes orientés $G(r,t)$ détaillés ci-après. Le type est constant sur chaque classe (propriété de la classe).

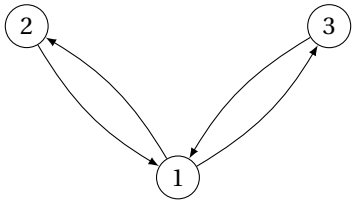


FIGURE 6.3 – Multi-graphe orienté associé à $\mathbb{E}(M_{12}M_{21}M_{13}M_{31})$. On a $r = 4$, $t = 3$, $i_1 = i_3 = 1$, $i_2 = 2$, $i_4 = 3$. Les arêtes successives sont 12, 21, 13, 31. Chaque arête présente, ainsi que l'arête de sens opposé, ne l'est qu'une fois. Le graphe non orienté squelette est l'arbre $2 \leftrightarrow 1 \leftrightarrow 3$. Ce multi-graphe orienté est donc de type 1.

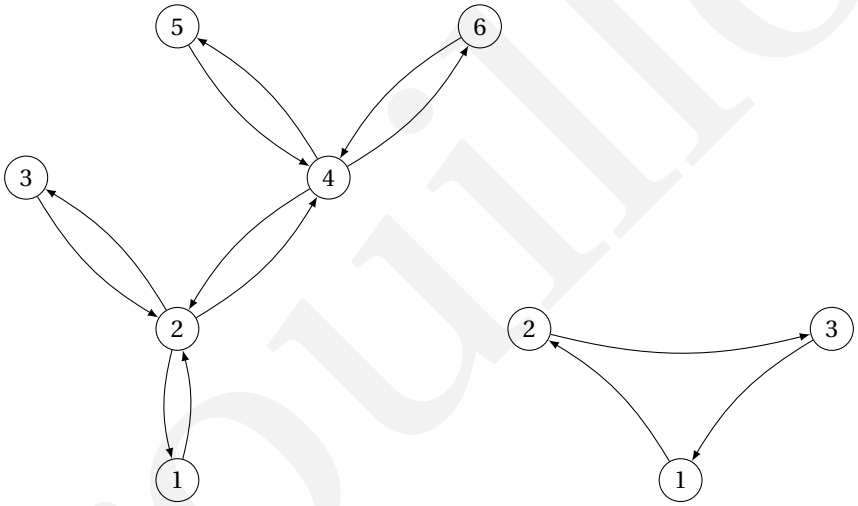
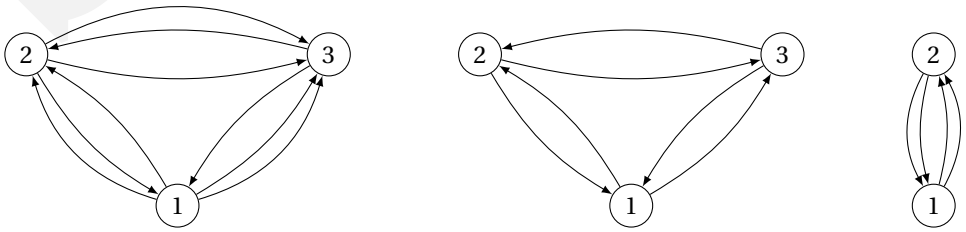


FIGURE 6.4 – En haut à gauche le multi-graphe orienté associé au long exemple $\mathbb{E}(M_{12}M_{23}M_{32}M_{24}M_{45}M_{54}M_{46}M_{64}M_{42}M_{21})$, qui est de type 1. En haut à droite le multi-graphe orienté associé à $\mathbb{E}(M_{12}M_{23}M_{31})$, qui est de type 2. En bas au milieu le multi-graphe orienté associé à $\mathbb{E}(M_{12}M_{21}M_{13}M_{32}M_{23}M_{31})$, et en bas à droite le multi-graphe orienté associé à $\mathbb{E}(M_{12}M_{21}M_{12}M_{21})$, tous deux de type 3 mais pour des raisons différentes : graphe squelette contient un cycle, et sur-multiplicité d'arêtes respectivement. Enfin, en bas à gauche, un exemple de type 3 avec un nombre impair d'arrêtes : $r = 9$. Enfin, il est utile de mentionner que pour les graphes de types 2 et 3, le sens de parcours n'est pas forcément unique.



— Classes de type 1 : chaque arête présente, ainsi que l'arête de sens opposé, ne l'est qu'une fois, et le graphe non-orienté squelette obtenu en effaçant les orientations et les multiplicités des arêtes est un arbre (c'est-à-dire sans cycles). Il s'agit d'arbres planaires en raison de la numérotation des sommets. Exemple : $\mathbb{E}(M_{12}M_{21}M_{13}M_{31}) = \mathbb{E}(M_{12}^2)\mathbb{E}(M_{13}^2) = 1$, voir figures 6.3 et 6.4.
— Classes de type 2 : une arête au moins n'apparaît qu'une seule fois et l'arête de sens opposé n'apparaît pas, comme par exemple $\mathbb{E}(M_{12}M_{23}M_{31}) = \mathbb{E}(M_{12})\mathbb{E}(M_{23})\mathbb{E}(M_{31}) = 0$, voir figure 6.4.
— Classes de type 3 : pas de type 1 ou 2. Exemple : $\mathbb{E}(M_{12}M_{21}M_{12}M_{21}) = \mathbb{E}(M_{12}^4) > 0$ car l'arête 12 (et 21) apparaît exactement deux fois. Autre exemple : $\mathbb{E}(M_{12}M_{21}M_{13}M_{32}M_{23}M_{31}) = \mathbb{E}(M_{12}^2)\mathbb{E}(M_{13}^2)\mathbb{E}(M_{23}^2) = 1$, car le graphe non orienté squelette associé est le cycle $1 \leftrightarrow 2 \leftrightarrow 3 \leftrightarrow 1$. La figure 6.4 les illustre.

Nous abordons à présent la phase finale :

— *Contribution des classes de type 1.* Si G est un multi-graphe orienté de type 1 alors $\mathbb{E}(M_G) = 1$ par indépendance et le fait que les coefficients hors diagonale de M sont de variance unité. Cela ramène le problème à la détermination du nombre N_1 de classes d'équivalences de multi-graphes orientés $G(r,t)$ de type 1. Si r est impair alors $N_1 = 0$. Si r est pair, disons $r = 2a$, alors $t = 1 + a = 1 + r/2$ car le nombre de sommets t d'un arbre est toujours égal à 1 plus le nombre d'arêtes a. Ainsi les classes de type 1 sont en bijection avec les arbres planaires enracinés à a arêtes[3] (et à $1+a$ sommets). Leur nombre vérifie l'équation de récurrence des nombres de Catalan : il y en a donc $N_1 = \frac{1}{a+1}\binom{2a}{a}$, et

$$\sum_{\text{cl. t. 1}} \frac{n(n-1)\cdots(n-t+1)}{n^{1+r/2}}\mathbb{E}(M_G) = \frac{n}{n}\cdots\frac{n-a}{n}\frac{1}{a+1}\binom{2a}{a} \xrightarrow[n\to\infty]{} \frac{1}{1+a}\binom{2a}{a}.$$

— *Les classes de type 2 ont une contribution nulle* car dans ce cas $\mathbb{E}(M_G) = 0$ par indépendance et centrage.
— *Les classes de type 3 ont une contribution asymptotiquement nulle.* En effet, le type 1 nécessite que $2t = r+2$. Si G est de type 3 alors la négation des contraintes des types 1 et 2 conduit après, examination attentive, à $2t < r+2$, c'est-à-dire $2t \le r+1$, d'où $n(n-1)\cdots(n-t+1) \le n^t \le n^{1/2+r/2}$. D'autre part, le nombre de classes d'équivalences de multi-graphes orientés $G(r,t)$ est majoré par $t^r = O(1)$. Comme les coefficients de M sont bornés à valeurs dans $[-C,C]$, on a également $\mathbb{E}(M_G) \le C^r = O(1)$, d'où enfin

$$\sum_{\text{cl. t. 3}} \frac{n(n-1)\cdots(n-t+1)}{n^{1+r/2}}\mathbb{E}(M_G) = O(n^{-1/2}) = o_{n\to\infty}(1).$$

□

6.3 Ensemble Gaussien Orthogonal (GOE)

Me 19/03

Abordé en cours :
— Lemme sur GOE et sa preuve
— Lemme sur Cauchy–Stieltjes commenté mais sans la preuve
— Preuve de Wigner pour GOE par transformée de Cauchy–Stieltjes

Lemme 6.3.1. Ensemble Gaussien Orthogonal (GOE).

Soit $(M_{ij})_{1\le i,j\le n}$ une matrice aléatoire symétrique. Ces propriétés sont équivalentes :

(i) $(M_{ij})_{1\le i\le j\le n}$ sont indépendantes, avec $M_{ii} \sim \mathcal{N}(0,2)$, $1 \le i \le n$, et $M_{ij} \sim \mathcal{N}(0,1)$, $1 \le i < j \le n$.

(ii) M a pour densité $s \mapsto \frac{1}{Z_n}\mathrm{e}^{-\frac{1}{4}\mathrm{Tr}(s^2)}$ avec l'identification $\mathrm{Sym}_{n\times n}(\mathbb{R}) \equiv \mathbb{R}^{\frac{n(n+1)}{2}}$.

De plus $OMO^\top$ a même loi que M, pour toute matrice $n \times n$ orthogonale O.

— Le terme « ensemble » est entre « loi » et « variable aléatoire ». La physique a du forger ses termes et concepts avant la mathématisation moderne des probabilités. Le terme « statistique » était synonyme d'aléatoire. On dirait peut-être « physique stochastique » et « mécanique stochastique ».

Démonstration. Pour l'équivalence : $\mathrm{Tr}(s^2) = \sum_i s_{ii}^2 + 2\sum_{i<j} s_{ij}^2$. Pour l'invariance en loi : la conjugaison par O est linéaire donc de jacobien constant tandis que la cyclicité de la trace rend invariante la densité. □

— Dans l'approche de Wigner pour la physique nucléaire, développée par Mehta, Gaudin, et Dyson, pour mieux rendre compte des écarts entre niveaux d'énergie des noyaux atomiques, l'opérateur est rendu

3. Ne pas confondre avec les arbres binaires planaires enracinés, qui sont également comptés par les nombres de Catalan! Ces nombres comptent une grande variétés d'objets combinatoires : parenthésages, excursions de la marche aléatoire simple, chemins de Dyck, etc.

aléatoire (physique statistique au secours de la mécanique quantique en quelque sorte), ce qui donne lieu à des matrices aléatoires : GOE correspond aux systèmes invariants par renversement du temps.

— Maxwell. GOE est une gaussienne matricielle, une mesure de Boltzmann–Gibbs. Elle possède aussi une caractérisation dans l'esprit du théorème 1.2.7 de Maxwell : c'est la seule loi sur les matrices symétriques qui est à la fois invariante par conjugaison unitaire et qui rend les coefficients indépendants.

— Wick. GOE vérifie le théorème de Wigner. De plus, comme GOE est un vecteur gaussien, la formule de Wick permet d'évaluer $\mathbb{E}(M_{i_1 i_2}\cdots M_{i_n i_1})$, et donc de développer la méthode des moments pour $\mathbb{E}\mu_n$.

— log-Sobolev. GOE est une gaussienne, qui vérifie une inégalité de log-Sobolev, donc une concentration de la mesure sous-gaussienne pour les fonctions Lipschitz. Or le spectre d'une matrice symétrique est une fonction Lipschitz de la matrice. On peut en déduire une concentration de μ_n autour de $\mathbb{E}\mu_n$, cf. TD.

— Grandes déviations. Si $M = ODO^\top$ est la diagonalisation du GOE, alors un calcul de jacobien (géométrie différentielle !) donne que O et D sont indépendantes avec O de loi uniforme sur le groupe orthogonal et $D = \mathrm{diag}(\lambda_1,\dots,\lambda_n)$ avec $\lambda_1 \geq \cdots \geq \lambda_n$ de densité (noter la répulsion entre valeurs propres [4])

$$\frac{1}{Z_n}\exp\Big(-\frac{1}{4}\sum_{i=1}^n \lambda_i^2\Big)\prod_{i<j}(\lambda_i-\lambda_j).$$

Cette loi, vue comme une mesure de Boltzmann–Gibbs, est un gaz de Coulomb [5] :

$$\frac{1}{Z_n}\exp\Big(-\Big(\sum_{i=1}^n \frac{\lambda_i^2}{4} + \sum_{i<j}\log\frac{1}{\lambda_i-\lambda_j}\Big)\Big).$$

Par dilatation, la loi du spectre de $\frac{1}{\sqrt{n}}M$ est donnée par

$$\frac{1}{Z_n}\exp\Big(-n^2\Big(\sum_{i=1}^n \frac{1}{n}\frac{\lambda_i^2}{4} + \frac{1}{n^2}\sum_{i<j}\log\frac{1}{\lambda_i-\lambda_j}\Big)\Big).$$

L'énergie et donc la densité est fonction (quadratique) de la mesure empirique [6] $\frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i}$. Bien que la loi du spectre $(\lambda_{n,1},\dots,\lambda_{n,n})$ de $\frac{1}{\sqrt{n}}M$ ne soit pas produit, il est possible d'établir un théorème de type Sanov pour l'analyse asymptotique en grande dimension : la mesure empirique $\frac{1}{n}\sum_{k=1}^n \delta_{\lambda_{n,k}}$ du spectre de $\frac{1}{\sqrt{n}}M$ vérifie un PGD de vitesse $\frac{1}{n^2}$ et de fonction de taux

$$\mu \mapsto \int \frac{x^2}{4}\mu(\mathrm{d}x) + \iint \log\frac{1}{|x-y|}\mu(\mathrm{d}x)\mu(\mathrm{d}y),$$

et il est possible de déduire de ce PGD le théorème de Wigner pour GOE, via le lemme de Borel–Cantelli.

— Par définition, la constante Z_n ci-dessus varie d'une mesure à l'autre, pour assurer la normalisation à 1.

6.4 Approche analytique : transformée de Cauchy–Stieltjes

La transformée de Cauchy–Stieltjes d'une mesure de probabilité μ sur $\mathbb{R}$ est définie par

$$s_\mu(z) := \int \frac{\mu(\mathrm{d}x)}{x-z}, \quad z \in \mathbb{C}_+ := \{z \in \mathbb{C} : \Im z > 0\}.$$

Comme $z \in \mathbb{C}_+$, l'application $x \in \mathbb{R} \mapsto 1/(x-z)$ est continue et uniformément bornée :

$$\frac{1}{|x-z|} = \frac{1}{\sqrt{(x-\Re z)^2+(\Im z)^2}} \leq \frac{1}{\Im z} \quad \text{d'où en particulier} \quad |s_\mu(z)| \leq \frac{1}{\Im z} \quad \text{qui est uniforme en } \mu.$$

L'intérêt de cette transformée en analyse spectrale vient du fait que si $A \in \mathscr{M}_{n,n}(\mathbb{R})$ est symétrique de spectre $\lambda_1 \geq \cdots \geq \lambda_n$, et si $\mu_A := \frac{1}{n}\sum_{k=1}^n \delta_{\lambda_k}$, alors pour tout $z \in \mathbb{C}_+$, en notant $R_A(z) := (A - zI_n)^{-1}$ la résolvante de A :

$$s_{\mu_A}(z) = \frac{1}{n}\sum_{k=1}^n \frac{1}{\lambda_k - z} = \frac{1}{n}\mathrm{Tr}(R_A(z)).$$

4. Pour une matrice symétrique $\begin{pmatrix} x_1 & x_3 \\ x_3 & x_2 \end{pmatrix}$, l'écart entre valeurs propres est $\geq 2|x_3|$, donc > 0 dès que la matrice n'est pas diagonale.

5. Comme $\log\frac{1}{r}$ est le potentiel électrostatique entre deux charges unité en dimension 2, il faut imaginer n particules de même charge, en dimension 2, sur un rail unidimensionnel, et subissant un champ extérieur confinant de potentiel quadratique.

6. La situation rappelle celle du modèle de Curie–Weiss, mais avec une vitesse n^2 au lieu de n, qui écrase l'entropie (ici le volume).

La transformée s_{μ_A} fait le lien entre la variable spectrale d'intérêt μ_A et la variable A qui porte les hypothèses dans le théorème de Wigner. Elle consiste à tester la mesure spectrale empirique sur la famille de fonctions tests $\{1/(x-z) : z \in \mathbb{C}_+\}$, de même que la méthode des moments repose sur la famille de fonctions test $\{x^r : r \in \mathbb{N}\}$. Une liaison entre s_μ et les moments de μ est la suivante, valable pour z dans $\mathbb{C}_+$ tel que $|z| > \sup\{|x| : x \in \mathrm{supp}(\mu)\}$:

$$s_\mu(z) = -z^{-1}\int \frac{\mu(\mathrm{d}x)}{1-\frac{x}{z}} = -z^{-1}\int \sum_{r=0}^{\infty}\left(\frac{x}{z}\right)^r \mu(\mathrm{d}x) = -\sum_{r=0}^{\infty} z^{-(r+1)}\int x^r \mu(\mathrm{d}x).$$

De plus, cette transformée, tout comme la transformée de Fourier, caractérise la mesure et la convergence :

Lemme 6.4.1. Transformée de Cauchy–Stieltjes : caractérisations.

Soit $\mathscr{P}(\mathbb{R})$ l'ensemble des mesures de probabilités sur $\mathbb{R}$.

(i) Pour tous $\mu, \nu \in \mathscr{P}(\mathbb{R})$: $\mu = \nu$ ssi $s_\mu = s_\nu$

(ii) Pour toute suite $(\mu_n)_{n\geq 1}$ tendue dans $\mathscr{P}(\mathbb{R})$, $\lim_{n\to\infty} s_{\mu_n}(z) = s(z)$ existe pour tout $z \in \mathbb{C}_+$ ssi il existe $\mu \in \mathscr{P}(\mathbb{R})$ telle que $s(z) = s_\mu(z)$ pour tout $z \in \mathbb{C}_+$ et $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ pour tout $f \in \mathscr{C}_b$.

(iii) Pour tout $z \in \mathbb{C}_+$, on a $s_{\mu^{\mathrm{DC}}}(z) = \frac{-z+\sqrt{z^2-4}}{2}$, solution de $s(z)^2 + zs(z) + 1 = 0$ vérifiant $\Im s(z)\Im z > 0$.

— Rappelons qu'une famille $(\mu_i)_{i\in I}$ de mesure de probabilités sur un même espace topologique muni de sa tribu borélienne est tendue lorsque pour tout $\varepsilon > 0$, il existe un compact K_ε tel que $\sup_{i\in I} \mu_i(K_\varepsilon^c) \leq \varepsilon$.
— Par convention on prend la $\sqrt{\cdot}$ à partie imaginaire ≥ 0 (ou partie réelle ≥ 0 si partie imaginaire nulle).
— En fait s_μ est analytique sur $\mathbb{C}_+$. Donc si $s_\mu = s_\nu$ sur un ouvert de $\mathbb{C}_+$, arbitrairement petit, alors l'unicité du prolongement analytique (cours d'analyse complexe) donne $s_\mu = s_\nu$ sur $\mathbb{C}_+$, et donc $\mu = \nu$.

Démonstration.

(i) Pour tout $z \in \mathbb{C}_+$, $\frac{1}{\pi}\Im s_\mu(z)$ est la densité en $\Re z$ de $X + (\Im z)Z$ avec $X \sim \mu$ et $Z \sim \mathrm{Cauchy}(0,1)$ indépendantes :

$$\frac{1}{\pi}\Im s_\mu(z) = \frac{1}{\pi}\int \frac{\Im z}{(\lambda - \Re z)^2 + (\Im z)^2}\mu(\mathrm{d}\lambda) = (\kappa * \mu)(\Re z) \quad \text{où} \quad \kappa(x) := \frac{\Im z}{\pi(x^2 + (\Im z)^2)}.$$

Comme $\varphi_Z(t) = \mathrm{e}^{-|t|} \neq 0$ pour tout $t \in \mathbb{R}$, il en découle que si $s_\mu = s_\nu$ alors $\mu = \nu$. Au passage, cela suggère de concevoir la loi de densité $\frac{1}{\pi}\Im s_\mu$ comme une régularisation de μ : par convergence dominée,

$$X + (\Im z)Z \xrightarrow[\Im z \to 0]{\text{loi}} \mu.$$

(ii) Si $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ pour tout $f \in \mathscr{C}_b$ alors $s_{\mu_n}(z) \to s_\mu(z)$ pour tout $z \in \mathbb{C}_+$ en prenant $f(x) = \frac{1}{x-z}$. Réciproquement, supposons que $(\mu_n)_{n\geq 1}$ est tendue et telle que $s(z) := \lim_{n\to\infty} s_{\mu_n}(z)$ existe pour tout $z \in \mathbb{C}_+$. Le théorème de Prokhorov sur les familles tendues affirme qu'elles sont séquentiellement relativement compactes pour la topologie de la convergence étroite, il suffit donc d'établir l'unicité de la limite de sous-suites convergentes étroitement. Or si $\mu_{n_k} \to \mu$ étroitement avec μ mesure de probabilité, alors $s_{\mu_{n_k}} \to s_\mu$ ponctuellement par la première partie, donc $s_\mu = s$, d'où l'unicité de la limite.

Remarque 6.4.2. Sans théorème de Prokhorov.

Il est possible de se passer du théorème de Prokhorov. En effet, la tension permet de supposer sans perte de généralité que les μ_n sont à support dans un compact fixé $K \subset \mathbb{R}$. Soit $Z \sim \mathrm{Cauchy}(0,1)$ et $X_n \sim \mu_n$ indépendantes. Pour tout $z = x + \mathrm{i}y \in \mathbb{C}_+$ et tout $t \in \mathbb{R}$, par indépendance,

$$\varphi_{X_n+yZ}(t) := \mathbb{E}(\mathrm{e}^{\mathrm{i}t(X_n+yZ)}) = \varphi_{X_n}(t)\mathrm{e}^{-y|t|}$$

tandis qu'au vu de la preuve du (i), par hypothèse et convergence dominée (K est compact)

$$\varphi_{X_n+yZ}(t) = \int \frac{1}{\pi}\Im s_{\mu_n}(x+\mathrm{i}y)\mathrm{e}^{\mathrm{i}tx}\mathrm{d}x \xrightarrow[n\to\infty]{} \int \frac{1}{\pi}\Im s(x+\mathrm{i}y)\mathrm{e}^{\mathrm{i}tx}\mathrm{d}x.$$

Notons que $|\Im s(x+\mathrm{i}y)| \leq |s(z)| = |\lim_{n\to\infty} s_{\mu_n}(z)| \leq 1/\Im z = 1/y$. Il en découle que pour tout $t \in \mathbb{R}$,

$$\varphi_{X_n}(t) \xrightarrow[n\to\infty]{} \mathrm{e}^{y|t|}\int \frac{1}{\pi}\Im s(x+\mathrm{i}y)\mathrm{e}^{\mathrm{i}tx}\mathrm{d}x,$$

et la quantité limite est continue en $t=0$ par convergence dominée. Donc d'après le théorème de continuité de Paul Lévy, il existe une mesure de probabilité μ telle que $\int f\mathrm{d}\mu_n \to \int f\mathrm{d}\mu$ pour tout $f\in\mathscr{C}_b$. En particulier $s_{\mu_n}(z)\to s_\mu(z)$ pour tout $z\in\mathbb{C}_+$, et par unicité de la limite, $s_\mu = s$ sur $\mathbb{C}_+$.

Remarque 6.4.3. Sans hypothèse de tension.

Si $(\mu_n)_{n\geq 1}$ est une suite de mesures de probabilités telle que $s(z) := \lim_{n\to\infty} s_{\mu_n}(z)$ existe pour tout $z\in\mathbb{C}_+$, alors il existe une sous-probabilité μ telle que $\mu_n\to\mu$ vaguement, et μ est une probabilité ssi $\lim_{y\to+\infty} \mathrm{i}y s_{\mu_n}(\mathrm{i}y)=1$, et dans ce cas $\mu_n\to\mu$ étroitement. C'est une sorte d'analogue au critère $\Phi_\mu(0)=1$ de Paul Lévy pour les fonctions caractéristiques. En effet, par le théorème de Helly, il existe une sous-probabilité μ et une sous-suite $(\mu_{n_k})_{k\geq 1}$ telles que $\mu_{n_k}\to\mu$ vaguement. Par la première partie, $s_\mu = s$, puis l'unicité de la limite et la relative compacité séquentielle de l'ensemble des sous-probabilités pour la topologie de la convergence vague donne $\mu_n\to\mu$ vaguement. Cet usage du théorème de Helly permet égalememnt un traitement alternatif lorsque la suite est tendue. Enfin le développement asymptotique de $s_\mu(z)$ en $z=\infty$ en terme de moments reste valable pour une sous-probabilité et donne en particulier $zs_\mu(z)\sim\mu(\mathbb{R})$ quand $z\to\infty$, en particulier μ est une probabilité ssi $\lim_{y\to\infty}\mathrm{i}y s_\mu(\mathrm{i}y)=1$. Dans ce cas $\mu_n\to\mu$ étroitement.

(iii) Il est possible d'utiliser la méthode des résidus du cours d'analyse complexe (exercice!). Alternativement, on peut se ramener aux moments de μ^{DC} fournis par le lemme 6.2.2 : pour tout $z\in\mathbb{C}_+$ avec $|z|>2$,

$$s_{\mu^{\mathrm{DC}}}(z) = -\sum_{r=0}^{\infty} z^{-(r+1)}\int x^r\mu^{\mathrm{DC}}(\mathrm{d}x) = -\sum_{r=0}^{\infty} z^{-(2r+1)}\frac{\binom{2r}{r}}{r+1} = -\frac{1}{z}G\Big(\frac{1}{z^2}\Big)$$

où $G(w) := \sum_{n=0}^{\infty}\frac{\binom{2n}{n}}{n+1}w^n = \sum_{n=0}^{\infty} C_n w^n$ est la fonction génératrice des nombres de Catalan $C_n := \frac{1}{n+1}\binom{2n}{n}$. Ces nombres vérifient la relation de récurrence dite de Segner :

$$C_0 = 1 \quad\text{et}\quad C_{n+1} = \sum_{k=1}^{n} C_k C_{n-k}, \quad n\geq 0.$$

Cette relation se traduit par une équation sur la fonction génératrice : $G(w) = 1 + wG(w)^2$. Cette équation a deux solutions : $\frac{1\pm\sqrt{1-4w}}{2w}$, ce qui donne[7] $G(w) = \frac{1-\sqrt{1-4w}}{2w}$ car l'autre solution ne prend pas la valeur 1 quand $w\to 0$. On a donc au bout du compte $s_{\mu^{\mathrm{DC}}}(z) = -\frac{1}{z}G(\frac{1}{z^2})$ avec $G(w)=\frac{1-\sqrt{1-4w}}{2w}$, pour des z dans un voisinage de ∞, formule qui se généralise à tout $z\in\mathbb{C}_+$ par unicité du prolongement analytique. L'équation sur G donne aussi $s_{\mu^{\mathrm{DC}}}(z)^2 + zs_{\mu^{\mathrm{DC}}}(z)+1=0$. Or l'équation $s^2+zs+1=0$ en s a deux racines $s_\pm(z) := \frac{1}{2}(-z\pm\sqrt{z^2-4})$, et s_- est impossible car pour toute mesure de probabilité μ sur $\mathbb{R}$,

$$\Im s_\mu(z)\Im z = \int\frac{(\Im z)^2}{|x-z|^2}\mu(\mathrm{d}x) > 0, \quad z\in\mathbb{C}_+.$$

□

Examinons à présent comment utiliser cet outil pour établir le théorème de Wigner dans le cas où $M\sim$ GOE.

Résolvante

Si A est une matrices $n\times n$ symétrique réelle, $z\in\mathbb{C}_+$, alors sa résolvante $R_A(z) := (A-zI_n)^{-1}$ est bornée :

$$\|R_A(z)\|_{\mathrm{op}} := \max_{\substack{w\in\mathbb{C}^n\\|w|=1}}|R_A(z)w| = \max_{\substack{w\in\mathbb{C}^n\\|w|=1}}\sqrt{\langle (R_A)^* R_A w, w\rangle} = \max\Big\{\frac{1}{|\lambda-z|} : \lambda\in\mathrm{spec}(A)\Big\} \leq \frac{1}{(\Im z)} \quad \text{(uniforme en } A\text{)},$$

car $|\lambda-z| = \sqrt{(\lambda-\Re z)^2+(\Im z)^2}\geq \Im z$. Si A et B sont $n\times n$ symétriques réelles de résolvante $R_A := (A-zI_n)^{-1}$ et $R_B := (B-zI_n)^{-1}$, $z\in\mathbb{C}_+$, alors l'identité immédiate $(A-zI_n)-(B-zI_n) = A-B$ donne l'identité de la résolvante :

$$R_B - R_A = -R_A(B-A)R_B.$$

7. Réciproquement, on peut retrouver la fonction génératrice avec un développement en série binomial : $(1+w)^\alpha = \sum_{n=0}^{\infty}\binom{\alpha}{n}w^n$.

En l'utilisant avec $A = G := \frac{1}{\sqrt{n}} M$, $M \sim \mathrm{GOE}$ et $B = 0$, on obtient, en notant $R := R_G = (G - zI_n)^{-1}$:

$$-I_n - z\mathbb{E}(R) = -\mathbb{E}(RG), \quad z \in \mathbb{C}_+.$$

En notant $\tau := \frac{1}{n}\mathrm{Tr}$ et $\mu_n := \frac{1}{n}\sum_{k=1}^n \delta_{\lambda_{n,k}}$ où $\lambda_{n,1} \geq \cdots \geq \lambda_{n,n}$ sont les valeurs propres de G, il vient

$$-1 - z\mathbb{E}(s_{\mu_n}(z)) = -\tau(\mathbb{E}(RG)), \quad z \in \mathbb{C}_+.$$

L'idée à présent est de tenter d'exprimer $\tau(\mathbb{E}(RG))$ en fonction de $s_{\mu_n} = \tau(R)$. Comme $\tau(\mathbb{E}(RG)) = \frac{1}{n}\sum_{p=1}^n \mathbb{E}((RG)_{pp})$ et comme $G = \frac{1}{\sqrt{n}} M$ est gaussien, il est naturel d'utiliser une intégration par parties pour faire disparaître G.

Intégration par parties

Si X est un vecteur aléatoire de $\mathbb{R}^d$ de densité $\mathrm{e}^{-V(x)}\mathrm{d}x$ avec V de classe $\mathscr{C}^1$ alors pour toute fonction $F : \mathbb{R}^d \to \mathbb{C}$ de classe $\mathscr{C}^1$ telle que $F(x)\mathrm{e}^{-V(x)} \to 0$ quand $|x| \to +\infty$, on a la formule d'intégration par parties

$$\mathbb{E}(F(X)\nabla V(X)) = -\int F \nabla \mathrm{e}^{-V}\mathrm{d}x = \int \nabla F \mathrm{e}^{-V}\mathrm{d}x = \mathbb{E}(\nabla F(X)).$$

Dans le cas où $X = G := \frac{1}{\sqrt{n}} M$, $M \sim \mathrm{GOE}$, on a $d = \frac{n(n+1)}{2}$, $V(x) = \frac{n}{4}\mathrm{Tr}(x^2)$, et on obtient, pour tous $1 \leq p, q \leq n$,

$$\partial_{x_{pq}} V(x) = \frac{n}{1 + \mathbb{1}_{p=q}} x_{pq}, \quad \text{d'où} \quad \mathbb{E}(F(G)G_{pq}) = \frac{1 + \mathbb{1}_{p=q}}{n}\mathbb{E}(\partial_{x_{pq}} F(G)).$$

On désire prendre à présent $F(A) = (R_A)_{pq}$. L'identité de la résolvante appliquée deux fois donne [8]

$$R_B - R_A = -R_A(B-A)R_B = -R_A(B-A)(R_A - R_A(B-A)R_B) = -R_A(B-A)R_A + o(\|A-B\|_{\mathrm{op}}),$$

où le $o(\|A-B\|_{\mathrm{op}})$ dépend de $z \in \mathbb{C}_+$ via $\|R_A\|_{\mathrm{op}}, \|R_B\|_{\mathrm{op}} \leq 1/(\Im z)$, d'où, pour tous $1 \leq a, b, p, q \leq n$,

$$\partial_{A_{pq}}(R_A)_{ab} = -(R_A)_{ap}(R_A)_{qb}.$$

Par conséquent, cette intégration par parties donne, pour tout $1 \leq p \leq n$, toujours avec $R := R_G := (G - zI_n)^{-1}$,

$$\mathbb{E}((RG)_{pp}) = \sum_{q=1}^n \mathbb{E}(R_{pq}G_{qp}) = \sum_{q=1}^n \mathbb{E}(R_{pq}G_{pq}) \overset{\mathrm{IPP}}{=} -\frac{1}{n}\sum_{q=1}^n \mathbb{E}(R_{pp}R_{qq}) - \frac{1}{n}\mathbb{E}((R_{pp})^2) = -\mathbb{E}(R_{pp}\tau(R)) - \frac{1}{n}\mathbb{E}((R_{pp})^2),$$

d'où, en sommant sur $1 \leq p \leq n$ et en divisant par n, et en utilisant le fait que $\mathrm{Tr}(R^2) \leq n\|R\|_{\mathrm{op}}^2 \leq n/(\Im z)^2$,

$$\mathbb{E}(\tau(RG)) = -\mathbb{E}((\tau(R))^2) + o_{n\to\infty}(1).$$

Ici et dans toute la suite, les $O_{n\to\infty}$ et $o_{n\to\infty}(1)$ dépendent de z. Ceci donne finalement, pour tout $z \in \mathbb{C}_+$,

$$-1 - z\mathbb{E}(s_{\mu_n}(z)) = \mathbb{E}(s_{\mu_n}(z)^2) + o_{n\to\infty}(1).$$

L'idée à présent consiste à dire que $\mathbb{E}(s_{\mu_n}(z)^2) = \mathbb{E}(\tau(R)^2)$ et $\mathbb{E}(s_{\mu_n}(z))^2 = \mathbb{E}(\tau(R))^2$ sont équivalents dans l'asymptotique de grande dimension $n \to \infty$, c'est-à-dire que $s_{\mu_n} = \tau(R)$ devient déterministe. Il s'agit d'un phénomène de concentration de la mesure, qu'on peut quantifier avec une inégalité de Poincaré par exemple.

Inégalité de Poincaré

Comme $G = \frac{1}{n} M$, $M \sim \mathrm{GOE}$, suit une loi gaussienne produit, de variances en $O(1/n)$, il vérifie une inégalité de Poincaré, conséquence de l'inégalité de log-Sobolev par linéarisation (cf. TD), de constante $O(1/n)$:

$$\forall f \in \mathscr{C}_b^2(\mathbb{R}^{\frac{n(n+1)}{2}}, \mathbb{C}), \quad \mathrm{Var}(f(G)) := \mathbb{E}(|f(G) - \mathbb{E}(f(G))|^2) \leq O\Big(\frac{1}{n}\Big)\mathbb{E}(|\nabla f|^2(G)).$$

Pour l'appliquer à $f(A) := \tau(R_A)$, on observe tout d'abord que

$$\partial_{A_{pq}}\tau(R_A) = \frac{1}{n}\sum_a \partial_{A_{pq}}(R_A)_{aa} = -\frac{1}{n}\sum_a (R_A)_{ap}(R_A)_{qa} = -\frac{1}{n}(R_A^2)_{qp}$$

8. On ne fait ici que calculer la dérivée de $A \mapsto (A - zI_n)^{-1}$ au moyen de l'identité de la résolvante. Il est possible de faire autrement.

d'où, en notant $\|A\|_{\mathrm{HS}}^2 = \sum_{i,j=1}^n |A_{ij}|^2 = \mathrm{Tr}(A\overline{A}^\top)$ la norme de Hilbert–Schmidt de A,

$$|\nabla f|^2 = \sum_{p,q} |\partial_{A_{pq}} \tau(R_A)|^2 = \frac{1}{n^2}\|R_A^2\|_{\mathrm{HS}}^2 \le \frac{1}{n}\|R_A^2\|_{\mathrm{op}}^2 \le \frac{1}{n}\|R_A\|_{\mathrm{op}}^4 \le \frac{1}{n(\Im z)^4} = O\Big(\frac{1}{n}\Big),$$

et on a donc, pour $R := R_G := (G - zI_n)^{-1}$ où $G := \frac{1}{\sqrt{n}}M$, $M \sim \mathrm{GOE}$,

$$\begin{aligned}
|\mathbb{E}(\tau(R)^2) - \mathbb{E}(\tau(R))^2| &= |\mathbb{E}((\tau(R) - \mathbb{E}(\tau(R)))^2)| \\
&\le \mathbb{E}(|\tau(R) - \mathbb{E}(\tau(R))|^2) \\
&= \mathrm{Var}(f(G)) \\
&= O_{n\to\infty}\Big(\frac{1}{n}\Big) O_{n\to\infty}\Big(\frac{1}{n}\Big) = o_{n\to\infty}(1),
\end{aligned}$$

ce qui donne enfin

$$\mathbb{E}(s_{\mu_n}(z))^2 + z\mathbb{E}(s_{\mu_n}(z)) + 1 = o_{n\to\infty}(1).$$

Comme $\mathbb{E}(s_{\mu_n}(z)) = s_{\mathbb{E}\mu_n}(z)$, il vient

$$s_{\mathbb{E}\mu_n}(z)^2 + z s_{\mathbb{E}\mu_n}(z)) + 1 = o_{n\to\infty}(1).$$

Donc toute valeur d'adhérence $s = s(z)$ de $(s_{\mathbb{E}\mu_n}(z))_{n\ge1}$ vérifie $s^2 + zs + 1 = 0$.

Conclusion de la preuve du théorème de Wigner pour GOE

Comme observé dans la preuve du lemme 6.4.1 (iii), on a $\Im s_{\mathbb{E}\mu_n}(z)\Im z > 0$ pour tout n, et cela sélectionne la solution de $s(z)^2 + zs(z) + 1 = 0$ telle que $\Im s(z)\Im z > 0$, qui est $\frac{1}{2}(-z + \sqrt{z^2-4}) = s_{\mu^{\mathrm{DC}}}(z)$. Ainsi $(s_{\mathbb{E}\mu_n}(z))_{n\ge1}$ a pour unique valeur d'adhérence $s_{\mu^{\mathrm{DC}}}(z)$, et comme cette suite est bornée en module par $\Im z$, elle converge vers cette valeur d'adhérence, et par le lemme 6.4.1 (ii), $\lim_{n\to\infty} \int f \mathrm{d}\mathbb{E}\mu_n = \int f \mathrm{d}\mu^{\mathrm{DC}}$ pour tout $f \in \mathscr{C}_b$. La tension de $(\mathbb{E}\mu_n)_{n\ge1}$ est garantie par exemple par la bornitude du moment d'ordre 2 : $\int x^2 \mathrm{d}\mathbb{E}\mu_n = \frac{1}{n^2}\mathbb{E}\mathrm{Tr}(G^2) = O_{n\to\infty}(1)$.

La méthode que nous avons utilisée est due à Pastur. Une approche alternative utilisée par Bail et Silverstein [5] consiste à concevoir le modèle G $n\times n$ comme un modèle $(n-1)\times(n-1)$ et un bruit ε_n, et, par inversion par bloc dans $s_{\mu_n}(z) = \frac{1}{n}\mathrm{Tr}((G - z\mathrm{I_n})^{-1})$, à obtenir une récurrence non-linéaire avec bruit $s_{\mu_n}(z) = F(s_{\mu_{n-1}}(z), \varepsilon_n)$.

6.5 Théorème de Marchenko–Pastur

Avec l'avènement des données massives à la fin du vingtième siècle et l'explosion de l'informatique et des réseaux, l'analyse asymptotique en grande dimension des matrices de covariance empiriques a pris tout son sens, quand à la fois la taille n de l'échantillon et la dimension d des données tendent vers l'infini.

Reprenons le modèle de matrice de covariance empirique du début du chapitre.

Théorème 6.5.1. de Marchenko–Pastur sur les matrices de covariance empiriques.

Soient $(Y_{ij})_{1\le i,j\ge1}$ des v.a.r. i.i.d. de moyenne 0 et de variance 1, et la matrice rectangulaire

$$Y = (Y_{ij})_{1\le i\le d_n, 1\le j\le n}.$$

Soit $\lambda_{n,1} \ge \cdots \ge \lambda_{n,d_n}$ le spectre de la matrice symétrique semi-définie positive $\frac{1}{n}YY^\top$. Supposons que

$$\lim_{n\to\infty} \frac{d_n}{n} = \rho \in (0, +\infty).$$

Alors presque sûrement,

$$\mu_n = \frac{1}{n}\sum_{k=1}^{d_n} \delta_{\lambda_{n,k}} \xrightarrow[n\to\infty]{\mathscr{C}_b} \mu_\rho^{\mathrm{MP}}$$

où μ_ρ^{MP} est la loi de Marchenko–Pastur sur $[a,b] := [(1-\sqrt{\rho})^2, (1+\sqrt{\rho})^2]$, de densité

$$\mu_\rho^{\mathrm{MP}} := q\delta_0 + \frac{\sqrt{(b-x)(x-a)}}{\rho 2\pi x}\mathbb{1}_{[a,b]}(x)dx \quad \text{où} \quad q := \max\Big(0, 1 - \frac{1}{\rho}\Big).$$

— Une illustration est donnée par les figures 6.5 et 6.5.
— Ce théorème exprime une universalité : la loi limite ne dépend pas de la loi des Y_{ij}.
— La loi de Marchenko–Pastur peut être vue comme un bruit grand dimensionnel en quelque sorte, autour de δ_1 qui est la distribution spectrale de la matrice de covariance de population $I_n = \mathbb{E}(\frac{1}{n} Y Y^\top)$. Le régime de la statistique classique correspond à d fixe et donc à $\rho = 0$, ce qui fait disparaître ce bruit. Le théorème de Marchenko–Pastur possède une généralisation pour la situation ou la matrice de covariance de population est quelconque, ce qui donne une version colorée du bruit grand dimensionnel.
— Un modèle signal-plus-bruit peut prendre la forme $\frac{1}{n}(X+Y)(X+Y)^\top$. Son analyse spectrale asymptotique en grande dimension révèle que lorsque X est une matrice déterministe de rang fini, un phénomène de seuil apparaît (détection de signal) : transition BBP, cf. partie physique du cours!
— Le théorème de Marchenko–Pastur peut être démontré avec les mêmes méthodes que le théorème de Wigner, bien que la mise en œuvre soit plus lourde en raison du caractère quadratique du modèle.
— La matrice de covariance empirique $\frac{1}{n} Y Y^\top$ suit la loi de Wishart. Pour $d = 1$ on retrouve une loi du χ^2.
— La loi μ_ρ^{MP} a pour moyenne 1, variance ρ, et ses moments sont données pour tout $r \geq 1$ par

$$\int x^r \, d\mu_\rho^{\mathrm{MP}}(x) = \sum_{k=0}^{r-1} \frac{\rho^k}{k+1} \binom{r}{k} \binom{r-1}{k}.$$

— La loi μ_ρ^{MP} est un mélange (une combinaison convexe) entre une masse de Dirac en 0 et une loi à densité. L'atome en 0 est dû au fait que la matrice n'est pas forcément de rang plein, et disparaît lorsque $\rho \leq 1$.
— Si $\rho = 1$ alors Y est en quelque sorte asymptotiquement carrée. Dans ce cas, $(a,b,q) = (0,4,0)$. Par changement de variable, presque sûrement la mesure de comptage du spectre de $(\frac{1}{n} Y Y^\top)^{1/2}$ converge étroitement quand $n \to \infty$ vers la loi du quart-de-cercle de densité $x \mapsto \frac{1}{\pi}\sqrt{4-x^2}\, \mathbb{1}_{[0,2]}(x)$.
— Si R est une matrice rectangulaire $d \times n$, les valeurs propres de la matrice $\sqrt{RR^\top}$ sont les valeurs singulières de R. La décomposition en valeurs singulières (SVD en anglais) est $R = ODP$ où O et P sont des matrices orthogonales $d \times d$ et $n \times n$, et où D est une matrice rectangulaire $d \times n$ diagonale dont les éléments diagonaux sont les valeurs singulières de R. La SVD conduit à une méthode de réduction de dimension, appelée analyse en composantes principales (ACP en français et PCA en anglais) qui consiste à éliminer les valeurs singulières en dessous d'un seuil. Il est possible de marier cette approche classique en analyse des données avec l'analyse en grande dimension des matrices de covariance empirique.
— L'image de μ_ρ^{MP} par l'application affine $x \mapsto \frac{x-(1+\rho)}{\sqrt{\rho}}$ est $\frac{\sqrt{4-x^2}}{2\pi(1+\rho+\sqrt{\rho}x)} \mathbf{1}_{[-2,2]}(x) \to \mu^{\mathrm{DC}}$ quand $\rho \to 0$.

6.6 Pour aller plus loin

— Dans l'esprit du chapitre 5, que se passe-t-il si les M_{ij} n'ont pas de variance? Bouchaud et Cizeau ont découvert que si la loi est à queue lourde, en loi de puissance $t^{-\alpha}$ avec $0 < \alpha < 2$, alors un résultat analogue au théorème de Wigner subsiste, à condition de normaliser par $n^{1/\alpha}$ au lieu de $\sqrt{n}$, et la loi limite est nouvelle, universelle, portée par tout $\mathbb{R}$, à queue lourde en loi de puissance $t^{-\alpha}$.
— Il est possible de raffiner le théorème de Wigner de différentes manières. En voici quelques unes :
  — Bord : pour GOE, on peut établir que presque sûrement, le support de la distribution spectrale empirique de $\frac{1}{\sqrt{n}} M$ converge vers le support $[-2,2]$ de la loi limite μ^{DC}, c'est-à-dire que

$$\lim_{n\to\infty} \lambda_{n,n} = -2 \quad \text{et} \quad \lim_{n\to\infty} \lambda_{n,1} = 2.$$

  De plus ce phénomène est universel : il reste vrai pour les matrices de Wigner, ssi les coefficients possèdent un moment d'ordre 4 fini. Enfin la fluctuation des valeurs propres extrêmes autour du bord est décrite par une loi de Tracy–Widom, qui n'est pas une lois des extrêmes de v.a.r. i.i.d.
  — Fluctuation : pour une fonction f suffisamment régulière, la fluctuation de $\mu_n(f) - \mathbb{E}\mu_n(f)$ quand $n \to \infty$ est gaussienne mais sa variance est plus petite que si les valeurs propres étaient indépendantes, et fait intervenir une norme de Sobolev de f.
  — Loi locale : les effets joints du confinement et de la répulsion singulière dans GOE font que les valeurs propres forment un nuage rigide dont l'espacement est proche de $\frac{1}{n}$ à l'intérieur de $[-2,2]$. L'analyse de la fluctuation du nuage à l'échelle microscopique fait apparaître une convergence vers un processus spécial, et ce phénomène est universel, au-delà du GOE : on parle de loi locale.
  — Grandes déviations : le PGD à la Sanov de GOE est permis par la connaissance de la loi jointe des valeurs propres dans ce cas. À ce jour, les PGD pour les matrices de Wigner restent mal compris.

Les matrices aléatoires constituent un vaste champ de recherche, encore bien vivant aujourd'hui. Les travaux les plus anciens sont dus principalement à John Wishart (années 1920) sur les matrices de covariance

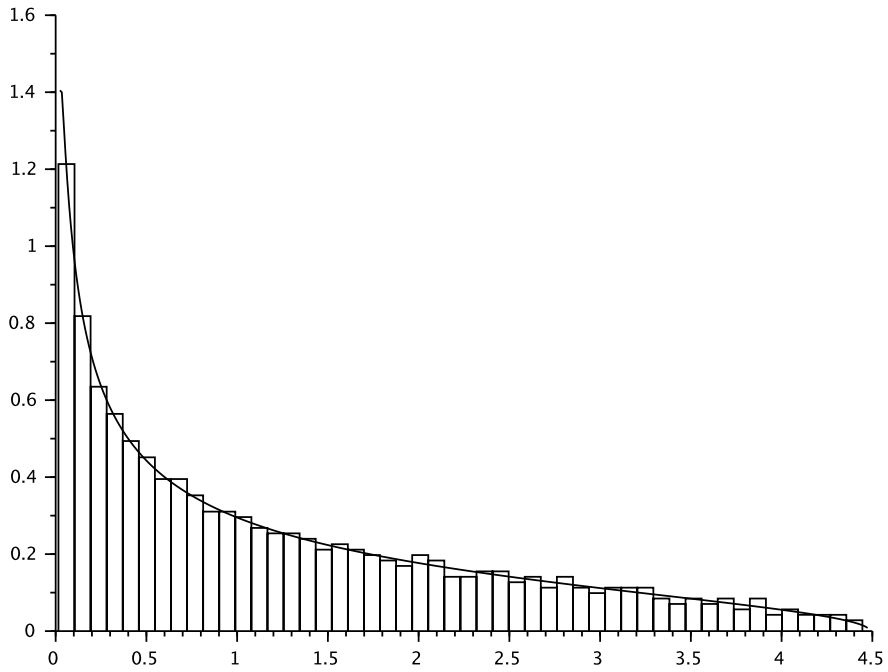


FIGURE 6.5 – Histogramme des valeurs propres d'une matrice de covariance empirique avec $d = 800$ et $n = 1000$ et loi de Marchenko–Pastur. Dans ce cas $d/n = 4/5 = 0.8$, à comparer avec μ_ρ^{MP} avec $\rho = 0.8$ dans la figure 6.5.

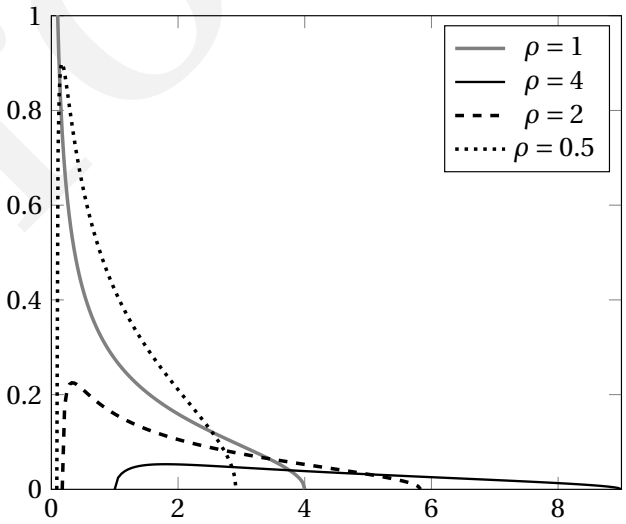


FIGURE 6.6 – Partie à densité de μ_ρ^{MP} pour différentes valeurs de ρ.

empiriques des échantillons gaussiens, à John von Neumann et Herman Goldstine (années 1940) sur l'analyse numérique matricielle, et à Eugene Wigner, Freeman Dyson, et Madan Lal Mehta (années 1950-1960) sur les niveaux d'énergie des noyaux atomiques en physique théorie et physique mathématique. Le théorème de Marchenko–Pastur a été obtenu par Vladimir Marchenko et Leonid Pastur dans les années 1960. La loi du demi-cercle constitue un analogue de la loi gaussienne dans la théorie des probabilités libres initiée par Dan-Virgil Voiculescu, en liaison avec les matrices aléatoires. Le théorème 6.1 de Wigner a été obtenu par Wigner dans les années 1950 sous des hypothèses plus restrictives. La version la plus aboutie date de la fin des années 1970.

La méthode des moments a été utilisée par Wigner lui-même, tandis que la méthode de la résolvante via la transformée de Cauchy–Stieltjes a été utilisée par Marchenko et Pastur puis raffinée par Pastur notamment. Ces deux méthodes ont été également développées par Bai et Silverstein notamment. Il y a enfin plusieurs manières de mettre en œuvre la méthode des moments et la méthode de la transformée de Cauchy–Stieltjes.

Quelques ouvrages accessibles ou de référence : [25], [65], [59], [1], [5], [41], [58].

Annexe A

Quelques rappels d'intégration et probabilités

Réflexes probabilistes de base :
— $\mathbb{E} = \int = \Sigma$
— $\mathbb{P}(A) = \mathbb{E}(\mathbb{1}_A)$
— $\sum_n \mathbb{1}_{A_n}$ pour compter (diviser pour régner)
— $\mathbb{E}(f) = \mathbb{E}(f\mathbb{1}_A) + \mathbb{E}(f\mathbb{1}_{A^c})$
— Si X et Y sont indépendantes alors par Fubini–Tonelli $\mathbb{E}(f(X,Y)) = \mathbb{E}(F(Y))$ où $F(y) = \mathbb{E}(f(X,y))$.

A.1 Inégalités : Hölder, Cauchy–Schwarz, Jensen, Markov

— Inégalité de Hölder (Cauchy–Schwarz si $p = 2$). Si $1 \le p \le \infty$, $\frac{1}{p} + \frac{1}{q} = 1$, alors

$$\mathbb{E}(|XY|) \le \mathbb{E}(|X|^p)^{1/p}\mathbb{E}(|Y|^q)^{1/q}.$$

De plus, si $1 < p < \infty$, alors il y a égalité ssi $|X|^p$ et $|Y|^q$ sont colinéaires.
— Inégalité de Jensen. Si $X \in L^1(\mathbb{R}^n)$ et $\varphi : \mathbb{R}^n \to \mathbb{R}$ convexe, alors

$$\mathbb{E}(\varphi(X)) \le \varphi(\mathbb{E}(X)).$$

De plus, si φ est strictement convexe, alors il y a égalité ssi X est constante.
— Inégalité de Markov. Si X v.a.r. ≥ 0 alors pour tout $r > 0$ et $p > 0$,

$$\mathbb{P}(X \ge r) \le \frac{\mathbb{E}(X^p)}{r^p}.$$

Plus généralement, si $\varphi : [0,\infty) \to [0,\infty)$ est croissante alors [1],

$$\mathbb{P}(X \ge r) = \mathbb{P}(\varphi(X) \ge \varphi(r)) = \mathbb{E}(\mathbb{1}_{\{\varphi(X) \ge \varphi(r)\}}) \le \mathbb{E}\Big(\frac{\varphi(X)}{\varphi(r)}\Big).$$

A.2 Caractérisation de la loi

— Fonctions tests indicatrices de $\{x : x_i \le t_i, 1 \le i \le n\}$, fonction de répartition.
— Fonctions tests continues et bornées, point de vue mesure de Radon.
— Fonctions tests $\mathscr{C}_c^\infty$, point de vue distributionnel.
— Fonctions tests trigonométriques $\{e^{it} : t \in \mathbb{R}\}$, fonction caractéristique $\varphi_X(t) := \mathbb{E}(e^{i\langle X,t\rangle}) = \varphi_{\langle X,t\rangle}(1)$, transformée de Fourier, théorème de Cramér–Wold, sondage par projection univariée, tomographie.
— Moments, sous condition de quasi-analycité de la transformée de Fourier, critère de Carleman, transformée de Cauchy–Stieltjes, potentiel logarithmique.

A.3 Convergences : presque sûre, en probabilité, en moyenne, en loi

— Convergence presque sûre. $X_n \xrightarrow{\text{p.s.}} X$ lorsque $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.
— Convergence en probabilité. $X_n \xrightarrow{\mathbb{P}} X$ lorsque $\forall \varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \ge \varepsilon) = 0$.
— Convergence en moyenne. Pour $1 \le p < \infty$, $X_n \xrightarrow{L^p} X$ lorsque $X \in L^p$ et $\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0$.

1. Par $a\mathbb{1}_{x \ge a} \le x$ pour $x \ge 0$ et monotonie de $\mathbb{E}$.

— Convergence en loi. Soit $X_n \sim \mu_n$ et $X \sim \mu$. Les propriétés suivantes sont équivalentes et on dit que $X_n \xrightarrow{\text{loi}} X$, ou $\mu_n \xrightarrow{\mathscr{C}_b} \mu$ (convergence étroite) lorsque $\lim_{n\to\infty} \mathbb{E}(f(X_n)) = \mathbb{E}(f(X))$ pour tout $f \in \mathscr{F}$ où $\mathscr{F}$ est l'une des classes de fonctions tests suivantes (liées au théorème porte-manteau [2]) :
  — $\mathscr{F} = \mathscr{C}_b(\mathbb{R},\mathbb{R})$
  — $\mathscr{F} = \mathscr{C}_c^\infty(\mathbb{R},\mathbb{R})$
  — $\mathscr{F} = \{\mathbb{1}_{(-\infty,x]} : x \text{ point de continuité de } F := \mathbb{P}(X \le \bullet) = \mu((-\infty,\bullet))\}$ (fonction de répartition)
  — $\mathscr{F} = \{x \mapsto \mathrm{e}^{\mathrm{i}tx} : t \in \mathbb{R}\}$ (transformée de Fourier ou fonction caractéristique)
  — $\mathscr{F} = \{x \mapsto \mathrm{e}^{-tx} : t \in [0,\infty)\}$ (transformée de Laplace transform, quand les X_n et X sont ≥ 0)
  — $\mathscr{F} = \{x \mapsto s^x : s \in [0,1]\}$ (fonction génératrice, quand les X_n et X sont discrètes à valeurs dans $\mathbb{N}$)

  En prenant une suite qui ne dépend pas de n on obtient des caractérisations de la loi.

La convergence en loi ne fait intervenir que la loi de X et la loi de X_n pour tout n, tandis que les trois autres modes de convergence nécessitent la loi du couple (X_n, X) et donc de définir la suite (X_n) et la limite X sur un même espace de probabilité sauf dans le cas spécial où X est constante.

Ces notions de convergence s'étendent naturellement aux vecteurs aléatoires en utilisant une distance, norme, ou produit scalaire, par exemple pour les fonctions caractéristiques, on remplace $\mathrm{i}tX$ par $\mathrm{i}\langle t, X\rangle$.

Ces notions de convergence s'étendent également naturellement aux espaces métriques.

$$\begin{array}{ccccc} & & \boxed{\text{CV L}^p} & & \\ & & \Downarrow & & \\ & & \boxed{\text{CV L}^1} & & \\ & & \Downarrow & & \\ \boxed{\text{CV p.s.}} & \Rightarrow & \boxed{\text{CV en } \mathbb{P}} & \Rightarrow & \boxed{\text{CV en loi}} \end{array}$$

Si X est constante alors la convergence en loi implique la convergence en probabilité.

La convergence dans L^1 est équivalente à la convergence en probabilité plus l'intégrabilité uniforme [3].

A.4 Convergence monotone, lemme de Fatou, convergence dominée

— Théorème de convergence monotone. Si $(X_n)_{n\ge1}$ est croissante et prend ses valeurs dans $[0,\infty]$ alors

$$\mathbb{E}(\lim_{n\to\infty} X_n) = \lim_{n\to\infty} \mathbb{E}(X_n) \in [0,\infty].$$

— Lemme de Fatou. Si $(X_n)_{n\ge1}$ prend ses valeurs dans $[0,\infty]$ alors

$$\mathbb{E}(\underline{\lim}_{n\to\infty} X_n) \le \underline{\lim}_{n\to\infty} \mathbb{E}(X_n) \in [0,\infty].$$

— Théorème de convergence dominée. Si $X_n \xrightarrow{\mathbb{P}} X$ et si [4] $\sup_n |X_n| \le Y$ avec $\mathbb{E}(Y) < \infty$, alors

$$X_n \xrightarrow[n\to\infty]{\mathrm{L}^1} X, \quad \text{en particulier} \quad \lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(\lim_{n\to\infty} X_n) = \mathbb{E}(X).$$

— Lemme de Slutsky. Si $X_n \xrightarrow{\text{loi}} X$ et $Y_n \xrightarrow{\text{loi}} Y$ et Y est constante alors $(X_n, Y_n) \xrightarrow{\text{loi}} (X, Y)$.
  En particulier $X_n Y_n \xrightarrow{\text{loi}} XY$, $X_n + Y_n \xrightarrow{\text{loi}} X + Y$, $X_n / Y_n \xrightarrow{\text{loi}} X/Y$ si $Y \ne 0$.

— Théorème de Fubini–Tonelli. Si $(\Omega_1, \mathscr{A}_1, \mu_1)$ et $(\Omega_2, \mathscr{A}_2, \mu_2)$ sont des espaces mesurés, et si $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ est mesurable, avec $f \ge 0$ ou $f \in \mathrm{L}^1(\mu_1 \otimes \mu_2)$, alors

$$\int f(x,y)\mathrm{d}(\mu_1\otimes\mu_2)(x,y) = \int \left(\int f(x,y)\mathrm{d}\mu_1(x)\right)\mathrm{d}\mu_2(y).$$

— Intégrabilité et queue de distribution (complète l'inégalité de Markov). Si $X \ge 0$ alors pour tout $p > 0$,

$$\mathbb{E}(X^p) = p\int_0^\infty t^{p-1}\mathbb{P}(X > t)\mathrm{d}t.$$

2. D'après Billingsley, le théorème porte-manteau est dû à Alexandrov. Dans la deuxième édition du classique *Convergence of Probability Measures*, Billingsley attribue le théorème à Jean-Pierre Portmanteau, de l'université de Felletin, dans un article de 4 pages que Jean-Pierre Portmanteau aurait publié en 1915 dans les Annales de l'Université de Felletin, sous le titre « Espoir pour l'ensemble vide ? ». Il s'agit d'un canular : il n'y a pas de mathématicien portant le nom de Jean-Pierre Portmanteau, et il n'y a jamais eu d'université à Felletin.

3. Voir `https://djalil.chafai.net/blog/2014/03/09/de-la-vallee-poussin-on-uniform-integrability/`

4. Cette condition de domination n'est qu'en fait une condition suffisante pour assurer l'intégrabilité uniforme.

Plus généralement, si $\varphi : [0,\infty) \to [0,\infty)$ est $\mathscr{C}^1$ et $\varphi(0) = 0$, alors [5]

$$\mathbb{E}(\varphi(X)) = \int \varphi(x)\mathrm{d}\mathbb{P}_X(x) = \int \Big(\int_0^x \varphi'(t)\mathrm{d}t\Big)\mathrm{d}\mathbb{P}(x) = \int_0^\infty \varphi'(t)\Big(\int \mathbb{1}_{0\le t\le x}\mathrm{d}\mathbb{P}(x)\Big)\mathrm{d}t = \int_0^\infty \varphi'(t)\mathbb{P}(X > t)\mathrm{d}t.$$

Dans la même veine : $\mathbb{E}(|X|) - 1 \le \sum_{n=1}^\infty \mathbb{P}(|X| \ge n) \le \mathbb{E}(|X|)$.

— Intégrabilité et troncature. $\mathbb{E}(|X|) < \infty$ ssi $\lim_{r\to\infty} \mathbb{E}(|X|\mathbb{1}_{|X|\ge r}) = 0$.

— Théorème de Paul Lévy. Si $(X_n)_{n\ge1}$ vecteurs aléatoires de $\mathbb{R}^d$ avec $\lim_{n\to\infty} \varphi_{X_n}(t) := \mathbb{E}(\mathrm{e}^{\mathrm{i}\langle t, X_n\rangle}) = \varphi(t)$ pour tout $t \in \mathbb{R}^d$, avec φ continue en 0, alors il existe X tel que $\varphi = \varphi_X$ et $X_n \xrightarrow{\text{loi}} X$ quand $n \to \infty$.

— Caractérisation de l'indépendance. X et Y indépendants ssi $\varphi_{(X,Y)}(s,t) = \varphi_X(s)\varphi_Y(t)$ pour tous $s, t \in \mathbb{R}^d$.

— Fonction caractéristique et moments. Si X a tous ses moments finis jusqu'à l'ordre m alors φ_X est dérivable en 0 jusqu'à l'ordre m et $\varphi^{(k)}(0) = \mathrm{i}^{k_1+\cdots+k_d}\mathbb{E}(X_1^{k_1}\cdots X_d^{k_d})$ pour tout $k \in \mathbb{N}^d$ avec $0 \le k_1+\cdots+k_d \le m$.

— Continuous mapping theorem. Si $X_n \xrightarrow{\text{loi}} X$ et f continue alors $f(X_n) \xrightarrow{\text{loi}} f(X)$.

— Méthode delta. Si $a_n(X_n - c) \xrightarrow[n\to\infty]{\text{loi}} L$, $a_n \to \infty$, et si $f : \mathbb{R} \to \mathbb{R}$ est continue, dérivable en c, alors

$$a_n(f(X_n) - f(c)) \xrightarrow[n\to\infty]{\text{loi}} \mathrm{Loi}(f'(c)Z) \quad \text{où} \quad Z \sim L.$$

A.5 Quelques autres lois classiques

— Gaussiennes isotropes. On note $\gamma_\sigma^n := \mathscr{N}(0, \sigma^2 I_n)$, $\sigma \ge 0$, de densité : $\frac{\mathrm{e}^{-\frac{|x|^2}{2\sigma^2}}}{(\sqrt{2\pi}\sigma)^n}$.

— Si X vecteur aléatoire de $\mathbb{R}^n$ de densité f alors pour tout $\sigma > 0$, σX a pour densité $\frac{1}{\sigma^n} f(\frac{\cdot}{\sigma})$

— Si X variable aléatoire positive de densité f alors pour tout $\alpha > 0$, X^α a pour densité $v \mapsto \frac{v^{\frac{1-\alpha}{\alpha}}}{\alpha} f(v^{\frac{1}{\alpha}})$.

— Loi uniforme sur $[a,b]$. Densité $u \in [a,b] \mapsto \frac{1}{b-a}$, moyenne $\frac{a+b}{2}$, variance $\frac{(b-a)^2}{12}$.

— Loi de Cauchy Cauchy(x_0, a), $x_0 \in \mathbb{R}$, $a > 0$, de densité $x \mapsto \frac{a}{\pi(a^2+(x-x_0)^2)}$, standard quand $(x_0, a) = (0,1)$.

— Loi Gamma(a,λ), densité $x \in [0,\infty) \mapsto \frac{\Gamma(a)}{\lambda} x^{a-1}\mathrm{e}^{-\lambda x}$, a et λ sont les paramètres de forme et d'échelle. Moyenne $= a/\lambda$, variance $= a/\lambda^2$. On a Gamma$(a_1,\lambda) * \cdots * $ Gamma$(a_n,\lambda) =$ Gamma$(a_1+\cdots+a_n,\lambda)$.

— Loi Beta(α,β), densité $r \in [0,1] \mapsto \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} r^{\alpha-1}(1-r)^{\beta-1}$, moyenne $\frac{\alpha}{\alpha+\beta}$, variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
Loi Beta(α,β) sur $[-1,1]$ de densité $t \in [-1,1] \mapsto \frac{\Gamma(\alpha+\beta)}{\Gamma(a)\Gamma(b)} 2^{1-a-b}(1+t)^{a-1}(1-t)^{b-1}$.

— Quelques Beta spéciales. Beta$(1,1) =$ Unif$([0,1])$, Beta$(\alpha,\infty) = \delta_0$, Beta$(\infty,\infty) = \delta_{1/2}$, Beta$(\infty,\beta) = \delta_1$.
Si $X_{\alpha,\beta} \sim$ Beta(α,β) alors $\sqrt{X_{\alpha,\beta}}$ a pour densité $r \in [0,1] \mapsto \frac{\Gamma(\alpha)(\Gamma(\beta)}{2\Gamma(\alpha+\beta)} r^\alpha(1-r^2)^{\beta-1}$.
En particulier $\sqrt{X_{0,1/2}} \sim$ ArcSinus$([0,1])$ et $\sqrt{X_{0,3/2}} \sim$ SemiCercle$([0,1])$.

— Loi exponentielle Exp$(\lambda) =$ Gamma$(1,\lambda)$, densité $x \in [0,\infty) \mapsto \frac{1}{\lambda}\mathrm{e}^{-\lambda x}$.

— Loi du khi-deux à n degrés de liberté est définie par $\chi^2(n) :=$ Loi$(|Z|^2)$ avec $Z \sim \mathscr{N}(0, I_n)$, et vérifie $\chi^2(n) =$ Gamma$(\frac{n}{2}, \frac{1}{2})$, en particulier $\chi^2(1) =$ Gamma$(\frac{1}{2}, \frac{1}{2})$ et $\chi^2(2) =$ Exp$(\frac{1}{2})$.

— Loi du khi à n degrés de liberté : $\chi(n) :=$ Loi$(|Z|) =$ Loi$(\sqrt{R})$ avec $Z \sim \mathscr{N}(0, I_n)$ et $R \sim \chi^2(n)$.

— Loi de Student [6] : si $X \sim \mathscr{N}(0,1)$ et $Y \sim \chi(n)$ alors $X/Y \sim t(n)$.

— Loi de Dirichlet : Dirichlet$(a_1,\dots,a_n) =$ Loi$\Big(\frac{(X_1,\dots,X_n)}{X_1+\cdots+X_n}\Big)$ avec $X_1,\dots,X_n$ indépendantes, $X_i \sim$ Gamma$(a_i,1)$.
Si $D \sim$ Dirichlet$(a_1,\dots,a_n)$ alors $D_i \sim$ Beta$(a_i, a_1+\cdots+a_n-a_i)$.
Si $I_1,\dots,I_k$ est une partition de $\{1,\dots,n\}$ alors $(\sum_{i\in I_1} D_i,\dots,\sum_{i\in I_k} D_i) \sim$ Dirichlet$(\sum_{i\in I_1} a_i,\dots,\sum_{i\in I_k} a_i)$.
En particulier pour tout $I \subset \{1,\dots,n\}$ distinct de $\varnothing$ et $\{1,\dots,n\}$, on a $\sum_{i\in I} D_i \sim$ Beta$(\sum_{i\in I} a_i, \sum_{i\notin I} a_i)$.
On peut dilater par $1/\lambda$, ce qui revient à considérer des Gamma(a_i,λ), d'où par exemple la propriété : si $A \sim \chi^2(a) =$ Gamma$(a, \frac{1}{2})$ et $B \sim \chi^2(b) =$ Gamma$(b, \frac{1}{2})$ sont indépendantes alors $A/(A+B) \sim$ Beta$(\frac{a}{2}, \frac{b}{2})$.

— Partition aléatoire et statistique d'ordre uniforme. Soit $d > 1$ et soient $U_1,\dots,U_{d-1}$ des v.a.r. sur $[0,1]$. Soit $U_{(0)} \le \cdots \le U_{(d)}$ leur réarrangement croissant, avec la convention $U_{(0)} := 0$ et $U_{(d)} := 1$. Alors on a $(U_{(1)} - U_{(0)},\dots,U_{(d)} - U_{(d-1)}) \sim$ Dirichlet$(1,\dots,1)$ ssi les $U_1,\dots,U_{d-1}$ sont i.i.d. de loi uniforme sur $[0,1]$.

A.6 Formules et fonctions spéciales

— Coordonnées sphériques (polaires si $n = 2$). $\mathrm{d}x = r^{n-1}\mathrm{d}r\mathrm{d}\theta$ avec r dans $[0,\infty)$ et θ dans $\mathbb{S}^{n-1}$.

— Fonction Gamma. $\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}\mathrm{d}t$, $\Gamma(z+1) = z\Gamma(z)$, $\Gamma(n+1) = n!$.

5. Par le théorème fondamental du calcul et le théorème de Fubini–Tonelli.

6. Le nom Student, utilisé par Ronald Fisher, est le pseudonyme utilisé par William Gosset dans son article publié en 1908 dans Biometrika, son employeur, les brasseries Guinness (Dublin), ne l'ayant pas autorisé à utiliser son nom pour ses publications scientifiques.

— Volume et surface. $|\mathbb{B}^n| = \dfrac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}$ et $|\mathbb{S}^{n-1}| = \partial_{r=1}|\mathbb{B}^n(r)| = |\mathbb{B}^n|\partial_{r=1} r^n = n|\mathbb{B}^n| = \dfrac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}$.

— Fonction de répartition gaussienne. $\Phi(x) = \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t = \operatorname{erfc}\Big(\dfrac{x}{\sqrt{2}}\Big)$.

— Fonction d'erreur. $\operatorname{erf}(x) = \dfrac{2}{\sqrt{\pi}}\displaystyle\int_0^x \mathrm{e}^{-t^2}\,\mathrm{d}t = 2\Phi(\sqrt{2}x) - 1$.

— Fonction d'erreur complémentaire. $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = 2\Phi(-\sqrt{2}x) = 2(1 - \Phi(\sqrt{2}x))$

— Estimées gaussiennes. $\operatorname{erfc}(x) = \dfrac{\mathrm{e}^{-x^2}}{\sqrt{\pi}x}\Big(1 - \dfrac{1}{2(1+x^2)} + \dfrac{1}{4(1+x^2)(2+x^2)} - \cdots\Big)$, et $\operatorname{erfc}(x) \le \mathrm{e}^{-x^2}$ si $x \ge 0$.

— Formules de Stirling. $\Gamma(x+1) \sim_{x\to\infty} \sqrt{2\pi x}\Big(\dfrac{x}{\mathrm{e}}\Big)^x$, $\Gamma(x+a) \sim_{x\to\infty} x^a\Gamma(x)$.

Annexe B

Lexique bilingue français/anglais

Français	Anglais
Loi des grands nombres (LGN)	Law of large numbers (LLN)
Théorème limite central (TLC)	Central limit theorem (CLT)
Corps convexe	Convex body
-	Continuous mapping theorem
En position isotrope	In isotropic position
Phénomène ou effet couche mince	Thin-shell phenomenon or effect
Inégalité de Sobolev logarithmique (ISL)	Logarithmic Sobolev inequality (LSI)
Inégalité de concentration	Concentration inequality
Inégalité de transport	Transportation inequality
Principe de grandes déviations	Large deviation principle
Principe de contraction	Contraction principle
Fonction à variation régulière	Regularly varying function
Fonction à variation lente	Slowly varying function
Distribution à queue lourde	Heavy tailed distribution
Loi stable	Stable law
Phénomène de grande dimension	High dimensional phenomenon
Méthode des moments	Moments method
Combinatoire	Combinatorics
Loi du demi-cercle	Semi-circle law

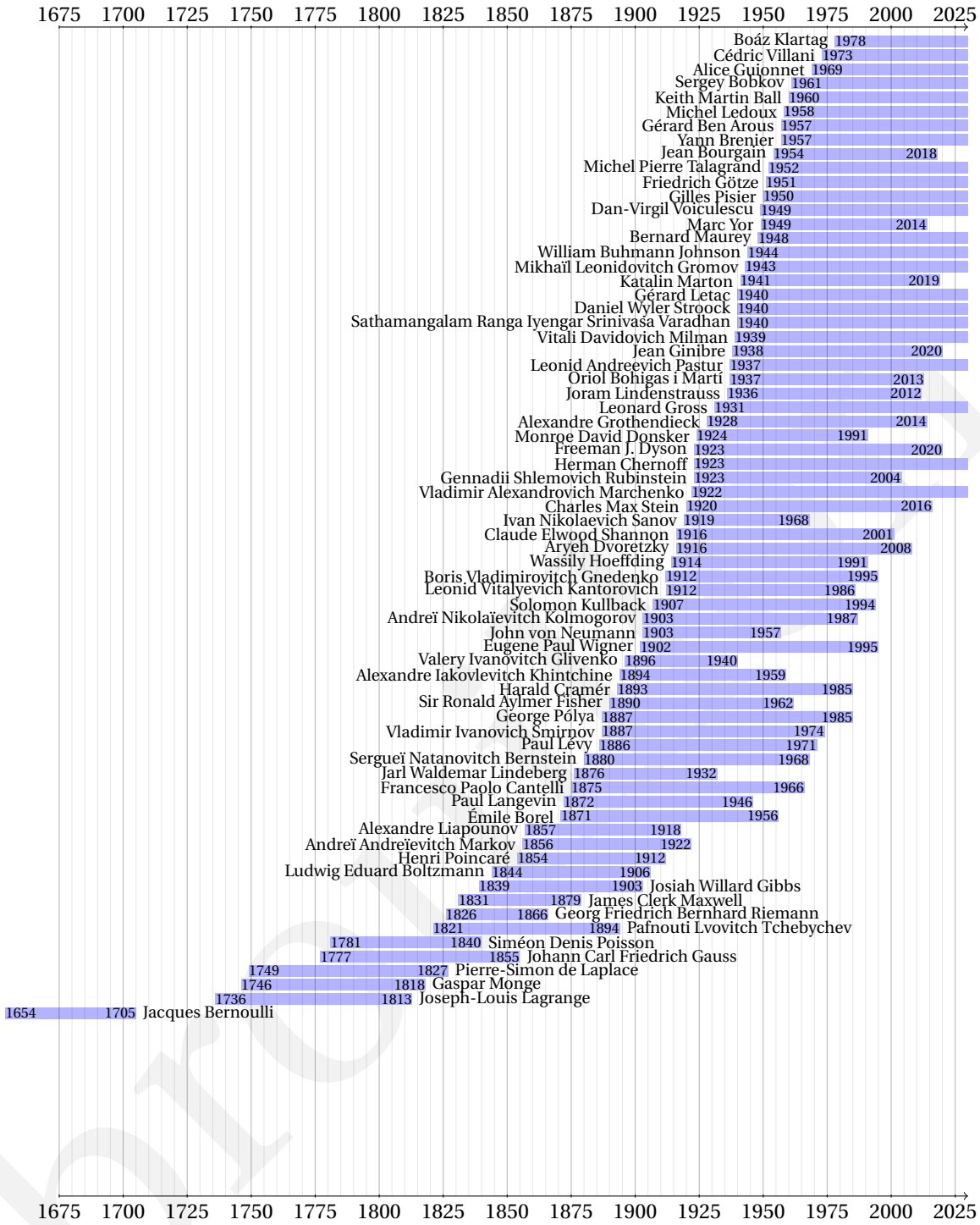
Annexe C

Chronologie

Quelques personnages historiques liés à ces notes [1].

Hors cadre : Archimède de Syracuse (-287 – -212).

1. Merci à Julien Guillod pour le partage de son code Tikz.


1675 1700 1725 1750 1775 1800 1825 1850 1875 1900 1925 1950 1975 2000 2025
Boáz Klartag 1978
Cédric Villani 1973
Alice Guionnet 1969
Sergey Bobkov 1961
Keith Martin Ball 1960
Michel Ledoux 1958
Gérard Ben Arous 1957
Yann Brenier 1957
Jean Bourgain 1954 2018
Michel Pierre Talagrand 1952
Friedrich Götze 1951
Gilles Pisier 1950
Dan-Virgil Voiculescu 1949
Marc Yor 1949 2014
Bernard Maurey 1948
William Buhmann Johnson 1944
Mikhaïl Leonidovitch Gromov 1943
Katalin Marton 1941 2019
Gérard Letac 1940
Daniel Wyler Stroock 1940
Sathamangalam Ranga Iyengar Srinivasa Varadhan 1940
Vitali Davidovich Milman 1939
Jean Ginibre 1938 2020
Leonid Andreevich Pastur 1937
Oriol Bohigas i Martí 1937 2013
Joram Lindenstrauss 1936 2012
Leonard Gross 1931
Alexandre Grothendieck 1928 2014
Monroe David Donsker 1924 1991
Freeman J. Dyson 1923 2020
Herman Chernoff 1923
Gennadii Shlemovich Rubinstein 1923 2004
Vladimir Alexandrovich Marchenko 1922
Charles Max Stein 1920 2016
Ivan Nikolaevich Sanov 1919 1968
Claude Elwood Shannon 1916 2001
Aryeh Dvoretzky 1916 2008
Wassily Hoeffding 1914 1991
Boris Vladimirovitch Gnedenko 1912 1995
Leonid Vitalyevich Kantorovich 1912 1986
Solomon Kullback 1907 1994
Andreï Nikolaïevitch Kolmogorov 1903 1987
John von Neumann 1903 1957
Eugene Paul Wigner 1902 1995
Valery Ivanovitch Glivenko 1896 1940
Alexandre Iakovlevitch Khintchine 1894 1959
Harald Cramér 1893 1985
Sir Ronald Aylmer Fisher 1890 1962
George Pólya 1887 1985
Vladimir Ivanovich Smirnov 1887 1974
Paul Lévy 1886 1971
Sergueï Natanovitch Bernstein 1880 1968
Jarl Waldemar Lindeberg 1876 1932
Francesco Paolo Cantelli 1875 1966
Paul Langevin 1872 1946
Émile Borel 1871 1956
Alexandre Liapounov 1857 1918
Andreï Andreïevitch Markov 1856 1922
Henri Poincaré 1854 1912
Ludwig Eduard Boltzmann 1844 1906
1839 1903 Josiah Willard Gibbs
1831 1879 James Clerk Maxwell
1826 1866 Georg Friedrich Bernhard Riemann
1821 1894 Pafnouti Lvovitch Tchebychev
1781 1840 Siméon Denis Poisson
1777 1855 Johann Carl Friedrich Gauss
1749 1827 Pierre-Simon de Laplace
1746 1818 Gaspar Monge
1736 1813 Joseph-Louis Lagrange
1654 1705 Jacques Bernoulli
1675 1700 1725 1750 1775 1800 1825 1850 1875 1900 1925 1950 1975 2000 2025

Bibliographie

[1] G. W. ANDERSON, A. GUIONNET et O. ZEITOUNI : *An introduction to random matrices*, vol. 118 de *Camb. Stud. Adv. Math.* Cambridge : Cambridge University Press, 2010.

[2] C. ANÉ, S. BLACHÈRE, D. CHAFAÏ, P. FOUGÈRES, I. GENTIL, F. MALRIEU, C. ROBERTO et G. SCHEFFER : *Sur les inégalités de Sobolev logarithmiques*, vol. 10 de *Panor. Synth.* Paris : Société Mathématique de France, 2000.

[3] G. AUBRUN : Vers la conjecture de Kannan–Lovász–Simonovits, d'après Yuansi Chen. Séminaire Bourbaki. Exposé 1192. À parraître dans Astérisque, 2022.

[4] G. AUBRUN, J. JENKINSON et S. J. SZAREK : Optimal constants in concentration inequalities on the sphere and in the gauss space. arXiv:2406.13581v2, 2024.

[5] Z. BAI et J. W. SILVERSTEIN : *Spectral analysis of large dimensional random matrices.* Springer Ser. Stat. Dordrecht : Springer, 2nd ed. édn, 2010.

[6] R. BALDASSO, R. I. OLIVEIRA, A. PEREIRA et G. REIS : A Proof of Sanov's Theorem via Discretizations. To appear in Journal of Theoretical Probability arXiv:2112.04280, 2022.

[7] P. BARBE et M. LEDOUX : *Probabilité.* Paris : EDP Sciences, nouvelle ed. édn, 2007.

[8] F. BARTHE : Un théorème de la limite centrale pour les ensembles convexes (d'après Klartag et Fleury-Guédon-Paouris). Num. 332, p. Exp. No. 1007, ix, 287–304. 2010. Séminaire Bourbaki. Volume 2008/2009. Exposés 997–1011.

[9] F. BARTHE, O. GUÉDON, S. MENDELSON et A. NAOR : A probabilistic approach to the geometry of the ℓ_p^n-ball. *Ann. Probab.*, 33(2):480–513, 2005.

[10] B. BEAUZAMY : *Archimedes modern works.* Real life Mathematics. Société de Calcul Mathématique SA, 2012.

[11] Q. BERGER, M. BIRKNER et L. YUAN : Collective vs. individual behaviour for sums of i.i.d. random variables : appearance of the one-big-jump phenomenon. preprint `http://arxiv.org/abs/2303.12505v1`arXiv :2303.12505v1, 2023.

[12] N. H. BINGHAM, C. M. GOLDIE et J. L. TEUGELS : *Regular variation.*, vol. 27 de *Encycl. Math. Appl.* Cambridge etc. : Cambridge University Press, paperback ed. édn, 1989.

[13] T. BODINEAU : Modélisation de phénomènes aléatoires : introduction aux chaînes de markov et aux martingales. note de cours `http://www.cmap.polytechnique.fr/~bodineau/MAP432.pdf`, 2022.

[14] C. BORDENAVE, P. CAPUTO et D. CHAFAÏ : Spectrum of large random reversible Markov chains : two examples. *ALEA, Lat. Am. J. Probab. Math. Stat.*, 7:41–64, 2010.

[15] V. S. BORKAR : *Probability theory. An advanced course.* New York, NY : Springer-Verlag, 1995.

[16] S. BOUCHERON, G. LUGOSI et P. MASSART : *Concentration inequalities. A nonasymptotic theory of independence.* Oxford : Oxford University Press, corrected paperback edition édn, 2016.

[17] A. BOVIER : *Statistical mechanics of disordered systems. A mathematical perspective.* Cambridge : Cambridge University Press, reprint of the 2006 hardback ed. édn, 2012.

[18] S. BRAZITIKOS, A. GIANNOPOULOS, P. VALETTAS et B.-H. VRITSIOU : *Geometry of isotropic convex bodies*, vol. 196 de *Math. Surv. Monogr.* Providence, RI : American Mathematical Society (AMS), 2014.

[19] W. BRYC : A remark on the connection between the large deviation principle and the central limit theorem. *Stat. Probab. Lett.*, 18(4):253–256, 1993.

[20] T. CAI, J. FAN et T. JIANG : Distributions of angles in random packing on spheres. *J. Mach. Learn. Res.*, 14:1837–1864, 2013.

[21] R. CERF : *On Cramér's theory in infinite dimensions*, vol. 23 de *Panor. Synth.* Paris : Société Mathématique de France (SMF), 2007.

[22] R. CERF et P. PETIT : A short proof of Cramér's theorem in $\mathbb{R}$. *Am. Math. Mon.*, 118(10):925–931, 2011.

[23] D. CHAFAÏ : From Boltzmann to random matrices and beyond. *Ann. Fac. Sci. Toulouse, Math. (6)*, 24(4):641–689, 2015.

[24] D. CHAFAÏ, O. GUÉDON, G. LECUÉ et A. PAJOR : *Interactions between compressed sensing random matrices and high dimensional geometry*, vol. 37 de *Panor. Synth.* Paris : Société Mathématique de France (SMF), 2012.

[25] D. CHAFAÏ, C. GIRAUD et S. MÉLÉARD : *Aléatoire.* Éditions de l'École Polytechnique, 2013. Actes des journées X-UPS 2013.

[26] D. CHAFAÏ et P.-A. ZITT : *Probabilités : Préparation à l'agrégation interne.* `https://hal.archives-ouvertes.fr/hal-01374158`, 2020.

[27] T. M. COVER et J. A. THOMAS : *Elements of information theory.* Hoboken, NJ : John Wiley & Sons, 2nd ed. édn, 2006.

[28] I. CSISZÁR : A simple proof of Sanov's theorem. *Bull. Braz. Math. Soc. (N.S.)*, 37(4):453–459, 2006.

[29] N. CURIEN : Yet another proof of the strong law of large numbers. *Am. Math. Mon.*, 129(10):972–974, 2022.

[30] A. DEMBO et O. ZEITOUNI : *Large deviations techniques and applications.*, vol. 38 de *Stoch. Model. Appl. Probab.* Berlin : Springer, 2 édn, 2010.

[31] F. DEN HOLLANDER : *Large deviations*, vol. 14 de *Fields Inst. Monogr.* Providence, RI : AMS, American Mathematical Society, 2000.

[32] J.-D. DEUSCHEL et D. W. STROOCK : *Large deviations.* Boston, MA etc. : Academic Press, Inc., rev. ed. édn, 1989.

[33] P. DUPUIS et R. S. ELLIS : *A weak convergence approach to the theory of large deviations.* Wiley Ser. Probab. Stat. Chichester : John Wiley & Sons, 1997.

[34] B. EFRON : Student's t-test under symmetry conditions. *J. Amer. Statist. Assoc.*, 64:1278–1302, 1969.

[35] R. S. ELLIS : *Entropy, large deviations, and statistical mechanics.* Class. Math. Berlin : Springer, reprint of the 1985 edition édn, 2006.

[36] P. EMBRECHTS, C. KLÜPPELBERG et T. MIKOSCH : *Modelling extremal events for insurance and finance*, vol. 33 de *Appl. Math. (N. Y.)*. Berlin : Springer, 1997.

[37] W. FELLER : An introduction to probability theory and its applications. I. New York-London-Sydney : John Wiley and Sons, Inc. XVIII, 509 p. (1968)., 1968.

[38] W. FELLER : *An introduction to probability theory and its applications. Vol II. 2nd ed.* Wiley Ser. Probab. Math. Stat. John Wiley & Sons, Hoboken, NJ, 1971.

[39] G. FERRÉ : A subexponential version of Cramer's theorem. preprint arXiv:2206.05791, 2022.

[40] A. FIGALLI et F. GLAUDO : *An invitation to optimal transport, Wasserstein distances, and gradient flows.* EMS Textb. Math. Berlin : European Mathematical Society (EMS), 2021.

[41] P. J. FORRESTER : *Log-gases and random matrices.*, vol. 34 de *Lond. Math. Soc. Monogr. Ser.* Princeton, NJ : Princeton University Press, 2010.

[42] S. FRIEDLI et Y. VELENIK : *Statistical mechanics of lattice systems. A concrete mathematical introduction.* Cambridge : Cambridge University Press, 2018.

[43] C. GIRAUD : *Introduction to high-dimensional statistics*, vol. 168 de *Monogr. Stat. Appl. Probab.* Boca Raton, FL : CRC Press, 2nd edition édn, 2022.

[44] B. V. GNEDENKO et A. N. KOLMOGOROV : *Limit distributions for sums of independent random variables.* Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., revised édn, 1968. Translated from the Russian, annotated, and revised by K. L. Chung, With appendices by J. L. Doob and P. L. Hsu.

[45] N. GOZLAN : A characterization of dimension free concentration in terms of transportation inequalities. *Ann. Probab.*, 37(6):2480–2498, 2009.

[46] O. GUÉDON : Concentration phenomena in high dimensional geometry. *ESAIM, Proc.*, 44:47–60, 2014.

[47] O. GUÉDON, P. NAYAR et T. TKOCZ : Concentration inequalities and geometry of convex bodies. *In Analytical and probabilistic methods in the geometry of convex bodies*, p. 9–86. Warsaw : Polish Academy of Science, Institute of Mathematics, 2014.

[48] A. GUT : *Probability : a graduate course.* Springer Texts Stat. New York, NY : Springer, 2nd ed. édn, 2013.

[49] M. IBRAGIMOV, R. IBRAGIMOV et J. WALDEN : *Heavy-tailed distributions and robustness in economics and finance*, vol. 214 de *Lect. Notes Stat.* Cham : Springer, 2015.

[50] J.-M. KANTOR et L. GRAHAM : *Au nom de l'infini. Une histoire vraie de mysticisme religieux et de création mathématique.* Belin, Paris, 2010.

[51] B. KLARTAG et J. LEHEC : Bourgain's slicing problem and KLS isoperimetry up to polylog. *Geom. Funct. Anal.*, 32(5):1134–1159, 2022.

[52] M. LEDOUX : Concentration of measure and logarithmic Sobolev inequalities. *In Séminaire de probabilités XXXIII*, p. 120–216. Berlin : Springer, 1999.

[53] M. LEDOUX : *The concentration of measure phenomenon*, vol. 89 de *Math. Surv. Monogr.* Providence, RI : American Mathematical Society (AMS), 2001.

[54] M. LEDOUX et M. TALAGRAND : *Probability in Banach spaces. Isoperimetry and processes.* Class. Math. Berlin : Springer, reprint of the 1991 hardback ed. édn, 2011.

[55] G. LETAC : From archimedes to statistics : the area of the sphere. Lecture notes for evenings at the Rovinj summer academy, available online, 2004.

[56] G. LIVAN, M. NOVAES et P. VIVO : *Introduction to random matrices. Theory and practice*, vol. 26 de *SpringerBriefs Math. Phys.* Cham : Springer, 2018.

[57] N. F. MARTIN et J. W. ENGLAND : Mathematical theory of entropy. Foreword by James K. Brooks. *In Encyclopedia of Mathematics and its Applications*, vol. 12. Addison-Wesley Publishing Company, 1981.

[58] M. L. MEHTA : *Random matrices.* Amsterdam : Elsevier, 3rd ed. édn, 2004.

[59] J. A. MINGO et R. SPEICHER : *Free probability and random matrices*, vol. 35 de *Fields Inst. Monogr.* Toronto : The Fields Institute for Research in the Mathematical Sciences ; New York, NY : Springer, 2017.

[60] J. NAIR, A. WIERMAN et B. ZWART : *The fundamentals of heavy tails. Properties, emergence, and estimation*, vol. 53 de *Camb. Ser. Stat. Probab. Math.* Cambridge : Cambridge University Press, 2022.

[61] P. PETIT : Cramér's theorem in Banach spaces revisited. *In Séminaire de probabilités XLIX*, p. 455–473. Cham : Springer, 2018.

[62] V. V. PETROV : *Limit theorems of probability theory. Sequences of independent random variables*, vol. 4 de *Oxf. Stud. Probab.* Oxford : Clarendon Press, 1995.

[63] E. J. G. PITMAN : On the behavior of the characteristic function of a probability distribution in the neighborhood of the origin. *J. Austral. Math. Soc.*, 8:423–443, 1968.

[64] J. PITMAN et N. ROSS : Archimedes, Gauss, and Stein. *Notices Am. Math. Soc.*, 59(10):1416–1421, 2012.

[65] M. POTTERS et J.-P. BOUCHAUD : *A first course in random matrix theory : for physicists, engineers and data scientists.* Cambridge : Cambridge University Press, 2021.

[66] F. RASSOUL-AGHA et T. SEPPÄLÄINEN : *A course on large deviations with an introduction to Gibbs measures*, vol. 162 de *Grad. Stud. Math.* Providence, RI : American Mathematical Society (AMS), 2015.

[67] S. I. RESNICK : *Heavy-tail phenomena. Probabilistic and statistical modeling.* Springer Ser. Oper. Res. Financ. Eng. New York, NY : Springer, 2007.

[68] S. I. RESNICK : *Extreme values, regular variation and point processes.* Springer Ser. Oper. Res. Financ. Eng. New York, NY : Springer, reprint of the 1987 original édn, 2008.

[69] B. RIDER et B. VIRÁG : The noise in the circular law and the Gaussian free field. *Int. Math. Res. Not.*, 2007(2):32, 2007. Id/No rnm006.

[70] C. P. ROBERT : *The Bayesian choice. From decision-theoretic foundations to computational implementation*. Springer Texts Stat. New York, NY : Springer, 2nd ed., 1st paperback ed. édn, 2007.

[71] C. P. ROBERT et G. CASELLA : *Monte Carlo statistical methods*. Springer Texts Stat. New York, NY : Springer, 2nd ed. édn, 2004.

[72] F. SANTAMBROGIO : *Optimal transport for applied mathematicians. Calculus of variations, PDEs, and modeling*, vol. 87 de *Prog. Nonlinear Differ. Equ. Appl.* Cham : Birkhäuser/Springer, 2015.

[73] G. SCHECHTMAN et J. ZINN : On the volume of the intersection of two ℓ_n^p balls. *Proc. Amer. Math. Soc.*, 110:217–224, 1990.

[74] N. N. T. TALEB : *Statistical Consequences of Fat Tails : Real World Preasymptotics, Epistemology, and Applications*. STEM Academic Press, 2023.

[75] H. TOUCHETTE : The large deviation approach to statistical mechanics. *Phys. Rep.*, 478(1-3):1–69, 2009.

[76] P. TOULOUSE : *Thèmes de probabilités et statistique*. Dunod, 1999.

[77] R. van HANDEL : Probability in High Dimension. Lecture notes `https://web.math.princeton.edu/~rvan/`, 2016.

[78] R. VERSHYNIN : *High-dimensional probability. An introduction with applications in data science*, vol. 47 de *Camb. Ser. Stat. Probab. Math.* Cambridge : Cambridge University Press, 2018.

[79] C. VILLANI : *Topics in optimal transportation*, vol. 58 de *Grad. Stud. Math.* Providence, RI : American Mathematical Society (AMS), 2003.

[80] M. J. WAINWRIGHT : *High-dimensional statistics. A non-asymptotic viewpoint*, vol. 48 de *Camb. Ser. Stat. Probab. Math.* Cambridge : Cambridge University Press, 2019.

[81] M. ZINSMEISTER : *Formalisme thermodynamique et systèmes dynamiques holomorphes*, vol. 4 de *Panoramas et Synthèses*. Société Mathématique de France. Paris, 1996.

```
# Julia Code (c) Djalil Chafai 2021 - GNU Public License GPL v3

using Plots # for plot(), savefig()

# Define the 2D Dyson Ornstein-Uhlenbeck process
Base.@kwdef mutable struct DOU2D
    n::Int = 2         # number of particles
    b::Float64 = 2.    # repulsion parameter
    T::Float64 = 10.   # terminal time
    dt::Float64 = 1E-3 # time increment
    r::Float64  = 1.   # initial conditions uniform grid on horizontal segment [-r,r]-3*im
    m = floor(Int, T / dt) # number of times
    x::Array{Complex{Float64},2} = zeros(Complex{Float64},m,n)
end # end struct

function compute!(X::DOU2D)
    X.x[1,:] = range(-X.r, X.r, length = X.n) .- 3 * im
    for i = 2:X.m
     dB = (randn(1,X.n) + im * randn(1,X.n))/sqrt(2)
     X.x[i,:] = copy(X.x[i-1,:])
     for j in 1:X.n # particles
         X.x[i,j] += sqrt(2/X.n) * dB[j] * sqrt(X.dt)
         X.x[i,j] += -2 * X.x[i-1,j] * X.dt
         for k in 1:X.n
             if (j == k) continue end
             X.x[i,j] += X.b * X.dt / (X.n * (conj(X.x[i-1,j] - X.x[i-1,k])))
         end # for
     end # for
    end # for
end # function

# process and graphics
dou = DOU2D(n = 66, r = 3., T = 5., dt = 1E-4)
compute!(dou)
#
pdou2d = plot()
for j in 1:dou.n # particles
    plot!(pdou2d, real(dou.x[:,j]), imag(dou.x[:,j]), aspect_ratio =:equal, legend = false)
end # for
savefig(pdou2d,"dou2d.png")
savefig(pdou2d,"dou2d.svg")
# EOF
```