Probabilités M1

Max Fathi[1]

March 30, 2023

[1]LJLL & LPSM, Université de Paris, France

Contents

Part I

Processus en temps discret

Chapter 1

Notions de processus

1.1 Généralités sur les processus

On se place sur un espace de probabilités $(\Omega, \mathcal{F}, \mathbb{P})$.

Par définition, un processus aléatoire en temps discret est une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ toutes définies sur $(\Omega, \mathcal{F}, \mathbb{P})$.

Le paramètre n joue le rôle du temps. Le but est en général de modéliser un phénomène qui évolue, comme par exemple le cours de la bourse, la propagation d'une épidémie, ou encore l'évolution d'une opinion mesurée semaine après semaine par un sondage.

NB. *Il est important de distinguer la loi d'un processus $(X_n)_{n\in\mathbb{N}}$ de la suite des lois des X_n. La première notion englobe l'information des corrélations entre les variables, mais pas la deuxième.*

Définition 1.1.1 (Filtrations)**.** *Une filtration de $(\Omega, \mathcal{F}, \mathbb{P})$ est une suite croissante $(\mathcal{F}_n)_{n\in\mathbb{N}}$ de sous-tribus de $\mathcal{F}$, i.e. $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 ... \subset \mathcal{F}$. On dit aussi que $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, \mathbb{P})$ est un espace de probabilité filtré.*

Un processus $(X_n)_{n\in\mathbb{N}}$ est dit adapté à la filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ si pour tout $n \in \mathbb{N}$, X_n est $\mathcal{F}_n$-mesurable.

Informellement, on peut penser à la sous-tribu $\mathcal{F}_n$ comme étant l'information acquise au temps n.

Exemple 1.1.1.

1. *Si $(X_n)_{n\in\mathbb{N}}$ est une suite (quelconque) de variables aléatoires définies sur $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{F}_n^X := \sigma(X_0, ..., X_n)$ définit une filtration, appelée filtration canonique du processus $(X_n)_{n\in\mathbb{N}}$. C'est la plus petite filtration rendant le processus $(X_n)_{n\in\mathbb{N}}$ adapté.*

2. *Prenons $\Omega = [0, 1)$, $\mathcal{F}$ la tribu borelienne et $\mathbb{P}$ la mesure de Lebesgue. Posons $\mathcal{F}_n := \sigma\left(\left[\frac{i-1}{2^n}, \frac{i}{2^n}\right), i = 1, ..., 2^n\right)$. C'est une filtration de $[0, 1)$, appelée filtration dyadique.*

NB. *Dans toute la suite, il y aura implicitement un espace de probabilité filtré, qui ne sera pas toujours rendu explicite, et les notions seront bien entendu relatives à cet espace.*

1.2 Quelques exemples de base

1.2.1 Marches aléatoires

L'exemple le plus emblématique de processus aléatoire est celui de la marche aléatoire en dimension un :

Exemple 1.2.1 (Marche aléatoire simple sur $\mathbb{Z}$). *On part du point* 0*, et à chaque instant* n *on choisit une direction uniformément au hasard (droite ou gauche), et on fait un pas dans cette direction.*

Plus exactement, $X_0 = 0$*, et* $(\varepsilon_n)_{n\in\mathbb{N}}$ *est une suite de variables de Rademacher, c'est à dire uniformes sur* $\{-1, 1\}$*, et* $X_{n+1} = X_n + \varepsilon_{n+1}$*. De manière équivalente,* $X_n = \sum_{i=1}^{n} \varepsilon_i$*.*

La filtration canonique de ce processus est donnée par $\sigma(X_0, ..., X_n) = \sigma(\varepsilon_0, ...\varepsilon_n)$.

De manière similaire, on peut définir une marche aléatoire biaisée de paramètre $p \in [0, 1]$ en sommant des variables ε_i avec $\mathbb{P}(\varepsilon_i = 1) = 1 - \mathbb{P}(\varepsilon_i = -1) = p$. On parlera alors de marche aléatoire simple asymmétrique de paramètre p.

On peut généraliser cet exemple à des lois de saut arbitraire, et en dimension arbitraire.

Exemple 1.2.2 (Marche aléatoires sur $\mathbb{R}^d$). *Soit* (Y_i) *une suite de v.a. i.i.d. sur* $\mathbb{R}^d$*, de loi* μ*. La marche aléatoire avec condition initiale* X_0 *et loi de saut* μ *est donnée par* $X_n := X_0 + \sum_{i=1}^{n} Y_i$*.*

On peut également considérer des marches aléatoires sur des graphes plus généraux que $\mathbb{Z}$ ou $\mathbb{Z}^d$:

Exemple 1.2.3 (Marches aléatoires sur les graphes). *Soit* $G = (S, A)$ *un graphe. Une marche aléatoire simple sur le graphe* G *est une suite de variables aléatoires* $(X_n)_{n\in\mathbb{N}}$ *à valeurs dans* G*, telle que conditionnellement à* $\{X_n = x\}$ *la loi de* X_{n+1} *est uniforme sur l'ensemble des voisins de* x*, c'est à dire*

$$\mathbb{P}(X_{n+1} = y | X_n = x) = \frac{\mathbb{1}_{(x,y)\in A}}{|\{z; (x, z) \in A\}|}.$$

1.2.2 Processus de Galton-Watson

Les processus de Galton-Watson sont utilisés pour modéliser l'évolution du nombre d'individus d'une espèce asexuée, où chaque individu d'une génération donnée se reproduit indépendamment des autres individus (duplication de cellules...).

Pour les définir, on considère une suite $(X_n^k)_{k,n\in\mathbb{N}^*}$ de v.a. i.i.d. à valeurs dans $\mathbb{N}$, où X_n^k représente le nombre d'enfants du k-ième individu de la génération n, ainsi qu'une taille de population initiale donnée par une v.a. N_0. Alors le processus est donné par

$$N_{n+1} := \sum_{k=1}^{N_n} X_n^k.$$

La filtration canonique de ce processus $\sigma(N_0, ..., N_n)$ est différente de $\sigma((X_i^k)_{i\leqslant n})$, car cette dernière contient plus d'information que juste les tailles totales des populations.

Exercice 1.2.1. *Calculer l'espérance et la variance de N_n en fonction de celles de X_1^1.*

Solution 1.2.1. *On a la relation de récurrence*

$$\mathbb{E}[N_{n+1}] = \mathbb{E}[\mathbb{E}[N_{n+1}|N_n]] = \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{N_n} X_n^k \mid N_n\right]\right] = \mathbb{E}[N_n]\mathbb{E}[X_1^1].$$

Donc $\mathbb{E}[N_n] = \mathbb{E}[N_0]\mathbb{E}[X_1^1]^n$.

De même, en posant $m = \mathbb{E}[X_1^1]$, *on a*

$$\mathbb{E}[N_{n+1}^2] = \mathbb{E}[\mathbb{E}[N_{n+1}^2|N_n]] = \mathbb{E}[N_n]\mathbb{E}[(X_1^1)^2] + \mathbb{E}[N_n(N_n-1)]\mathbb{E}[X_1^1]^2 = E[N_n^2]m^2 + \mathbb{E}[N_n]\operatorname{Var}(X_1^1).$$

Posons $u_n = \mathbb{E}[N_n^2]m^{-2n}$. *On a*

$$u_{n+1} = u_n + \mathbb{E}[N_0]\operatorname{Var}(X_1^1)m^{-n-2}$$

Donc

$$u_n = \mathbb{E}[N_0^2] + \mathbb{E}[N_0]\operatorname{Var}(X_1^1)m^{-2}\frac{1-m^{-n}}{1-m^{-1}}$$

et

$$\mathbb{E}[N_n^2] = \mathbb{E}[N_0^2]m^{2n} + \mathbb{E}[N_0]\operatorname{Var}(X_1^1)\frac{m^{2n}-m^n}{m^2-m}.$$

D'où

$$\operatorname{Var}(N_n) = \operatorname{Var}(N_0)\mathbb{E}[X_1^1]^{2n} + \mathbb{E}[N_0]\operatorname{Var}(X_1^1)\frac{\mathbb{E}[X_1^1]^{2n} - \mathbb{E}[X_1^1]^n}{\mathbb{E}[X_1^1]^2 - \mathbb{E}[X_1^1]}.$$

L'espérance de X gouverne donc la croissance de la taille de la population.

1.2.3 Files d'attente

Un exemple de modèle (très basique) pour le nombre de personnes en train d'attendre dans une file est celui d'une suite de variables aléatoires sur $\mathbb{N}$, où X_n représente le nombre de personnes dans la file à l'instant n, et avec la loi conditionnelle

$$\mathbb{P}(X_{n+1} = k+1|X_n = k) = p;\ \mathbb{P}(X_{n+1} = k-1|X_n = k) = q;\ \mathbb{P}(X_{n+1} = k|X_n = k) = r$$

avec $p + q + r = 1$.

On peut aussi voir ce processus comme une marche aléatoire (possiblement asymétrique) sur $\mathbb{N}$.

Pour modéliser le nombre de gens attendant à des caisses de supermarché, on peut considérer plusieurs files d'attente en interaction, où les probabilités d'arriver à un instant donné sont dépendantes du nombre de gens dans chaque file (par exemple, il ne peut arriver quelqu'un que dans la file avec le moins de monde).

1.3 Temps d'arrêt

Définition 1.3.1. *Une v.a. $T : \Omega \longrightarrow \bar{N} = \mathbb{N} \cup \{\infty\}$ est appelé temps d'arrêt de la filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ si pour tout n on a $\{T = n\} \in \mathcal{F}_n$.*

Interprétation : on joue à un jeu modélisé par des v.a. Un temps d'arrêt représente un instant auquel on décide de s'arrêter, qui doit être fonction de l'information connue à ce moment là (i.e. on ne connaît pas le futur).

Remarque 1.3.1. *Il est équivalent de demander $\{T \leqslant n\} \in \mathcal{F}_n$, on pourra utiliser ces deux définitions de manière interchangeable. En effet, si pour tout n $\{T \leqslant n\} \in \mathcal{F}_n$, alors $\{T = n\} = \{T \leqslant n\} \cap \{T \leqslant n-1\}^c \in \mathcal{F}_n$, en utilisant la propriété de filtration. Réciproquement, si pour tout n on a $\{T = n\} \in \mathcal{F}_n$ alors $\{T \leqslant n\} = \cup_{k\text{`}n}\{T = k\} \in \mathcal{F}_n$. De même, on peut utiliser $\{T > n\}$ dans la définition au lieu de $\{T = n\}$, par passage au complémentaire.*

La valeur ∞ est autorisée. En effet, $\{T = \infty\} = \left(\bigcup_{n\in\mathbb{N}}\{T = n\}\right)^c$, donc $\{T = \infty\} \in \sigma\left(\bigcup_{n\in\mathbb{N}} \mathcal{F}_n\right)$.

On aura souvent à prouver $T < \infty$ p.s. dans des cas concrets.

Exemple 1.3.1.

1. *$T = k \in \mathbb{N}$ fixé est un temps d'arrêt.*
2. *Si (X_n) est un processus adapté à valeurs dans $\mathbb{R}^d$ et A un borelien de $\mathbb{R}^d$, alors*
$$T_A := \inf\{n \in \mathbb{N}; Y_n \in A\}$$
est un temps d'arrêt, appelé temps d'atteinte de A. En effet,
$$\{T_A = n\} = \{X_0 \notin A, X_1 \notin A, ..., X_{n-1} \notin A, X_n \in A\} \in \sigma(X_0, ..., X_n) \subset \mathcal{F}_n.$$
3. *En revanche, pour N fixé, $L_A = \sup\{n \leqslant N, X_n \in A\}$ n'est pas un temps d'arrêt en général. En effet, pour déterminer que le processus ne reviendra pas dans l'ensemble A après un temps donné demanderait de connaître le futur. Interprétation : il n'est pas possible de vendre des actions à leur prix maximal sur l'année.*

Regardons un exemple plus en détails. Soit (X_n) une marche aléatoire simple symétrique sur $\mathbb{Z}$, issue de 0. Soit $\varepsilon_n := X_n - X_{n-1}$, de sorte que $(\varepsilon_n)_{n\in\mathbb{N}^*}$ est une suite de variables iid uniformes sur $\{-1, 1\}$. On considère le temps d'arrêt $T := \inf\{n \geqslant 1; X_n = 0\}$, appelé temps de premier retour en 0. Montrons que $T < \infty$ p.s.

Tout d'abord, nous allons montrer que pour tout $p \in \mathbb{N}$, on a $\mathbb{P}(-p \leqslant \inf X_n \leqslant \sup X_n \leqslant p) = 0$. Notons B_p cet évènement. Soit $k > 2p$, et posons

$$A_j = \{\varepsilon_{kj+1} = 1, \varepsilon_{kj+2} = 1, ..., \varepsilon_{k(j+1)} = 1\}.$$

On a $\bigcup A_j \subset B_p^c$. Les A_j sont indépendants, et de probabilité strictement positive. Donc, par le lemme de Borel-Cantelli, $\mathbb{P}(\bigcup A_j) = 1$, et donc $\mathbb{P}(B_p) = 0$. On a donc

$$\begin{aligned}
&\mathbb{P}[\inf X_n < -p] + \mathbb{P}[\sup X_n > p] \geqslant 1 \\
&\Rightarrow \mathbb{P}[\inf X_n = -\infty] + \mathbb{P}[\sup X_n = +\infty] \geqslant 1 \\
&\Rightarrow \mathbb{P}[\inf X_n = -\infty] = \mathbb{P}[\sup X_n = +\infty] > 0.
\end{aligned}$$

où pour la dernière étape on a utilisé la symétrie du processus. Or

$$\{\sup X_n = +\infty\} \subset \bigcap_{k\geqslant 0} \sigma(X_k, X_{k+1}, ...)$$

donc on peut utiliser la loi du 0-1 de Kolmogorov, et en déduire que $\sup X_n = +\infty$ presque sûrement, et de même $\inf X_n = -\infty$ presque sûrement. On a donc bien $T < \infty$ p.s., puisque pour passer de valeurs positives à des valeurs négatives le processus doit repasser par l'origine.

Proposition 1.3.2. *1. Si S et T sont deux temps d'arrêt, alors $\max(S,T)$ et $\min(S,T)$ en sont aussi.*

2. Si $(T_k)_{k\in\mathbb{N}}$ est une suite de temps d'arrêt, alors $\inf T_k$, $\sup T_k$, $\liminf T_k$ et $\limsup T_k$ en sont aussi.

Proof. 1. On a

$$\{\min(S,T) \leqslant n\} = \{S \leqslant n\} \cup \{T \leqslant n\} \in \mathcal{F}_n;$$

$$\{\max(S,T) \leqslant n\} = \{S \leqslant n\} \cap \{T \leqslant n\} \in \mathcal{F}_n.$$

2. Nous allons montrer que $\inf T_k$ et $\liminf T_k$ sont des temps d'arrêt, les deux autres cas se prouvent de manière similaire.

$$\{\inf T_k \leqslant n\} = \bigcup_k \{T_k \leqslant n\} \in \mathcal{F}_n;$$

$$\{\liminf T_k \leqslant n\} = \bigcap_j \left(\bigcup_{k\geqslant j} \{T_k \leqslant n\} \right) \in \mathcal{F}_n.$$

□

Définition 1.3.3. *Soit T un temps d'arrêt. La tribu du passé jusqu'à l'instant T est*

$$\mathcal{F}_T := \{A \in \mathcal{F};\ \forall n \in \mathbb{N},\ A \cap \{T = n\} \in \mathcal{F}_n\}.$$

$\mathcal{F}_T$ est bien une tribu, et si $T = k$, on a bien $\mathcal{F}_T = \mathcal{F}_k$. En effet,

$$\Omega \cap \{T = n\} = \{T = n\} \in \mathcal{F}_n,$$

et comme

$$A \in \mathcal{F}_T \Leftrightarrow \forall n,\ A \cap \{T = n\} \in \mathcal{F}_n$$

et $A^c \cap \{T = n\} = \{T = n\} \backslash (A \cap \{T = n\}) \in \mathcal{F}_n$, on a bien $A \in \mathcal{F}_T \Rightarrow A^c \in \mathcal{F}_T$.

Enfin, si (A_k) est une suite d'évenment de $\mathcal{F}_T$, alors

$$\bigcup_k (A_k \cap \{T = n\}) = \left(\bigcup_k A_k \right) \cap \{T = n\} \in \mathcal{F}_n.$$

Proposition 1.3.4. *Soient S et T deux temps d'arrêt vérifiant $S \leqslant T$. Alors $\mathcal{F}_S \subset \mathcal{F}_T$.*

Proof. Soit $A \in \mathcal{F}_S$. On a

$$A \cap \{T = n\} = A \cap \{S \leqslant n\} \cap \{T = n\} = \bigcup_{k=0}^{n} (A \cap \{S = k\}) \cap \{T = n\},$$

or $A \cap \{S = k\} \in \mathcal{F}_k \subset \mathcal{F}_n$, donc on a bien $A \cap \{T = n\} \in \mathcal{F}_n$. □

Proposition 1.3.5. *Soit (Y_n) un processus adapté à valeurs réelles, et T un temps d'arrêt. La variable aléatoire*

$$\mathbb{1}_{\{T<\infty\}} Y_T(\omega) := \begin{cases} Y_n(\omega) \text{ si } T(\omega) = n; \\ 0 \text{ si } T(\omega) = \infty \end{cases}$$

est $\mathcal{F}_T$-mesurable.

Si $T < \infty$ p.s., on notera Y_T cette variable aléatoire, pour simplifier les notations. En particulier, une variable aléatoire de la forme $Y_{n\wedge T}$ est $\mathcal{F}_{n\wedge T}$ mesurable, et donc $\mathcal{F}_T$ mesurable via la Proposition 1.3.4.

Proof. Soit B un borelien. On a

$$\{\mathbb{1}_{T<\infty} Y_T \in B\} \cap \{T = n\} = \{Y_n \in B\} \cap \{T = n\} \in \mathcal{F}_n,$$

et donc on a bien $\{\mathbb{1}_{T<\infty} Y_T \in B\} \in \mathcal{F}_T$. □

1.4 Rappels sur la convergence de variables aléatoires

Définition 1.4.1 (Convergence presque sûre)**.** *Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ converge presque sûrement vers une variable aléatoire X (définie sur le même espace) si*

$$\mathbb{P}(\{\omega \in \Omega; X_n(\omega) \longrightarrow X(\omega)\}) = 1.$$

Définition 1.4.2 (Convergence L^p)**.** *Soit $p \geqslant 1$. Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ converge au sens L^p vers une variable aléatoire X (définie sur le même espace) si*

$$\mathbb{E}[|X_n - X|^p] \longrightarrow 0.$$

Définition 1.4.3 (Convergence en probabilité)**.** *Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ définies sur un même espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ converge au sens L^p vers une variable aléatoire X (définie sur le même espace) si pour tout $\varepsilon > 0$ on a*

$$\mathbb{P}(|X_n - X| > \varepsilon) \longrightarrow 0.$$

Définition 1.4.4 (Convergence en loi)**.** *Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ converge en loi vers une variable aléatoire X si pour toute fonction continue bornée on a*

$$\mathbb{E}[f(X_n)] \longrightarrow \mathbb{E}[f(X)].$$

NB. *La convergence en loi est une notion de convergence d'une nature différente que les trois premières : elle ne dépend que de la suite des lois, sans demander à ce que les variables vivent sur un même espace $(\Omega, \mathcal{F}, \mathbb{P})$, ou qu'on tiennent compte de corrélations.*

Proposition 1.4.5 (Relations entre les notions de convergence). *1. La convergence L^p implique la convergence en probabilité;*

2. La convergence p.s. implique la convergence en probabilité;

3. La convergence en probabilité implique la convergence en loi.

Exemple 1.4.1. *1. Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de variables indépendantes, avec $\mathbb{P}(X_n = 0) = 1 - 1/n$ et $\mathbb{P}(X_n = n) = 1/n$. Alors X_n converge en probabilité (et même presque sûrement) vers 0, mais pas dans L^p, pour n'importe quel $p \geqslant 1$.*

2. Soit $\Gamma = \{(n, j); 0 < j \leqslant n\}$, qu'on numérote dans l'ordre lexicographique. Pour $m = (n, j) \in \Gamma$, on pose $X_m = \mathbb{1}_{((j-1)/n, j/n]}$. Les X_m forment une suite de variables aléatoires sur $((0, 1], \mathrm{Leb})$. Comme la largeur des intervalles tend vers 0, la suite $(X_m)_{m\in\mathbb{N}}$ converge en probabilité vers 0. Toutefois, il existe des indices arbitrairement grand tels que $X_m = 1$, et donc $\limsup X_m = 1$ p.s. En particulier, $(X_m)_{m\in\mathbb{N}}$ ne converge pas p.s. vers 0.

3. N'importe quelle suite (X_n) de variables aléatoires iid et non constante p.s. converge en loi (puisque la loi ne change jamais), mais ne converge pas p.s. En effet, la convergence en loi ne tient pas compte des corrélations entre les X_n, mais la convergence p.s. peut en dépendre.

Chapter 2

Martingales

Dans toute cette partie, on ne considérera que des processus à valeurs réelles.

2.1 Définitions et premières propriétés

Définition 2.1.1. *Soit $(X_n)_{n\in\mathbb{N}}$ un processus adapté à une filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, et intégrable.*

1. *$(X_n)_{n\in\mathbb{N}}$ est une martingale si $\forall n$ on a $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$.*
2. *$(X_n)_{n\in\mathbb{N}}$ est une surmartingale si $\forall n$ on a $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leqslant X_n$.*
3. *$(X_n)_{n\in\mathbb{N}}$ est une sous-martingale si $\forall n$ on a $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \geqslant X_n$.*

Lorsque la filtration est la filtration canonique du processus, on parlera de (sur/sous)-martingale, sans plus de précisions.

Une martingale est donc un processus dont en moyenne on n'attend pas d'évolution. On interprète souvent une martingale comme un jeu équitable : si X_n est l'avoir du joueur au temps n, et $\mathcal{F}_n$ l'information dont le joueur dispose, alors $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n$ signifie qu'en moyenne le joueur ne perd ni ne gagne. Une surmartingale est un jeu défavorable (du point de vue du joueur), et une sous-martingale un jeu favorable.

Remark 2.1.1.
1. *Par récurrence sur $\ell - n$, on peut montrer que si $(X_n)_{n\in\mathbb{N}}$ est une martingale, alors pour tout $\ell \geqslant n \geqslant 0$ on a $\mathbb{E}[X_\ell|\mathcal{F}_n] = X_n$.*
2. *Par conséquence, on a aussi $\mathbb{E}[X_\ell] = \mathbb{E}[X_n]$, i.e. l'espérance d'une martingale reste constante au cours du temps.*
3. *Le processus $(X_n)_{n\in\mathbb{N}}$ est une sous-martingale ssi $(-X_n)_{n\in\mathbb{N}}$ est une surmartingale. En conséquence, on énoncera souvent les théorèmes pour les surmartingales, et l'analogue pour les sous-martingales suivra automatiquement.*

Exemple 2.1.1. *La marche aléatoire simple de paramètre p est une martingale si $p = 1/2$, une sous martingale si $p \geqslant 1/2$, et une surmartingale si $p \leqslant 1/2$.*

Pour le montrer, on calcule

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}[X_n + \varepsilon_{n+1}|\mathcal{F}_n] = X_n + \mathbb{E}[\varepsilon_{n+1}] = X_n + 2p - 1.$$

Plus généralement, un processus de la forme $X_n = \sum_{i=1}^n Y_i$ avec les Y_i iid à valeurs réelles est une martingale ssi $\mathbb{E}[Y] = 0$, une surmartingale ssi $\mathbb{E}[Y] \leqslant 0$, et une sous-martingale ssi $\mathbb{E}[Y] \geqslant 0$.

Exemple 2.1.2. *Le processus de population de Galton-Watson est une martingale si $\mathbb{E}[X_i^j] = 1$, une sous-martingale si $\mathbb{E}[X_i^j] \geqslant 1$ et une surmartingale si $\mathbb{E}[X_i^j] \leqslant 1$.*

En effet, si $N_{n+1} = \sum_{i=1}^{N_n} X_i^{n+1}$ avec les X_i^k iid à valeurs entières et positives, alors $\mathbb{E}[N_{n+1}|\mathcal{F}_n] = N_n \mathbb{E}[X_1^1]$.

Exemple 2.1.3. *Soit $(\mathcal{F}_n)_{n\in\mathbb{N}}$ une filtration et X une v.a. L^1 et mesurable par rapport à $\mathcal{F}_\infty = \sigma(\cup_n \mathcal{F}_n)$. Alors le processus défini par $X_n := \mathbb{E}[X|\mathcal{F}_n]$ est une martingale. Un processus de cette forme est appelé une martingale fermée.*

Cette propriété est une conséquence immédiate de la croissance des filtrations : $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}_{n+1}]|\mathcal{F}_n] = \mathbb{E}[X|\mathcal{F}_n] = X_n$. On verra plus tard que beaucoup de martingales sont implicitement de cette forme.

Exemple 2.1.4. *Si (X_n) est une suite décroissante et adaptée de variables aléatoires intégrables, alors c'est une surmartingale : $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leqslant \mathbb{E}[X_n|\mathcal{F}_n] = X_n$.*

Exercice 2.1.1. *Montrer que si (X_n) est une martingale avec $X_n \in L^2$ pour tout $n \geqslant 0$, alors $\forall m, n$ on a*

$$\mathbb{E}[(X_{n+m} - X_n)^2] = \sum_{k=n}^{n+m-1} \mathbb{E}[(X_{k+1} - X_k)^2].$$

Solution 2.1.1. *Procédons par récurrence sur m. L'identité est trivialement vraie pour $m = 1$. Supposons qu'elle est vraie pour un certain $m \geqslant 1$. On a alors*

$$\mathbb{E}[(X_{n+m+1}-X_n)^2] = \mathbb{E}[(X_{n+m}-X_n)^2]+\mathbb{E}[(X_{n+m+1}-X_{n+m})^2]+2\mathbb{E}[(X_{n+m+1}-X_{n+m})(X_{n+m}-X_n)],$$

donc il suffit de montrer que le dernier terme à droite est nul. En conditionnant et en utilisant la propriété de martingale

$$\begin{aligned}\mathbb{E}[(X_{n+m+1} - X_{n+m})(X_{n+m} - X_n)] &= \mathbb{E}[\mathbb{E}[(X_{n+m+1} - X_{n+m})(X_{n+m} - X_n)|\mathcal{F}_{n+m}]] \\ &= \mathbb{E}[(X_{n+m} - X_n)\mathbb{E}[(X_{n+m+1} - X_{n+m})|\mathcal{F}_{n+m}]] = 0,\end{aligned}$$

ce qui conclut la preuve.

Une variante de cet argument montre que les incréments d'une martingale forment une famille orthogonale dans L^2. L'identité finale est alors une conséquence du théorème de Pythagore.

Exercice 2.1.2. *Soit U une variable aléatoire uniforme sur $[0, 1)$. Pour tout n, soit I_n l'intervalle (aléatoire) de la forme $[i/2^n, (i + 1)/2^n)$ qui contient U. Soit f une fonction continue bornée sur $[0, 1]$, et on pose*

$$X_n := \frac{1}{|I_n|} \int_{I_n} f dx.$$

1. *Montrer que le processus $(I_n)_{n\in\mathbb{N}}$ est adapté par rapport à la filtration dyadique;*

2. *Montrer que $(X_n)_{n\in\mathbb{N}}$ est une martingale par rapport à cette filtration, et qu'elle converge p.s. vers $f(U)$.*

Solution 2.1.2. 1. *On a bien $\{U \in [i/2^n, (i+1)/2^n)\} \in \mathcal{F}_n$, et les évènements de cette forme suffisent à déterminer I_n.*

2. *Conditionnellement à $I_n = [i/2^n, (i+1)/2^n)$, avec probabilité $1/2$ on a $I_{n+1} = [2i/2^{n+1}, (2i+1)/2^{n+1})$, et avec probabilité $1/2$ on a $I_{n+1} = [(2i+1)/2^{n+1}, (2i+2)/2^{n+1})$. On a donc*

$$\mathbb{E}[X_{n+1}|I_n] = \frac{2^{n+1}}{2}\int_{2i/2^{n+1}}^{(2i+1)/2^{n+1}} f dx + \frac{2^{n+1}}{2}\int_{(2i+1)/2^{n+1}}^{(2i+2)/2^{n+1}} f dx = X_n.$$

Pour la convergence p.s., comme f est continue sur le compact $[0,1]$, elle est uniformément continue. Soit $\varepsilon > 0$ et $\delta > 0$ tel que si $|x-y| < \delta$, alors $|f(x) - f(y)| < \varepsilon$. Alors si $2^{-n} < \delta$, pour tout $x \in I_n$, on a $|x-y| < \delta$, et donc

$$\left| |I_n|^{-1} \int_{I_n} f dx - f(U) \right| < \varepsilon.$$

On en déduit la convergence p.s. de X_n vers $f(U)$.

2.2 Quelques outils

2.2.1 Transformations de martingales

Proposition 2.2.1. *Soit (X_n) un processus adapté et φ une fonction convexe telle que $\mathbb{E}[|\varphi(X_n)|] < \infty$ pour tout n.*

1. *Si $(X_n)_{n\in\mathbb{N}}$ est une martingale, alors $(\varphi(X_n))_{n\in\mathbb{N}}$ est une sous-martingale.*
2. *Si φ est de plus croissante et si $(X_n)_{n\in\mathbb{N}}$ est une sous-martingale, alors $(\varphi(X_n))_{n\in\mathbb{N}}$ est une sous-martingale.*

Exemple 2.2.1. *Si $(X_n)_{n\in\mathbb{N}}$ est une martingale, alors $(X_n^2)_{n\in\mathbb{N}}$ est une sous-martingale. En particulier, la variance d'une martingale croît avec le temps.*

Proof. 1. D'après l'inégalité de Jensen,

$$\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n] \geqslant \varphi(\mathbb{E}[X_{n+1}|\mathcal{F}_n]) = \varphi(X_n).$$

2. Toujours en appliquant l'inégalité de Jensen, et la monotonie

$$\mathbb{E}[\varphi(X_{n+1})|\mathcal{F}_n] \geqslant \varphi(\mathbb{E}[X_{n+1}|\mathcal{F}_n]) \geqslant \varphi(X_n)$$

car $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \geqslant X_n$.

□

Définition 2.2.2. *Une famille (H_n) de variables aléatoires est dite prévisible si $\forall n \geqslant 1$ H_n est L^1 et $\mathcal{F}_{n-1}$-mesurable.*

Informellement, une famille est prévisible si au temps n on connaît déjà H_{n+1}.

Proposition 2.2.3. *Soit $(X_n)_{n\in\mathbb{N}}$ un processus adapté, et $(H_n)_{n\in\mathbb{N}}$ une famille prévisible et bornée. On pose $(H\cdot X)_0 := 0$ et pour tout $n \geqslant 1$*

$$(H\cdot X)_n := \sum_{i=1}^{n} H_i(X_i - X_{i-1}).$$

Alors

1. *Si $(X_n)_{n\in\mathbb{N}}$ est une martingale alors $((H\cdot X)_n)_{n\in\mathbb{N}}$ l'est aussi.*
2. *Si $(X_n)_{n\in\mathbb{N}}$ est une surmartingale et les H_n sont tous positifs, alors $((H\cdot X)_n)_{n\in\mathbb{N}}$ est aussi une surmartingale.*

Proof. Tout d'abord, comme les H_n sont bornés, les variables $(H\cdot X)_n$ sont bien intégrables, et le processus est adapté, par construction.

Supposons que $(X_n)_{n\in\mathbb{N}}$ est une martingale. Soit $n \in \mathbb{N}$. On a bien

$$\begin{aligned}\mathbb{E}[(H\cdot X)_{n+1} - (H\cdot X)_n|\mathcal{F}_n] &= \mathbb{E}[H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n] \\ &= H_{n+1}\mathbb{E}[X_{n+1} - X_n|\mathcal{F}_n] \\ &= 0.\end{aligned}$$

De même, si $(X_n)_{n\in\mathbb{N}}$ est une surmartingale et les H_n sont positifs,

$$\begin{aligned}\mathbb{E}[(H\cdot X)_{n+1} - (H\cdot X)_n|\mathcal{F}_n] &= \mathbb{E}[H_n(X_{n+1} - X_n)|\mathcal{F}_n] \\ &= H_n\mathbb{E}[X_{n+1} - X_n|\mathcal{F}_n] \\ &\leqslant 0.\end{aligned}$$

□

2.2.2 Décomposition de Doob-Meyer

Théorème 2.2.4. *Soit (X_n) un processus adapté et L^1. Il existe un unique processus prévisible (A_n) et une unique martingale (M_n) tels que $A_0 = M_0 = 0$ et pour tout $n \geqslant 1$ on ait $X_n = X_0 + A_n + M_n$. De plus, A_n est donnée par*

$$A_n = \sum_{k=0}^{n-1} \mathbb{E}[X_{k+1} - X_k|\mathcal{F}_k]. \tag{2.1}$$

Proof. Par définition, (A_n) est bien un processus prévisible, donc il nous suffit de montrer que le processus défini par $M_n = X_n - X_0 - A_n$ est bien une martingale. Tout d'abord, comme X_n et A_n sont intégrables, M_n l'est aussi. De plus,

$$\begin{aligned}\mathbb{E}(M_{n+1} - M_n|\mathcal{F}_n) &= \mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n) - \mathbb{E}(A_{n+1} - A_n|\mathcal{F}_n) \\ &= \mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n) - \mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n) = 0,\end{aligned}$$

et donc (M_n) est bien une martingale.

Montrons maintenant l'unicité. Supposons qu'il existe une martingale M' et un processus prévisible A' tels que $X = X_0 + M' + A'$. Alors $A'_{n+1} - A'_n = X_{n+1} - X_n - (M'_{n+1} - M'_n)$. Comme A' est prévisible, on a alors

$$A'_{n+1} - A'_n = \mathbb{E}[A'_{n+1} - A'_n|\mathcal{F}_n] = \mathbb{E}[X_{n+1} - X_n|\mathcal{F}_n] - \mathbb{E}[M'_{n+1} - M'_n|\mathcal{F}_n].$$

Comme M' est une martingale, le second terme est nul. En sommant sur n on voit que A' est bien donné par la formule (2.1), donc coincide avec le processus A, et nécessairement on a alors aussi $M' = M$. □

Exercice 2.2.1. *Quelle est la décomposition de Doob-Meyer d'une marche aléatoire simple asymmétrique, de paramètre p? Et pour un processus de Galton-Watson?*

Solution 2.2.1. *Soit (X_n) une telle marche aléatoire. On a $\mathbb{E}[X_{k+1} - X_k|\mathcal{F}_k] = 2p-1$, donc $A_n = (2p-1)n$ et $M_n = X_n - (2p-1)n$.*

Pour un processus de Galton-Watson, en reprenant les notations de la Section 1.2.2, on a

$$\mathbb{E}[N_{n+1} - N_n|\mathcal{F}_n] = N_n(\mathbb{E}[X_1^1] - 1)$$

et donc la partie prévisible est $A_n = (\mathbb{E}[X_1^1]-1)\left(\sum_{k=0}^{n-1} N_k\right)$, et la partie martingale est $N_n - (\mathbb{E}[X_1^1] - 1)\left(\sum_{k=0}^{n-1} N_k\right)$.

2.2.3 Inégalités maximales

Théorème 2.2.5 (Première inégalité de Doob)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une sous-martingale positive. Alors pour tout $n \in \mathbb{N}$ et $r > 0$ on a*

$$\mathbb{P}\left[\max_{0\leqslant k\leqslant n} X_k \geqslant r\right] \leqslant \frac{\mathbb{E}\left[X_n \mathbb{1}_{\max_{0\leqslant k\leqslant n} X_k \geqslant r}\right]}{r} \leqslant \frac{\mathbb{E}[X_n]}{r}.$$

Proof. Soit T le temps d'arrêt $\inf\{n; X_n \geqslant r\}$, de sorte que $\{\max_{0\leqslant k\leqslant n} X_k \geqslant r\} = \{T \leqslant n\}$. On a alors

$$r\mathbb{1}_{T=k} \leqslant X_k\mathbb{1}_{T=k} \leqslant \mathbb{E}[X_n|\mathcal{F}_k]\mathbb{1}_{T=k} = \mathbb{E}[X_n\mathbb{1}_{T=k}|\mathcal{F}_k],$$

et donc $r\mathbb{P}[T = k] \leqslant \mathbb{E}[X_n\mathbb{1}_{T=k}]$. On conclut en sommant sur $k \leqslant n$. □

Exemple 2.2.2. *Soit $M_n = \sum_{i=1}^n Y_i$ avec les Y_i des v.a. i.i.d. Gaussiennes centrées réduites. Alors $\mathbb{P}(\max_{k=1,\dots,n} M_k \geqslant t) \leqslant \exp(-t^2/(2n))$.*

En effet, come $x \to \exp(\lambda x)$ est une fonction croissante convexe pour tout $\lambda > 0$, $\exp(\lambda M_n)$ est une sous-martingale positive. Donc

$$\begin{aligned}\mathbb{P}\left(\max_{k=1,\dots,n} M_k \geqslant t\right) &= \mathbb{P}\left(\max_{k=1,\dots,n} \exp(\lambda M_k) \geqslant \exp(\lambda t)\right)\\ &\leqslant e^{-\lambda t}\mathbb{E}[\exp(\lambda M_n)]\\ &= e^{-\lambda t}\mathbb{E}[\exp(\lambda Y_1)]^n\\ &= \exp(-\lambda t + n\lambda^2/2).\end{aligned}$$

La conclusion suit en prenant $\lambda = t/n$ (qui est le choix optimal).

Théorème 2.2.6 (Deuxième inégalité de Doob)**.** *Soit $(M_n)_{n\in\mathbb{N}}$ une martingale. Pour tout $n \in \mathbb{N}$ et $p > 1$ on a*

$$\mathbb{E}\left[\sup_{0\leqslant k\leqslant n} |M_k|^p\right] \leqslant \left(\frac{p}{p-1}\right)^p \mathbb{E}[|M_n|^p].$$

En particulier,

$$\mathbb{E}\left[\sup_{n\geqslant 0} |M_k|^p\right] \leqslant \left(\frac{p}{p-1}\right)^p \sup_{n\geqslant 0} \mathbb{E}[|M_n|^p].$$

NB. *Les prefacteurs dans ces inégalités ne dépendent pas de n, ce qui sera utile lorsqu'on étudiera le comportement en temps long des martingales.*

Proof. La seconde inégalité est une conséquence immédiate de la première, par convergence monotone. Pour la première inégalité, posons $S_n := \max_{0\leqslant k\leqslant n} |M_k|$. Comme $|M_n|$ est une sous-martingale positive, par la première inégalité de Doob on a $r\mathbb{P}(S_n \geqslant r) \leqslant \mathbb{E}[|M_n|\mathbb{1}_{S_n\geqslant r}]$. En intégrant, on a donc

$$\begin{aligned}
&\int_0^\infty r\mathbb{P}(S_n \geqslant r) p r^{p-2} dr \leqslant \int_0^\infty \mathbb{E}[|M_n|\mathbb{1}_{S_n\geqslant r}] p r^{p-2} dr \\
\Leftrightarrow &\mathbb{E}\left[\int_0^\infty \mathbb{1}_{S_n\geqslant r} p r^{p-1} dr\right] \leqslant \mathbb{E}\left[\int_0^\infty |M_n|\mathbb{1}_{S_n\geqslant r} p r^{p-2} dr\right] \\
\Leftrightarrow &\mathbb{E}\left[\int_0^{S_n} \mathbb{1}_{S_n\geqslant r} p r^{p-1} dr\right] \leqslant \frac{p}{p-1}\mathbb{E}\left[\int_0^{S_n} |M_n|(p-1) r^{p-2} dr\right] \\
\Leftrightarrow &\mathbb{E}[S_n^p] \leqslant \left(\frac{p}{p-1}\right)\mathbb{E}[|M_n| S_n^{p-1}]
\end{aligned}$$

où on a utilisé le théorème de Fubini-Tonelli pour échanger l'ordre des intégrales. En utilisant l'inégalité de Hölder, on peut majorer le terme de droite dans la dernière inégalité :

$$\mathbb{E}[|M_n| S_n^{p-1}] \leqslant \mathbb{E}[|M_n|^p]^{1/p}\mathbb{E}[S_n^p]^{(p-1)/p},$$

ce qui nous permet de conclure qu'on a bien

$$\mathbb{E}[S_n^p] \leqslant \left(\frac{p}{p-1}\right)^p \mathbb{E}[|M_n|^p].$$

□

2.3 Théorème d'arrêt

Le but dans cet section va être d'étudier les variables de la forme X_T, lorsque $(X_n)_{n\in\mathbb{N}}$ est une martingale, et T un temps d'arrêt.

Théorème 2.3.1. *Soit $(X_n)_{n\in\mathbb{N}}$ une martingale, et T un temps d'arrêt. Alors $(X_{n\wedge T})_{n\in\mathbb{N}}$ est aussi une martingale.*

De même, si $(X_n)_{n\in\mathbb{N}}$ est une surmartingale, alors $(X_{n\wedge T})_{n\in\mathbb{N}}$ l'est aussi.

Une des applications de ce théorème est de permettre de calculer l'espérance de X_T lorsque T est borné p.s. :

Corollaire 2.3.2. *Si $(X_n)_{n\in\mathbb{N}}$ est une martingale et T est un temps d'arrêt borné p.s., alors $X_T \in L^1$, et $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.*

De même, si $(X_n)_{n\in\mathbb{N}}$ est une surmartingale et T est un temps d'arrêt borné p.s., alors $X_T \in L^1$, et $\mathbb{E}[X_T] \leqslant \mathbb{E}[X_0]$.

Un exemple d'interprétation est que si on joue à un jeu d'argent équilibré (i.e. modélisé par une martingale), il est impossible de trouver une stratégie qui gagne de l'argent en moyenne à un horizon de temps fini.

Proof. Soit $n \geqslant 1$. On pose $H_n := \mathbb{1}_{T \geqslant n} = 1 - \mathbb{1}_{T \leqslant n-1}$. Le processus $(H_n)_{n\in\mathbb{N}}$ est prévisible. Le processus

$$(H \cdot X) := \sum_{i=1}^{n} H_i(X_i - X_{i-1}) = X_{n\wedge T} - X_0$$

est alors une martingale, et donc $(X_{n\wedge T})_{n\in\mathbb{N}}$ l'est aussi. Si de plus $T \leqslant N$ p.s., alors

$$\mathbb{E}[X_T] = \mathbb{E}[X_{N\wedge T}] = \mathbb{E}[X_0].$$

La preuve pour les surmartingales fonctionne de la même manière. □

A noter que l'hypothèse T bornée n'est pas anodyne, et le corollaire peut être faux sans. Par exemple, si on considère une marche aléatoire simple sur $\mathbb{Z}$ issue de 0, et le temps d'arrêt $T := \inf\{n \geqslant 1; X_n = 1\}$. On a vu que $\limsup X_n = \infty$ p.s., et donc ce temps d'arrêt est fini p.s. Toutefois,

$$1 = X_T = \mathbb{E}[X_T] \neq \mathbb{E}[X_0] = 0.$$

On voit donc que on ne peut pas remplacer l'hypothèse de T borné par T fini p.s. Toutesfois, sous des conditions d'intégrabilité supplémentaires, c'est quand même possible :

Exercice 2.3.1. *Soit $(M_n)_{n\in\mathbb{N}}$ un martingale telle que il existe $L < \infty$ avec $|M_n| \leqslant L$ p.s., pour tout $n \in \mathbb{N}$. Montrer que si T est un temps d'arrêt fini p.s., alors $\mathbb{E}[M_T] = \mathbb{E}[M_0]$.*

Solution 2.3.1. *Pour tout $n \in \mathbb{N}$, d'après le théorème d'arrêt on a $\mathbb{E}[M_{n\wedge T}] = \mathbb{E}[M_0]$. On peut ensuite passer à la limite en n, en utilisant la borne L^∞ pour appliquer le théorème de convergence dominée.*

Nous verrons plus tard une version du théorème d'arrêt autorisant des temps d'arrêt non bornés, mais requérant une hypothèse supplémentaire sur la martingale à la place, qui sera toutefois plus faible que la borne L^∞ de l'énoncé ci-dessus.

On peut généraliser le théorème d'arrêt au résultat suivant :

Exercice 2.3.2. *Soient S et T deux temps d'arrêt bornés, tels que $S \leqslant T$, et $(X_n)_{n\in\mathbb{N}}$ une sous-martingale. Montrer que $\mathbb{E}[X_S] \leqslant \mathbb{E}[X_T]$.*

Le cas $S = 0$ correspond au théorème d'arrêt.

Solution 2.3.2. *S et T sont bornés, donc X_S et X_T sont bien L^1. Posons $H_n := \mathbb{1}_{S<n\leqslant T} = \mathbb{1}_{S\leqslant n-1} - \mathbb{1}_{T\leqslant n-1}$. Ca définit bien un processus prévisible et positif, donc $(H \cdot X)_n$ est une sous-martingale, et donc pour tout n on a $\mathbb{E}[(H \cdot X)_n] \geqslant 0$. Et si on prend $N \geqslant T$ fixé, on a $(H \cdot X)_N = X_T - X_S$, ce qui permet de conclure.*

2.4 Convergence de martingales

Dans ce chapitre, nous allons étudier le comportement en temps long des martingales. Le but est de montrer que sous des conditions très générales, si $(X_n)_{n\in\mathbb{N}}$ est une martingale alors la suite X_n converge vers une variable aléatoire limite. Comme il y a plusieurs notions de limite (limite p.s., limite L^p), nous aurons plusieurs types de résultats.

2.4.1 Convergence L^2

Théorème 2.4.1. *Soit $(X_n)_{n\in\mathbb{N}}$ une martingale bornée dans L^2. Alors il existe une variable aléatoire X_∞ telle que $X_n \longrightarrow X_\infty$ p.s. et dans L^2.*

Remark 2.4.1. *Ce théorème reste vrai si on remplace l'hypothèse de borne L^2 par une borne L^p avec $p > 1$. En revanche, il est faux si on suppose seulement une borne L^1. Nous verrons plus tard des contre-exemples.*

Proof. Sans perdre de généralité, on peut supposer $X_0 = 0$. Comme on l'a vu à l'exercice 2.1.1, on a pour tout $n, m \in \mathbb{N}$ l'identité

$$\mathbb{E}[(X_{n+m} - X_n)^2] = \sum \mathbb{E}[(X_{n+1} - X_n)^2].$$

En particulier, comme (X_n) est bornée dans L^2, cela implique que $\sum_n \mathbb{E}[(X_{n+1} - X_n)^2] < \infty$, et que (X_n) est une suite de Cauchy dans L^2. On en déduit alors que (X_n) converge dans L^2 vers une limite X_∞.

Montrons maintenant que (X_n) converge aussi p.s. Posons $R_n := \sup_{k,\ell\geqslant n} |X_k - X_\ell|$ et $R'_n := \sup_{k\geqslant n} |X_k - X_n|$. On a $R_n \leqslant 2R'_n$. (R_n) est une suite monotone, donc elle converge p.s. vers une limite R_∞. Si on montre que cette limite est nulle, on en déduirait que (X_n) est presque sûrement une suite de Cauchy, et donc convergerait p.s. Pour cela, nous allons montrer que $\lim_n \mathbb{E}[(R'_n)^2] = 0$. D'après la deuxième inégalité de Doob (Théorème 2.2.6) appliqué à la martingale $(X_k - X_n)_{k\geqslant n}$, on a

$$\mathbb{E}[(R'_n)^2] \leqslant 4 \sup_{k\geqslant n} \mathbb{E}[(X_k - X_n)^2] = \sum_{k\geqslant n} \mathbb{E}[(X_{k+} - X_k)^2].$$

La série $\sum_k \mathbb{E}[(X_{k+} - X_k)^2]$ étant finie, on en déduit que $\mathbb{E}[(R'_n)^2] \longrightarrow 0$, et donc que (X_n) converge p.s. vers une limite X'_∞.

Il nous reste à démontrer que la limite p.s. et la limite L^2 coïncident, mais cela est une conséquence directe du lemme de Fatou appliqué à $|X_n - X_\infty|$, qui converge p.s. vers $|X_\infty - X'_\infty|$, et dans L^2 vers 0. □

2.4.2 Convergence presque sûre

Nous allons maintenant étudier la convergence presque sûre des martingales. On a déjà vu que les martingales bornées dans L^2 convergenent p.s., le but ici va être de se passer de l'hypothèse de borne L^2.

Théorème 2.4.2. *Soit $(X_n)_{n\in\mathbb{N}}$ une surmartingale positive. Alors X_n converge p.s. vers une limite X_∞, à valeurs dans $[0, \infty]$.*

Proof. Comme $t \longrightarrow \exp(-t)$ est convexe et décroissante, le processus défini par $Y_n := \exp(-X_n)$ est une sous-martingale, d'après la Propositon 2.2.1. On considère sa décomposition de Doob-Meyer

$$Y_n = Y_0 + M_n + A_n$$

avec M une martingale et A un processus prévisible, tous deux issus de 0. La formule définissant A_n dans la décompositon de Doob-Meyer implique que, comme Y_n est une sous-martingale, A_n est croissante, et donc converge p.s. vers une limite A_∞, à valeurs dans $[0, +\infty]$. Pour prouver la converge p.s. de X_n, il nous suffit

donc de montrer la convergence p.s. de la martingale (M_n). Pour cela, nous allons montrer que cette martingale est bornée dans L^2, et appliquer le Théorème 2.4.1.

Tout d'abord, on remarque que comme Y_n est bornée, A_n et Y_n sont L^2 pour n fixé, et donc M_n est L^2. On peut donc écrire le développement

$$\mathbb{E}[(M_n)^2] = \sum_{k=0}^{n-1} \mathbb{E}[(M_{k+1} - M_k)^2].$$

On peut aussi développer

$$Y_n^2 = Y_0^2 + \sum_{k=0}^{n-1} Y_{k+1}^2 - Y_k^2.$$

Comme $Y_{k+1} - Y_k = A_{k+1} - A_k + M_{k+1} - M_k$, on a alors

$$Y_n^2 = Y_0^2 + \sum_{k=0}^{n-1} (M_{k+1}-M_k)^2 + (A_{k+1}-A_k)^2 + 2Y_k(M_{k+1}-M_k) + 2Y_k(A_{k+1}-A_k) + 2(M_{k+1}-M_k)(A_{k+1}-A_k).$$

Comme $Y_k \geqslant 0$ et $(A_{k+1} - A_k) \geqslant 0$, en négligeant des termes positifs on obtient

$$\sum_{k=0}^{n-1} (M_{k+1} - M_k)^2 + 2\sum_{k=0}^{n-1} (Y_k + A_{k+1} - A_k)(M_{k+1} - M_k) \leqslant Y_n^2 \leqslant 1.$$

De plus, comme A est un processus prévisible,

$$\mathbb{E}[(Y_k + A_{k+1} - A_k)(M_{k+1} - M_k)] = \mathbb{E}[(Y_k + A_{k+1} - A_k)\mathbb{E}[(M_{k+1} - M_k)|\mathcal{F}_k]] = 0.$$

Donc $\sum_{k=0}^{n-1} \mathbb{E}[(M_{k+1} - M_k)^2] \leqslant 1$, et donc M est une martingale bornée dans L^2, ce qui permet de conclure la preuve via le Théorème 2.4.1. □

Exemple 2.4.1. *Un processus de Galton-Watson avec $\mathbb{E}[X_1^1] \leqslant 1$ (on parle de processus sous-critique) converge p.s.*

Exercice 2.4.1. *Soit S_n une marche aléatoire simple, symétrique sur $\mathbb{Z}$, issue de $S_0 = 1$. Soit $T := \inf\{n; S_n = 0\}$.*

1. *Montrer que $S_{n\wedge T}$ converge p.s. vers une limite que l'on déterminera.*
2. *Montrer que $(S_{n\wedge T})_{n\in\mathbb{N}}$ est bornée dans L^1, mais qu'il n'y a pas convergence L^1.*

Solution 2.4.1. *1. $S_{n\wedge T}$ est une martingale positive, donc elle converge p.s. Comme T est fini p.s., sa limite est 0*

On peut aussi montrer cela sans savoir a priori que T est fini p.s. En effet, si $S_{n\wedge T} \neq 0$, alors $|S_{(n+1)\wedge T} - S_{n\wedge T}| = 1$. Donc la seule limite possible est 0, et a fortiori T est fini p.s.

2. *D'après le théorème d'arrêt, on a pour tout n $\mathbb{E}[S_{n\wedge T}] = \mathbb{E}[S_0] = 1 \neq 0$. Il n'y a donc pas convergence L^1. Toutefois, comme $S_{n\wedge T}$ est positive, $\sup_n \mathbb{E}[|S_{n\wedge T}|] = \sup_n \mathbb{E}[S_{n\wedge T}] = 1$.*

Théorème 2.4.3. *Soit $(X_n)_{n\in\mathbb{N}}$ une sous-martingale bornée dans L^1. Alors elle converge p.s. vers une limite X_∞, qui est elle-même L^1.*

NB. *En revanche, sous ces hypothèses la convergence L^1 n'est pas nécessairement vraie, comme le montre l'exercice précédent. On a seulement $\mathbb{E}[|X_\infty|] \leqslant \lim \mathbb{E}[|X_n|]$ via le lemme de Fatou.*

Proof. Le but va être de se ramener au cas du Théorème 2.4.2, en écrivant X comme la différence de deux surmartingales positives. Comme $x \longrightarrow \max(x, 0)$ est convexe, $X_n^+ := \max(X_n, 0)$ est une sous-martingale positive. Soit

$$X_n^+ = X_0^+ + M_n + A_n$$

sa décomposition de Doob. (A_n) est une suite croissante, donc elle converge p.s. vers une limite A_∞. De plus, $\mathbb{E}(A_n) = \mathbb{E}(X_n^+) - \mathbb{E}(X_0^+)$ est uniformément bornée par $\sup \mathbb{E}(|X_n|)$, donc par convergence monotone A_∞ est L^1. Le processus

$$Y_n := X_0^+ + M_n + \mathbb{E}[A_\infty|\mathcal{F}_n]$$

est alors une martingale, et est positive car $A_\infty \geqslant A_n$ et donc

$$Y_n \geqslant X_0^+ + M_n + A_n = X_n^+ \geqslant 0.$$

Donc Y_n converge p.s., et on peut vérifier que sa limite est L^1.

Le processus $Z_n := Y_n - X_n$ est alors une surmartingale, car c'est la différence entre une martingale et une sous-martingale. De plus

$$Z_n \geqslant X_n^+ - X_n \geqslant 0.$$

C'est donc une surmartingale positive, elle converge p.s., et on peut aussi vérifier que sa limite est L^1. On en déduit que X_n converge p.s., comme différence de deux suites convergentes.

□

Exercice 2.4.2. *Soit $(X_n)_{n\in\mathbb{N}}$ le processus issu de $X_0 = 1$, tel que $X_n \in \{0, 2^n\}$ pour tout n, et défini par les opérations suivantes : si $X_n = 0$, alors $X_{n+1} = 0$. Sinon, X_{n+1} suit une loi uniforme sur $\{0, 2^n\}$.*

1. *Montrer que (X_n) est une martingale.*
2. *Montrer qu'elle est bornée dans L^1.*
3. *Montrer qu'elle converge p.s. vers une limite qu'on déterminera. A-t-on convergence L^1 ?*

Solution 2.4.2.

1. *On calcule directement $\mathbb{E}[X_{n+1}|X_n = 0] = 0$ et $\mathbb{E}[X_{n+1}|X_n = 2^n] = 2^n$.*
2. *Comme les X_n sont positifs et en utilisant la propriété de martingale*

$$E[|X_n|] = \mathbb{E}[X_n] = \mathbb{E}[X_0] = 1 \quad \forall n.$$

3. *En appliquant le théorème 2.4.3, on a la convergence p.s. vers une limite X_∞. Comme de plus $\mathbb{P}(\liminf_k X_k > 0) \leqslant \mathbb{P}(X_n > 0) = 2^{-n} \longrightarrow 0$, la limite est nulle. On voit que l'espérance de la limite est différente de la limite des espérances, donc il n'y a pas convergence L^1.*

2.4.3 Martingales uniformément intégrables

Le but de cette section va être de mieux comprendre quand est-ce qu'on a convergence L^1 des martingales qui sont seulement bornées dans L^1.

Définition 2.4.4. *Une famille $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires est dite uniformément intégrable si*

$$\lim_{R\longrightarrow\infty}\sup_n \mathbb{E}[|X_n|\mathbb{1}_{|X_n|\geqslant R}] = 0.$$

Exemple 2.4.2. *Une famille bornée dans L^p pour $p > 1$ est uniformément intégrable. Et si une famille est uniformément intégrable, alors elle est bornée dans L^1.*

Pour la première assertion, si $\sup_n \mathbb{E}[|X_n|^p] \leqslant C$, alors en prenant q tel que $p^{-1} + q^{-1} = 1$, on

$$\begin{aligned}\sup_n \mathbb{E}[|X_n|\mathbb{1}_{|X_n|\geqslant R}] &\leqslant \sup_n \mathbb{E}[|X_n|^p]^{1/p}\mathbb{P}[|X_n| \geqslant R]^{1/q}\\ &\leqslant \sup_n C^{1/p}\mathbb{E}[|X_n|^p]^{1/q}R^{-p/q}\\ &\leqslant CR^{-p/q}\end{aligned}$$

et donc on a bien $\lim_{R\longrightarrow\infty}\sup_n \mathbb{E}[|X_n|\mathbb{1}_{|X_n|\geqslant R}] = 0$.

Pour la deuxième assertion, on a

$$\sup_n \mathbb{E}[|X_n|] \leqslant R + \sup_n \mathbb{E}[|X_n|\mathbb{1}_{|X_n|\geqslant R}] < \infty.$$

Exercice 2.4.3. *Soit $(X_n)_{n\in\mathbb{N}}$ une famille de variables aléatoires. On suppose qu'il existe une variable aléatoire Z L^1, et telle que pour tout n on a $|X_n| \leqslant Z$. Montrer que $(X_n)_{n\in\mathbb{N}}$ est uniformément intégrable.*

Solution 2.4.3. *On a $\sup_n \mathbb{E}[|X_n|\mathbb{1}_{|X_n|\geqslant R}] \leqslant \mathbb{E}[|Z|\mathbb{1}_{|Z|\geqslant R}]$, et comme Z est L^1, on a $\lim_{R\to\infty}\mathbb{E}[|Z|\mathbb{1}_{|Z|\geqslant R}] = 0$.*

Une manière de voir cet énoncé est que le théorème de convergence dominé est un critère d'uniforme intégrabilité.

On peut un peu étendre la notion d'uniforme intégrabilité de la manière suivante :

Proposition 2.4.5. *Soit $(X_n)_{n\in\mathbb{N}}$ une famile bornée dans L^1. Les propriétés suivantes sont équivalentes :*

1. *La famille $(X_n)_{n\in\mathbb{N}}$ est uniformément intégrable;*
2. *Pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour tout évènement $A \in \mathcal{F}$ avec $P(A) \leqslant \delta$, on ait*
$$\sup_n \mathbb{E}[|X_n|\mathbb{1}_A] \leqslant \varepsilon.$$

Proof. $(ii) \Rightarrow (i)$ est le sens facile, car d'après l'inégalité de Markov, en prenant $C = \sup_n \mathbb{E}[|X_n|]$, on a

$$\sup_n \mathbb{P}(|X_n| \geqslant R) \leqslant \frac{C}{R}.$$

Donc pour tout $\varepsilon > 0$ et δ associé, si on prend R tel que $C/R < \delta$, on a bien

$$\sup_n \mathbb{E}[|X_n| \mathbb{1}_{|X_n| \geqslant R}] \leqslant \varepsilon.$$

Prouvons maintenant que $(i) \Rightarrow (ii)$. Soit $\varepsilon > 0$. On peut choisir a suffisemment grand tel que

$$\sup_n \mathbb{E}[|X_n| \mathbb{1}_{|X_n| > a}] < \varepsilon/2.$$

Alors en prenant $\delta = \varepsilon/(2a)$, si $P(A) \leqslant \delta$, on a pour tout $n \in \mathbb{N}$

$$\mathbb{E}[|X_n| \mathbb{1}_A] = \mathbb{E}[|X_n| \mathbb{1}_{A \cap \{|X_n| \leqslant a\}}] + \mathbb{E}[|X_n| \mathbb{1}_{A \cap \{|X_n| > a\}}] \leqslant aP(A) + \mathbb{E}[|X_n| \mathbb{1}_{|X_n| > a}] < \varepsilon,$$

ce qui conclut la preuve. □

Théorème 2.4.6. *Soit $(X_n)_{n \in \mathbb{N}}$ une martingale. Les propriétés suivantes sont équivalentes :*

1. *$(X_n)_{n \in \mathbb{N}}$ est uniformément intégrable;*
2. *$(X_n)_{n \in \mathbb{N}}$ est une martingale fermée;*
3. *$(X_n)_{n \in \mathbb{N}}$ converge p.s. et dans L^1.*

Proof. Commençons par montrer que (1) ⇒ (3). Si la martingale est UI, on a déjà vu qu'elle est bornée dans L^1, et donc elle converge p.s. vers une limite X_∞ qui est L^1. Reste à montrer la convergence L^1. Nous allons montrer que la suite est de Cauchy dans L^1. En utilisant la Propositon 2.4.5, on peut voir que la famile dénombrable $(X_n - X_m)_{n,m \in \mathbb{N}}$ est aussi uniformément intégrable. Donc pour $\varepsilon > 0$, il existe a tel que

$$\sup_{n,m} \mathbb{E}[|X_n - X_m| \mathbb{1}_{|X_n - X_m| \geqslant a}] < \varepsilon.$$

On a ensuite

$$\begin{aligned}
\mathbb{E}[|X_n - X_m|] &\leqslant \mathbb{E}[|X_n - X_m| \mathbb{1}_{|X_n - X_m| \geqslant a}] \\
&\quad + \mathbb{E}[|X_n - X_m| \mathbb{1}_{\varepsilon < |X_n - X_m| < a}] + \mathbb{E}[|X_n - X_m| \mathbb{1}_{|X_n - X_m| \leqslant \varepsilon}] \\
&\leqslant 2\varepsilon + a\mathbb{P}(|X_n - X_m| > \varepsilon)
\end{aligned}$$

Comme $(X_n)_{n \in \mathbb{N}}$ converge p.s. (et donc en probabilité) vers X_∞, on a

$$\mathbb{P}(|X_n - X_m| > \varepsilon) \leqslant \mathbb{P}(|X_n - X_\infty| > \varepsilon/2) + \mathbb{P}(|X_m - X_\infty > \varepsilon/2) \underset{n,m \to \infty}{\longrightarrow} 0.$$

On en déduit que $\limsup_{n,m \to \infty} \mathbb{E}[|X_n - X_m|] \leqslant 2\varepsilon$, pour tout $\varepsilon > 0$. La suite $(X_n)_{n \in \mathbb{N}}$ est donc bien une suite de Cauchy dans L^1. Sa limite L^1 est toujours X_∞, par application du lemme de Fatou.

Montrons maintenant que (3) ⇒ (2). Par la propriété de martingale, pour tout $m \geqslant n$, on a $X_n = \mathbb{E}[X_m | \mathcal{F}_n]$. Comme on a toujours $\mathbb{E}[|\mathbb{E}[Y|\mathcal{F}_n|] \leqslant \mathbb{E}[|Y|]$, l'application $Y \longrightarrow \mathbb{E}[Y|\mathcal{F}_n]$ est continue pour la topologie L^1, et donc on peut faire tendre m vers l'infini, et on obtient que pour tout n $X_n = \mathbb{E}[X_\infty | \mathcal{F}_n]$, où X_∞ est la limite p.s. et L^1. Donc la martingale est fermée.

Enfin, montrons que (2) ⇒ (1). Soit Z une variable aléatoire L^1 telle que pour tout n on ait $X_n = \mathbb{E}[Z|\mathcal{F}_n]$. Comme le singleton Z ets uniformément intégrable,

d'après la Propositon 2.4.5, pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que si $\mathbb{P}(A) < \delta$, on a $\mathbb{E}[|Z|\mathbb{1}_A] < \varepsilon$. Comme pour tout $a > 0$

$$\mathbb{P}(|\mathbb{E}[Z|\mathcal{F}_n]| > a) \leqslant \frac{\mathbb{E}[|\mathbb{E}[Z|\mathcal{F}_n]|]}{a} \leqslant \frac{\mathbb{E}[|Z|]}{a}$$

si on prend a suffisemment grand pour que $\mathbb{E}[|Z|]/a < \delta$, on a bien

$$\mathbb{E}[|\mathbb{E}[Z|\mathcal{F}_n]|\mathbb{1}_{|\mathbb{E}[Z|\mathcal{F}_n]|>a}] \leqslant \mathbb{E}[|X|\mathbb{1}_{|\mathbb{E}[Z|\mathcal{F}_n]|>a}] < \varepsilon,$$

et la famille $\mathbb{E}[Z|\mathcal{F}_n]$ est donc bien uniformément intégrable. □

On notera que dans cette preuve, la partie suivante n'a pas utilisé la structure de martingale, et est donc valable sans cette hypothèse :

Théorème 2.4.7. *Soit $(X_n)_{n\in\mathbb{N}}$ une famille de variable uniformément intégrable, et qui converge p.s. Alors la convergence est aussi L^1.*

On conclut par un théorème d'arrêt pour les martingales uniformément intégrables :

Corollaire 2.4.8 (Théorème d'arrêt pour les martingales UI)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une martingale uniformément intégrable, et T un temps d'arrêt fini p.s. Alors $\mathbb{E}[X_T] = \mathbb{E}[X_0]$.*

Proof. Montrons d'abord que X_T est L^1. On a

$$\begin{aligned}
\mathbb{E}[|X_T|] &= \sum_{n\in\mathbb{N}\cup\{\infty\}} \mathbb{E}[|X_n|\mathbb{1}_{T=n}] \\
&= \sum_{n\in\mathbb{N}\cup\{\infty\}} \mathbb{E}[|\mathbb{E}[X_\infty|\mathcal{F}_n]|\mathbb{1}_{T=n}] \\
&\leqslant \sum_{n\in\mathbb{N}\cup\{\infty\}} \mathbb{E}[\mathbb{E}[|X_\infty||\mathcal{F}_n]\mathbb{1}_{T=n}] \\
&= \sum_{n\in\mathbb{N}\cup\{\infty\}} \mathbb{E}[|X_\infty|\mathbb{1}_{T=n}] \\
&= \mathbb{E}[|X_\infty|].
\end{aligned}$$

Comme X_T est L^1, en appliquant le théorème de Fubini, si $A \in \mathcal{F}_T$

$$\mathbb{E}[X_T\mathbb{1}_A] = \sum_{n\in\mathbb{N}\cup\{\infty\}} \mathbb{E}[X_T\mathbb{1}_{A\cap\{T=n\}}] = \sum_{n\in\mathbb{N}\cup\{\infty\}} \mathbb{E}[X_n\mathbb{1}_{A\cap\{T=n\}}]$$

$$= \sum_{n\in\mathbb{N}\cup\{\infty\}} \mathbb{E}[\mathbb{E}[X_\infty|\mathcal{F}_n]\mathbb{1}_{A\cap\{T=n\}}] = \sum_{n\in\mathbb{N}\cup\{\infty\}} \mathbb{E}[X_\infty\mathbb{1}_{A\cap\{T=n\}}] = \mathbb{E}[X_\infty\mathbb{1}_A].$$

Donc

$$X_T = \mathbb{E}[X_\infty|\mathcal{F}_T]$$

et la conclusion suit. □

Exercice 2.4.4 (Ruine du joueur)**.** *Soit S_n une marche aléatoire simple symétrique issue de $k > 0$, et soit $\ell > k$. Alors la probabilité que S_n atteigne le site ℓ avant de visiter le site 0 est égale à k/ℓ.*

Solution 2.4.4. *On considère les temps d'arrêt* $T_p = \inf\{n; S_n = p\}$ *et* $T = T_0 \wedge T_\ell$. *Comme* $S_{n\wedge T}$ *est une martingale bornée, elle est uniformément intégrable. De plus, le temps d'arrêt* T *est fini p.s., et donc* $S_{n\wedge T}$ *converge vers* S_T. *On peut donc appliquer le Théorème 2.4.8 pour obtenir*

$$\mathbb{E}[S_T] = \mathbb{E}[S_0] = k.$$

Or

$$\mathbb{E}[S_T] = \ell\mathbb{P}[T_\ell < T_0] + 0 \times \mathbb{P}[T_0 < T_\ell] = \ell\mathbb{P}[T_\ell < T_0].$$

On en déduit qu'on a $\mathbb{P}[T_\ell < T_0] = k/\ell$, *ce qui est le résultat voulu.*

2.4.4 Une application en analyse : le théorème de Rademacher

Théorème 2.4.9. *Soit* $f : [0,1] \longrightarrow$ *une fonction lipchitzienne. Alors il existe une fonction* g *mesurable bornée telle que pour tout* $x \in \mathbb{R}$ *on ait*

$$f(x) = f(0) + \int_0^x g(t)dt.$$

En particulier, f *est dérivable presque partout.*

Proof. Soit $L > 0$ telle que f soit L-lipschitz. On considère X une v.a. uniforme sur $[0,1]$, et les variables

$$X_n = 2^{-n}\lfloor 2^n X \rfloor; \quad Z_n := 2^n(f(X_n + 2^{-n}) - f(X_n)).$$

On note $(\mathcal{F}_n)$ la filtration canonique associée aux X_n.

Alors (Z_n) est une $\mathcal{F}_n$-martingale bornée par L. En effet, conditionnellement à la valeur de X_n, on a deux valeurs équiprobables pour X_{n+1}, qui sont X_n et $X_n + 2^{-n-1}$. Alors

$$\begin{aligned}\mathbb{E}[Z_{n+1}|\mathcal{F}_n] &= 2^{n+1}\left(\frac{f(X_n + 2^{-n-1}) - f(X_n)}{2} + \frac{f(X_n + 2^{-n-1} + 2^{-n-1}) - f(X_n + 2^{-n-1})}{2}\right)\\ &= 2^n(f(X_n + 2^{-n}) - f(X_n)) = Z_n.\end{aligned}$$

Elle converge donc p.s. et dans L^1 vers une v.a. Z, qui est $\sigma(X)$-mesurable, car $\sigma(X) = \cap_n \sigma(X_n, X_{n+1}...)$. Il existe donc une fonction mesurable g telle que $Z = g(X)$. Comme de plus Z est bornée par L, on peut prendre g bornée par L. En particulier,

$$Z_n = \mathbb{E}[g(X)|X_n].$$

Or, conditionnellement à X_n, X est uniformément distribuée sur $[X_n, X_n + 2^{-n-1}[$, donc p.s.

$$Z_n = 2^n \int_{X_n}^{X_n + 2^{-n}} g(u)du.$$

On en déduit que p.s.

$$f(X_n + 2^{-n}) - f(X_n) = \int_{X_n}^{X_n + 2^{-n}} g(u)du.$$

Donc pour tout $k < 2^n$ on a

$$f((k+1)2^{-n}) - f(k2^{-n}) = \int_{k2^{-n}}^{(k+1)2^{-n}} g(u)du.$$

En sommant, on obtient pour tout $k < 2^n$

$$f(k2^{-n}) = f(0) + \int_0^{k2^{-n}} g(u)du.$$

En faisant tendre $k2^{-n}$ vers x, par continuité de f on en déduit qu'on a bien

$$f(x) = f(0) + \int_0^x g(u)du.$$

□

Chapter 3

Chaînes de Markov

Dans ce chapitre, $\mathcal{X}$ sera toujours un ensemble au plus dénombrable.

3.1 Définitions et premières propriétés

Définition 3.1.1.

$$Q : \begin{cases} \mathcal{X} \times \mathcal{X} & \longrightarrow [0,1] \\ (x,y) & \longrightarrow Q(x,y) \end{cases}$$

est une matrice stochastique sur $\mathcal{X}$ si pour tout $x \in \mathcal{X}$ on a $\sum_y Q(x,y) = 1$.

De manière équivalente, une matrice stochastique est une collection $(Q(x,\cdot))_{x\in\mathcal{X}}$ de mesures de probabilité sur $\mathcal{X}$.

Notation. *On peut définir la matrice stochastique $Q_n := Q^n$ par récurrence avec avec $Q_1 = Q$ et $Q_{n+1}(x,y) = \sum_z Q_n(x,z)Q(z,y)$.*

Pour tout $f : \mathcal{X} \longrightarrow \mathbb{R}$, on note $Qf(x) := \sum_y Q(x,y)f(y)$.

Si on identifie f à un vecteur colonne de $\mathbb{R}^{|\mathcal{X}|}$, alors Qf est obtenue en appliquant la matrice Q à f.

Définition 3.1.2. *Soient Q une matrice stochastique sur $\mathcal{X}$, et $(X_n)_{n\in\mathbb{N}}$ un processus aléatoire à valeurs dans $\mathcal{X}$. On dit que $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov de matrice de transition Q si pour tout $n \geqslant 0$ la loi conditionnelle de X_{n+1} connaissant $(X_0, ..., X_n)$ est $Q(X_n, \cdot)$.*

De manière équivalente,

$$\mathbb{P}(X_{n+1} = y | X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = Q(x_n, y)$$

pour tout $x_0, ..., x_n$ tels que $\mathbb{P}(X_0 = x_0, ..., X_n = x_n) > 0$.

Interprétation : pour un processus adapté général, la loi de X_{n+1} conditionnellement à $(X_0, ..., X_n)$ dépend a priori de tous les X_i, i.e. de tout le passé. Pour une chaîne de Markov, elle ne dépend que du dernier élément X_n. C'est un processus avec une absence de mémoire, le futur ne dépend pas du passé, conditionnellement au présent.

NB. *Ce n'est pas la même chose que de ne pas du tout dépendre du passé. Il y a des corrélations entre passé et futur, mais leur information prédictive est entièrement contenue dans celle du présent.*

On peut voir les chaînes de Markov comme l'analogue stochastique des suites récurrentes $x_{n+1} = f(x_n)$. En particulier, elles (et leur analogue en temps continu) jouent le rôle pour la physique statistique que les équations différentielles ordinaires jouent pour la mécanique classique.

Exemple 3.1.1. *1. La marche aléatoire simple sur $\mathbb{Z}$ de paramètre p est une chaîne de Markov, de matrice de transition*

$$Q(i, i+1) = p; \quad Q(i, i-1) = 1-p; \quad Q(k, \ell) = 0 \text{ si } |k-\ell| \geqslant 2.$$

2. Une marche aléatoire simple sur un graphe (S, A) est une chaîne de Markov de matrice de transition

$$Q(x,y) = \begin{cases} \frac{1}{d_x} & \text{si } (x,y) \in A; \\ & \text{sinon;} \end{cases}$$

où $d_x = \mathrm{Card}(\{y; (x,y) \in A\})$.

Exercice 3.1.1. *Montrer qu'un processus de Galton-Watson est une chaîne de Markov sur $\mathbb{N}$, et calculer sa matrice de transition dans le cas où les X_i^k sont des variables de Bernoulli de paramètre p.*

Solution 3.1.1. *On a*

$$\begin{aligned} \mathbb{P}(N_{n+1} = k_{n+1} | N_0 = k_0, ..., N_n = k_n) &= \mathbb{P}\left(\sum_{i=1}^{N_n} X_i^n = k_{n+1} \mid N_0 = k_0, ..., N_n = k_n \right) \\ &= \mathbb{P}\left(\sum_{i=1}^{k_n} X_i^n = k_{n+1} \right). \end{aligned}$$

C'est donc bien une chaîne de Markov. Dans le cas où les X_i suivent une loi de Bernoulli de paramètre p, conditonnellement à $\{N_n = k_n\}$, N_{n+1} suit une loi binomiale de paramètres (k_n, p). Donc

$$Q(k, \ell) = \binom{k}{\ell} p^\ell (1-p)^{k-\ell}; \quad 0 \leqslant \ell \leqslant k.$$

Proposition 3.1.3. *Un processus $(X_n)_{n\in\mathbb{N}}$ à valeurs dans $\mathcal{X}$ est une chaîne de Markov de matrice de transition Q ssi pour tout $n \in \mathbb{N}$ et $x_0, ..., x_n \in \mathcal{X}$ on a*

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = \mathbb{P}(X_0 = x_0) \prod_{i=1}^{n} Q(x_{i-1}, x_i). \tag{3.1}$$

En particulier, si $\mathbb{P}(X_0 = x) > 0$ on a

$$\mathbb{P}(X_n = y | X_0 = x) = Q_n(x, y). \tag{3.2}$$

Proof. Si $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov, on peut démontrer (3.1) par récurrence sur n, en utilisant la formule de conditionnement

$$\mathbb{P}(X_0 = x_0, ..., X_{n+1} = x_{n+1}) = \mathbb{P}(X_0 = x_0, ..., X_n = x_n) \mathbb{P}(X_{n+1} = x_{n+1} | X_0 = x_0, ..., X_n = x_n).$$

Inversement, si (3.1) est vraie, alors

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_0 = x_0, ..., X_n = x_n) = \frac{\mathbb{P}(X_0 = x_0, ..., X_{n+1} = x_{n+1})}{\mathbb{P}(X_0 = x_0, ..., X_n = x_n)} = Q(x_n, x_{n+1}).$$

L'identité (3.2) s'en déduit via le développement

$$Q_n(x_0, x_n) = \sum_{x_1, ..., x_{n-1} \in \mathcal{X}} \prod_{i=0}^{n-1} Q(x_i, x_{i+1}).$$

□

La loi de $(X_0, ..., X_n)$ lorsque les X_i forment une chaîne de Markov est donc complètement déterminée par la loi de la condition initiale X_0 et la matrice de transition Q.

Notation. *On utilisera les notations $\mathbb{E}_\mu$ et $\mathbb{P}_\mu$ pour l'espérance et la loi d'une chaîne de Markov lorsque la condition initiale est donnée par la mesure μ. En particulier, $\mathbb{P}_x$ sera la loi de la chaîne lorsque la loi de la condition initiale est δ_x.*

Proposition 3.1.4. *Soit $f : \mathcal{X} \longrightarrow \mathbb{R}$ une fonction mesurable, positive ou bornée. Alors*

$$\mathbb{E}[f(X_{n+1}) | X_0, ..., X_n] = \mathbb{E}[f(X_{n+1}) | X_n] = Qf(X_n). \tag{3.3}$$

Plus généralement,

$$\mathbb{E}[f(X_{n+p}) | X_0, ..., X_n] = Q_p f(X_n). \tag{3.4}$$

Proof. Commençons par démontrer (3.3). On a

$$\begin{aligned}
\mathbb{E}[f(X_{n+1}) | X_0, ..., X_n] &= \sum_{y \in \mathcal{X}} f(y) \mathbb{P}(X_{n+1} = y | X_0, ..., X_n) \\
&= \sum_{y \in \mathcal{X}} f(y) Q(X_n, y) \\
&= Qf(X_n).
\end{aligned}$$

Pour démontrer (3.4), on procède par récurrence sur p. Nous venons de le démontrer pour $p = 1$. Supposons que c'est vrai pour un $p \geqslant 1$ fixé. On a alors

$$\begin{aligned}
\mathbb{E}[f(X_{n+p+1}) | X_0, ..., X_n] &= \mathbb{E}[\mathbb{E}[f(X_{n+p+1}) | X_0, ..., X_n, X_{n+1}] | X_0, ..., X_n] \\
&= \mathbb{E}[Q_p f(X_{n+1}) | X_0, ..., X_n] \\
&= Q_{p+1} f(X_n).
\end{aligned}$$

□

On conclut ce chapitre avec un résultat général d'existences des lois de chaînes de Markov. Il ne sera pas explicitement utilisé par la suite, mais garantit la non-vacuité des énoncés qu'on fera dans ce cours. On n'en donnera pas la preuve ici.

Théorème 3.1.5. *Soit Q une matrice stochastique sur $\mathcal{X}$.*

1. *Il existe un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ sur lequel pour tout $x \in \mathcal{X}$ il existe un processus X_n^x qui est une chaîne de Markov de matrice de transition Q, issue de $X_0 = x$.*

2. *Pour tout $x \in \mathcal{X}$, il existe une unique mesure de probabilité $\mathbb{P}_x$ sur $\Omega = \mathcal{X}^{\mathbb{N}}$ telle que sous $\mathbb{P}_x$ le processus des coordonnées $(X_n)_{n \in \mathbb{N}}$ est une chaîne de Markov de matrice de transition Q, et avec $\mathbb{P}_x(X_0 = x) = 1$.*

3.2 Propriété de Markov

L'absence de mémoire des chaînes de Markov fait que, conditionnellement à la valeur de X_n, la loi du futur $(X_k)_{k\geqslant n}$ est la même que celle d'une chaîne de Markov de condition initiale X_n. Ceci est formalisé par le résultat suivant :

Théorème 3.2.1 (Propriété de Markov simple). *Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov de matrice de transition Q, et $(\mathcal{F}_n)_{n\in\mathbb{N}}$ sa filtration canonique. Alors pour tout $n \in \mathbb{N}$ et $x \in \mathcal{X}$, conditionnellement à $\{X_n = x\}$ le processus $(X_{n+p})_{p\in\mathbb{N}}$ est une chaîne de Markov de matrice de transition Q, de loi initial δ_x et indépendante de $(X_0, ..., X_n)$, i.e. pour tout $A \in \mathcal{F}_n$ on a*

$$\mathbb{P}(A\cap\{X_{n+1} = x_1, ..., X_{n+p} = x_p\}|X_n = x) = \mathbb{P}(A|X_n = x)\mathbb{P}_x(X_1 = x_1, ..., X_p = x_p).$$

On peut aussi dire que la loi conditionnelle de $(X_k)_{k\geqslant n}$ connaissant $(X_0, ..., X_n)$ est $\mathbb{P}_{X_n}$.

Proof. Il nous suffit de montrer le résultat pour A de la forme $\{X_0 = y_0, ..., X_n = y_n\}$, puisque cette famille engendre la tribu $\mathcal{F}_n$ (car $\mathcal{X}$ est au plus dénombrable). De plus, si $y_n \neq x$, les deux probabilités sont nulles, donc égales, et il suffit donc de considérer le cas $y_n = x$. On a alors

$$\begin{aligned}
&\mathbb{P}(A \cap \{X_{n+1} = x_1, ..., X_{n+p} = x_p\}|X_n = x)\\
&= \frac{\mathbb{P}(X_0 = y_0, ..., X_{n-1} = y_{n-1}, X_n = x, X_{n+1} = x_1, ..., X_{n+p} = x_p)}{\mathbb{P}(X_n = x)}\\
&= \frac{\mathbb{P}(A)}{\mathbb{P}(X_n = x)}Q(x, x_1)Q(x_1, x_2)...Q(x_{p-1}, x_p)\\
&= \mathbb{P}(A|X_n = x)\mathbb{P}_x(X_1 = x_1, ..., X_p = x_p).
\end{aligned}$$

□

On peut généraliser ce résultat, en remplaçant le temps n par un temps d'arrêt aléatoire :

Théorème 3.2.2 (Propriété de Markov forte). *Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov de matrice de transition Q, $(\mathcal{F}_n)_{n\in\mathbb{N}}$ sa filtration canonique et T un temps d'arrêt adapté. Alors pour tout $x \in \mathcal{X}$, conditionnellement à $\{X_T = x\} \cap \{T < \infty\}$ le processus $(X_{T+p})_{p\in\mathbb{N}}$ est une chaîne de Markov de matrice de transition Q, de loi initial δ_x et indépendante de $(X_0, ..., X_T)$, i.e. pour tout $A \in \mathcal{F}_T$ on a*

$$\mathbb{P}(A\cap\{X_{T+1} = x_1, ..., X_{T+p} = x_p\}|X_T = x, T < \infty) = \mathbb{P}(A|X_T = x, T < \infty)\mathbb{P}_x(X_1 = x_1, ..., X_p = x_p).$$

Proof. Soit $n \in \mathbb{N}$. En appliquant la propriété de Markov simple, on a

$$\begin{aligned}
&\mathbb{P}(A \cap \{T = n\} \cap \{X_{T+1} = x_1, ..., X_{T+p} = x_p\}|X_T = x, T < \infty)\\
&= \mathbb{P}(A \cap \{T = n\}|X_T = x, T < \infty)\mathbb{P}_x(X_1 = x_1, ..., X_p = x_p)
\end{aligned}$$

et on conclut en sommant sur $n \in \mathbb{N}$. □

Comme première application, on peut utiliser cette propriété de Markov forte pour étudier le nombre de retours en un point :

Proposition 3.2.3. *Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov. On pose*

$$N_x := \sum_{n=0}^{\infty} \mathbb{1}_{X_n=x}$$

le nombre de visites en x, et

$$H_x := \inf\{n \geqslant 1; X_n = x\}$$

le temps de premier passage en x après l'instant initial. Alors

$$\mathbb{P}_x[N_x = \infty] = 1 \Leftrightarrow \mathbb{P}_x[H_x < \infty] = 1,$$

i.e. si en partant de x on y revient p.s. une fois, on y revient une infinité de fois.

Proof. L'implication $\Rightarrow$ est triviale, pusique bien sur si un état est visité une infinité de fois, il est a fortiori visité au moins une fois après l'instant initial.

Pour l'implication $\Leftarrow$, comme $\mathbb{P}_x(H_x < \infty) = 1$,

$$N_x(X_0, X_1...) \overset{loi}{=} 1 + N_x(X_{H_x}, X_{H_x+1}, ...).$$

Mais d'après la propriété de Markov forte, $(X_{H_x}, X_{H_x+1}, ...)$ a la même loi que $(X_0, X_1, ...)$, i.e. $\mathbb{P}_x$. Donc

$$N_x \overset{loi}{=} 1 + N_x$$

et donc $N_x = +\infty$ p.s. □

Exemple 3.2.1. *La marche aléatoire simple sur $\mathbb{Z}$ issue de 0 revient une infinité de fois en 0.*

Exercice 3.2.1. *Montrer que ce n'est pas le cas pour les marches aléatoires asymétriques sur $\mathbb{Z}$.*

3.3 Mesures invariantes et classification des états

3.3.1 Classification des états

Définition 3.3.1 (Etats récurrents, états transitoires)**.** *Soit $x \in \mathcal{X}$. En reprenant les notations de la Proposition 3.2.3*

- *Si $\mathbb{P}_x(H_x < \infty) = 1$, on dit que x est récurrent.*
- *Si $\mathbb{P}_x(H_x < \infty) < 1$, on dit que x est transitoire.*

Cette terminologie est justifiée par la Proposition 3.2.3, puisque si la chaîne part d'un point récurrent, elle y revient une infinité de fois.

Proposition 3.3.2. *Si x est transitoire, alors*

$$\mathbb{E}_x(N_x) = \frac{1}{\mathbb{P}_x(H_x = \infty)}.$$

Proof. En utilisant la propriété de Markov forte, on a

$$\begin{aligned}\mathbb{P}_x(N_x \geqslant k+1) &= \mathbb{E}_x\left(\mathbb{1}_{H_x<\infty}((X_k)_{k\geqslant 0})\mathbb{1}_{N_x\geqslant k}((X_k)_{k\geqslant H_x})\right)\\ &= \mathbb{E}_x(\mathbb{1}_{H_x<\infty})\mathbb{E}_x(\mathbb{1}_{N_x\geqslant k})\\ &= \mathbb{P}_x(H_x<\infty)\mathbb{P}_x(N_x\geqslant k).\end{aligned}$$

Par récurrence, comme $\mathbb{P}_x(N_x \geqslant 1) = 1$ (par définition de N_x), on a donc

$$\mathbb{P}_x(N_x \geqslant k) = \mathbb{P}_x(H_x<\infty)^{k-1}.$$

On reconnait une loi géométrique de paramètre $\mathbb{P}_x(H_x<\infty) = 1-\mathbb{P}_x(H_x=\infty)$, et

$$\mathbb{E}_x(N_x) = \sum_{k=1}^{\infty}\mathbb{P}_x(N_x\geqslant k) = \frac{1}{\mathbb{P}_x(H_x=\infty)}.$$

□

Définition 3.3.3 (Noyau potentiel)**.** *Le noyau potentiel de la chaîne de Markov est*

$$U : \begin{cases} \mathcal{X}^2 & \longrightarrow [0,+\infty] \\ (x,y) & \longrightarrow \mathbb{E}_x(N_y). \end{cases}$$

Proposition 3.3.4. *1. Pour tout $x, y \in \mathcal{X}$, on a $U(x,y) = \sum_{n\in\mathbb{N}} Q_n(x,y)$, avec la convention $Q_0 = \mathrm{Id}$.*

2. $U(x,x) = +\infty \Leftrightarrow$ x est récurrent.

3. Pour tout $x \neq y$, on a $U(x,y) = \mathbb{P}_x(H_y<\infty)U(y,y)$.

Proof. 1. $U(x,y) = \sum_{n=0}^{\infty}\mathbb{E}_x(\mathbb{1}_{X_n=y}) = \sum\mathbb{P}_x(X_n=y) = \sum Q_n(x,y)$.

2. Immédiat par définition de U.

3. Par propriété de Markov forte,

$$\mathbb{E}_x(N_y) = \mathbb{E}_x(\mathbb{1}_{H_y<\infty}N_y((X_k)_{k\geqslant H_y})) = \mathbb{P}_x(H_y<\infty)\mathbb{E}_y(N_y).$$

□

Remarque 3.3.1. *$U(x,y) = 0$ ssi y est inaccessible depuis x.*
$U(x,y) > 0$ ssi il y a une probabilité positive de visiter y si l'on part de x.

Exemple 3.3.1. *Pour la marche aléatoire simple sur $\mathbb{Z}$ de paramètre p, $U(1,0) = 0$ si $p = 1$, sinon $U(1,0) > 0$.*

Exercice 3.3.1. *Soit $(Y_n^k)_{n\in\mathbb{N}}$, $k = 1,...d$ d copies indépendantes d'une marche aléatoires simple symétrique issue de 0. Montrer que $\mathbf{Y_n} := (\mathbf{Y_n^1},...,\mathbf{Y_n^d})$ est une chaîne de Markov, et calculer sa matrice de transition. Montrer ensuite que 0 est récurrent ssi $d < 3$.*

Solution 3.3.1. *On peut explicitement calculer*

$$\mathbb{P}(\mathbf{Y}_{n+1} = \mathbf{y}|\mathbf{Y}_1,...,\mathbf{Y}_n) = \frac{1}{2^d}\prod_{i=1}^{d}\mathbb{1}_{|Y_n^i - y^i|=1}.$$

Donc c'est bien une chaîne de Markov, de matrice de transition $Q(\mathbf{x},\mathbf{y}) = \frac{\mathbf{1}}{\mathbf{2^d}} \prod_{\mathbf{i=1}}^{\mathbf{d}} \mathbb{1}_{|\mathbf{x^i}-\mathbf{y^i}|=\mathbf{1}}$.

On a alors $Q_{2n+1}(0,0) = 0$ *pour tout* n, *et, par indépendance des coordonnées,*

$$Q_{2n}(0,0) = \mathbb{P}(Y_{2n}^1 = 0)^d \text{ et } \mathbb{P}(Y_{2n}^1 = 0) = 2^{-2n}\binom{2n}{n} \approx \frac{1}{\sqrt{\pi n}}$$

où on a utilisé la formule de Stirling. Donc $\sum Q_n(0,0) < \infty$ *ssi* $\sum n^{-d/2} < \infty$, *et donc ssi* $d \geqslant 3$.

Lemme 3.3.5. *Soit* $x \in \mathcal{X}$ *un point récurrent et* $y \in \mathcal{X}$ *tel que* $U(x,y) > 0$. *Alors* y *est récurrent, et* $\mathbb{P}_y(H_x < \infty) = 1$. *En particulier,* $U(y,x)$ *est alors aussi strictement positif.*

Une manière de comprendre ce lemme est que il n'est pas possible de visiter un état transitoire après un état récurrent.

Proof. Par hypothèse,

$$\begin{aligned} 0 &= \mathbb{P}_x(N_x < \infty) \\ &\geqslant \mathbb{P}_x(H_y < \infty,\ H_x((X_k)_{k \geqslant H_y}) = \infty) \\ &= \mathbb{P}_x(H_y < \infty)\mathbb{P}_y(H_x = \infty), \end{aligned}$$

où on a utilisé la propriété de Markov forte. Or comme $\mathbb{P}_x(H_y < \infty) \geqslant Q_p(x,y)$ pour tout $p \in \mathbb{N}^*$, comme $U(x,y) > 0$ on a $\mathbb{P}_x(H_y < \infty) > 0$. Donc c'est $\mathbb{P}_y(H_x = \infty)$ qui est nul.

Il existe alors $n_1, n_2 \in \mathbb{N}^*$ tels que $Q_{n_1}(x,y) > 0$ et $Q_{n_2}(y,x) > 0$. Comme

$$Q_{n_2+p+n_1}(y,y) \geqslant Q_{n_2}(y,x)Q_p(x,x)Q_{n_1}(x,y)$$

on a alors

$$U(y,y) \geqslant \sum_p Q_{n_2+p+n_1}(y,y) \geqslant Q_{n_2}(y,x)U(x,x)Q_{n_1}(x,y) > 0.$$

Donc y est bien récurrent. □

Théorème 3.3.6 (Classification des états)**.** *Soit* $\mathcal{R}$ *l'ensemble des points récurrents d'une chaîne de Markov donnée. Il existe une partition* $\mathcal{R} = \bigcup_{i \in I} \mathcal{R}_i$ *telle que*

- *Pour tout* $i \in I$, *si* $x \in \mathcal{R}_i$, *alors* $\mathbb{P}_x$*-p.s. on a* $N_y = +\infty$ *pour tout* $y \in \mathcal{R}_i$ *et* $N_y = 0$ *pour tout* $y \in \mathcal{X}\backslash\mathcal{R}_i$.
- *pour tout* $x \in \mathcal{X}\backslash\mathcal{R}$, *avec* $T = \inf\{n \in \mathbb{N}; X_n \in \mathcal{R}\}$, *on a soit* $T = +\infty$ *et* $N_y < \infty$ *pour tout* $y \in \mathcal{X}$, *soit* $T < \infty$ *et alors il existe* $j \in I$ *aléatoire, tel que que* $\forall n \geqslant T\ X_n \in \mathcal{R}_j$.

En résumé, on peut partitionner l'espace des points récurrents de manière à ce que l'on ne puisse pas sortir des $\mathcal{R}_i$. Les $\mathcal{R}_i$ sont appelés classes de récurrence de la chaîne de Markov.

Proof. Tout d'abord, pour $x, y \in \mathcal{R}$, la relation

$$x \equiv y \quad \Leftrightarrow \quad U(x,y) > 0$$

définit une relation d'équivalence sur $\mathcal{R}$, grâce au Lemme 3.3.5. On peut donc partitionner $\mathcal{R}$ en ses classes d'équivalences, qui définissent les $\mathcal{R}_i$ du théorème.

Soit $i \in I$ est $x \in \mathcal{R}_i$. Alors pour tout $y \in \mathcal{X}\backslash\mathcal{R}_i$ on a $U(x,y) = 0$, et donc $N_y = 0$ $\mathbb{P}_x$-p.s. Sinon, pour $y \in \mathcal{R}_i$, on a $\mathbb{P}_x(H_y < \infty) = 1$ et

$$\begin{aligned}\mathbb{P}_x(N_y = \infty) &= \mathbb{E}_x[\mathbb{1}_{H_y<\infty}\mathbb{1}_{N_y((X_k)_{k\geqslant H_y})=\infty}] \\ &= \mathbb{P}_x(H_y < \infty)\mathbb{P}_y(N_y = \infty) \\ &= 1.\end{aligned}$$

Si $x \in \mathcal{X}\backslash\mathcal{R}$ et $T = \infty$, on ne visite jamais $\mathcal{R}$, et pour $y \in \mathcal{X}\backslash\mathcal{R}$, on a

$$\mathbb{E}_x(N_y) = \mathbb{P}_x(H_y < \infty)\mathbb{E}_y(N_y) < \infty.$$

Enfin, si $x \in \mathcal{X}\backslash\mathcal{R}$ et $T < \infty$, soit j tel que $X_T \in \mathcal{R}_j$. Par propriété de Markov forte appliquée à T, la première partie du théorème implique que pour tout $n \geqslant T$ on a $X_n \in \mathcal{R}_j$. □

Exemple 3.3.2. *Pour la marche aléatoire simple sur $\mathbb{Z}$ de paramètre p, tous les états sont récurrents si $p = 1/2$, et ils sont tous transitoires si $p \neq 1/2$.*

Définition 3.3.7 (Chaînes irréductibles)**.** *Une chaîne de Markov est dite irréductible si pour tout $x, y \in \mathcal{X}$ on a $U(x,y) > 0$.*

Une chaîne irréductible est une chaîne où il est possible d'accéder à n'importe quel état, quelque soit la condition initiales. Il n'y a pas de découpage en plusieurs classes de récurrence.

Proposition 3.3.8. *Si une chaîne de Markov est irréductible, on a l'alternative suivante :*

- *Ou bien tous les états sont transitoires, et pour tout $x \in \mathcal{X}$ on a*

$$\mathbb{P}_x(N_y < \infty\ \forall y) = 1;$$

- *Ou bien tous les états sont récurrents, il y a une unique classe de récurrence, et pour tout $x \in \mathcal{X}$ on a*

$$\mathbb{P}_x(N_y = \infty\ \forall y) = 1.$$

Lorsque $\mathcal{X}$ est fini, seul le second cas peut se produire, car $\sum_y N_y = \infty$.

Proof. Si il existe x récurrent, alors tous les points le sont, d'après le Lemme 3.3.5, et par constructions des classes de récurrence il ne peut y en avoir qu'une. Le reste découle du théorème de classification des états. □

Exemple 3.3.3. *Une marche aléatoire sur un graphe fini connexe est irréductible et récurrente.*

Exercice 3.3.2. *Soit*

$$Q := \begin{pmatrix} 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 0 & 0 & 0 & 1/2 \end{pmatrix}.$$

Déterminer les états transitoires et les classes de récurrence de Q.

Solution 3.3.2. *L'ensemble des points transitoires est $\{2,4\}$, et les classes de récurrence sont $\{1,5\}$ et $\{3\}$.*

3.3.2 Mesures invariantes

On cherche maintenant à étudie les lois stationnaires des chaînes de Markov, c'est à dire les lois μ telles que si $X_n \equiv \mu$, alors $X_{n+1} \equiv \mu$.

Définition 3.3.9 (Mesures invariantes)**.** *On considère une chaîne de Markov de matrice de transition Q. Soit μ une mesure positive sur $\mathcal{X}$, telle que $\mu(x) < \infty$ pour tout $x \in \mathcal{X}$.*

On dit que μ est invariante pour la chaîne de Markov si pour tout $y \in \mathcal{X}$ on a

$$\sum_x \mu(x)Q(x,y) = \mu(y).$$

Interprétation : si $\mu(\mathcal{X}) < \infty$, quitte à remplacer μ par $\mu(\mathcal{X})^{-1}\mu$, on peut supposer $\mu(\mathcal{X}) = 1$. Alors

$$\mathbb{E}_\mu[f(X_1)] = \sum_x \mu(x) \sum_y Q(x,y)f(y) = \sum_y \mu(y)f(y) = \mathbb{E}_\mu[f(X_0)]$$

c'est à dire que μ est stationnaire pour la chaîne de Markov. Mais attention, cette définition autorise des mesures positives de masse arbitraire, y compris infinie.

Exemple 3.3.4. *Sur $\mathbb{Z}^d$, si on considère une matrice de transition de la forme $Q(x,y) = \gamma(y-x)$, alors la mesure de comptage est invariante.*

Notation. *Si on identifie μ à un vecteur ligne, une mesure invariante vérifie $\mu Q = \mu$. On utilisera cette notation par la suite. Par une récurrence immédiate, si μ est invariante, alors pour tout n on a $\mu Q_n = \mu$.*

Attention, l'ordre est important, si on identifie μ à un vecteur colonne, on n'a pas en général $Q\mu = \mu$.

Exercice 3.3.3. *Quelles sont les mesures de probabilité invariantes pour une chaîne de Markov sur $\{1,2,3\}$ de matrice de transition*

$$Q = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \end{pmatrix}$$

Solution 3.3.3. *On résoud le système linéaire associé à $\mu Q = \mu$ pour trouver $\mu(1) = \mu(3) = 2/5$ et $\mu(2) = 1/5$.*

Définition 3.3.10 (Mesures réversibles)**.** *Une mesure positive μ est dite réversible pour la chaîne de Markov si $\forall x, y$ on a $\mu(x)Q(x,y) = \mu(y)Q(y,x)$.*

Interprétation : sous la mesure μ, le courant le long d'une arrête est nul : $\mathbb{P}_\mu(X_0 = x, X_1 = y) = \mathbb{P}_\mu(X_0 = y, X_1 = x)$.

Proposition 3.3.11. *Une mesure réversible est invariante.*

Proof. On a bien

$$\mu(x) = \sum_y Q(x,y)\mu(x) = \sum_y Q(y,x)\mu(y).$$

□

En revanche, il peut exister des mesures invariantes non-réversibles.

Exemple 3.3.5. *Soit γ une mesure de probabilité sur $\mathbb{Z}$, non symétrique (i.e. il existe $x \in \mathbb{Z}$ tel que $\gamma(x) \neq \gamma(-x)$). On considère la marche aléatoire sur $\mathbb{Z}$ de loi de sauts γ. On a vu que la mesure de comptage est invariante, mais elle n'est pas réversible, car $Q(0,x) = \gamma(x) \neq Q(x,0)$.*

En revanche, si γ est symétrique, la mesure de comptage est réversible.

Exemple 3.3.6. *Soit $p \in (0,1)$ et $q = 1-p$. On définit la matrice de transition sur $\mathbb{Z}$*

$$Q(i, i+1) = p; \quad Q(i, i-1) = q$$

i.e. la matrice de transition de la marche aléatoire simple. La mesure de comptage est invariante, mais la mesure

$$\mu(i) = \left(\frac{p}{q}\right)^i; \quad i \in \mathbb{Z}$$

l'est aussi, car elle est réversible.

Les mesures invariantes ne sont donc pas nécessairement uniques, même à constante multiplicative près.

Exemple 3.3.7. *Soit $G = (S, A)$ un graphe. La mesure $\mu(x) = \mathrm{Card}\{y; (x,y) \in A\}$ est réversible pour la marche aléatoire sur le graphe. En effet, pour tout x, y, on a $\mu(x)Q(x,y) = \mathbb{1}_{x \equiv y}$.*

Théorème 3.3.12. *Soit x un point récurrent. Alors la mesure définie par*

$$\mu(y) := \mathbb{E}_x \left(\sum_{k=0}^{H_x - 1} \mathbb{1}_{X_k = y} \right)$$

est invariante. De plus, $\mu(y) > 0$ ssi y est dans la même classe de récurrence que x.

En général, ce n'est pas une mesure de probabilité, car on a choisit la normalisation $\mu(x) = 1$.

En particulier, si la chaîne de Markov a plusieurs classes de récurrence, on définit ainsi plusieurs mesures invariantes, à supports disjoints. En revanche, si la chaîne est récurrente irréductible, on verra plus tard qu'il n'y a qu'une seule mesure invariante (à constante multiplicative près).

Ce théorème est en pratique rarement utile pour calculer des mesures invaraintes, il est plutôt utilisé pour établir des propriétés des temps d'arrêt H_x.

Proof. Tout d'abord, si y n'est pas dans la classe de récurrence de x, alors $\mathbb{P}_x$ p.s. y ne sera jamais visité, et donc $\mu(y) = 0$.

Soit y dans la même classe de récurrence que x. Comme sous $\mathbb{P}_x$ on a $X_{H_x} = X_0$, on peut changer l'indexation

$$\mu(y) = \mathbb{E}_x\left(\sum_{k=0}^{H_x-1} \mathbb{1}_{X_k=y}\right) = \mathbb{E}_x\left(\sum_{k=1}^{H_x} \mathbb{1}_{X_k=y}\right).$$

On a ensuite

$$\begin{aligned}\mu(y) &= \sum_z \mathbb{E}_x\left(\sum_{k=1}^{H_x} \mathbb{1}_{X_k=y, X_{k-1}=z}\right) \\ &= \sum_z \sum_{k=1}^{\infty} \mathbb{E}_x\left(\mathbb{1}_{X_k=y}\mathbb{1}_{k\leqslant H_x, X_{k-1}=z}\right)\end{aligned}$$

Comme $\{k \leqslant H_x, X_{k-1} = z\}$ est $\mathcal{F}_{k-1}$-mesurable, on peut appliquer la propriété de Markov simple au temps $k-1$, et obtenir

$$\mathbb{E}_x\left(\mathbb{1}_{X_k=y}\mathbb{1}_{k\leqslant H_x, X_{k-1}=z}\right) = \mathbb{E}_x\left(\mathbb{1}_{k\leqslant H_x, X_{k-1}=z}\right) Q(z,y).$$

On a donc

$$\begin{aligned}\mu(y) &= \sum_z \sum_{k=1}^{\infty} \mathbb{E}_x\left(\mathbb{1}_{k\leqslant H_x, X_{k-1}=z}\right) Q(z,y) \\ &= \sum_z \mathbb{E}_x\left(\sum_{k=1}^{H_x} \mathbb{1}_{X_{k-1}=z}\right) Q(z,y) \\ &= \sum_z \mu(z) Q(z,y).\end{aligned}$$

μ est donc bien une mesure invariante pour la chaîne de Markov.

Pour finir, il nous reste à montrer que si y est dans la classe de récurrence de x, alors $\mu(y) > 0$ et est fini. Comme $\mu Q_n = \mu$, pour tout n on a

$$\mu(y) = \sum_z \mu(z) Q_n(z,y) \geqslant \mu(x) Q(x,y).$$

Or par définition des classes de récurrence il existe n tel que $Q_n(x,y) > 0$. Et comme par définition $\mu(x) = 1$, on a bien $\mu(y) > 0$. De même, il existe m tel que $Q_m(y,x) > 0$, et donc

$$1 = \mu(x) \geqslant \mu(y) Q_m(y,x)$$

et donc $\mu(y)$ est bien fini. □

Proposition 3.3.13. *Si une chaîne de Markov irréductible récurrente admet deux mesures invariantes μ_1 et μ_2, alors il existe une constante C telle que $\mu_1 = C\mu_2$.*

En particulier, il y a au plus une seule mesure de probabilité invariante (mais il peut n'y avoir que des mesures invariantes de masse totale infinie, et donc aucune mesure de probabilité invariante).

Proof. Soit μ une mesure invariante non nulle. Nous allons commencer par montrer par récurrence que pour tout $p \in \mathbb{N}$, on a pour tout $x, y \in \mathcal{X}$

$$\mu(y) \geqslant \mu(x)\mathbb{E}_x\left(\sum_{k=0}^{p\wedge(H_x-1)} \mathbb{1}_{X_k=y}\right).$$

Pour $p = 0$, ou si $x = y$, cette inégalité est trivialement vraie. Supposons la vérifiée pour un $p \in \mathbb{N}$ donné. Alors, pour $x \neq y$, on a

$$\begin{aligned}
\mu(y) &= \sum_z \mu(z)Q(z,y) \\
&\geqslant \mu(x)\sum_z \mathbb{E}_x\left(\sum_{k=0}^{p\wedge(H_x-1)} \mathbb{1}_{X_k=z}\right) Q(z,y) \\
&= \mu(x)\sum_z \sum_{k=0}^{p} \mathbb{E}_x\left(\mathbb{1}_{X_k=z,k\leqslant H_x-1}\mathbb{1}_{X_{k+1}=y}\right) \\
&= \mu(x)\mathbb{E}_x\left(\sum_{k=0}^{p\wedge(H_x-1)} \mathbb{1}_{X_{k+1}=y}\right) \\
&= \mu(x)\mathbb{E}_x\left(\sum_{k=0}^{(p+1)\wedge(H_x-1)} \mathbb{1}_{X_k=y}\right)
\end{aligned}$$

où on a utilisé à la dernière étape $x \neq y$. Si on pose $\nu_x(y) := \mathbb{E}_x\left(\sum_{k=0}^{H_x-1} \mathbb{1}_{X_k=y}\right)$ la mesure invariante donnée par le Théorème 3.3.12, en faisant tendre p vers l'infini on obtient alors

$$\mu(y) \geqslant \mu(x)\nu_x(y) \quad \forall x, y \in \mathcal{X}.$$

Donc pour tout $n \geqslant 1$, comme $\nu_x(x) = 1$ et $\nu_x Q_n = \nu$, on a

$$\mu(x) = \sum_z \mu(z)Q_n(z,x) \geqslant \mu(x)\sum_z \nu_x(z)Q_n(z,x) = \mu(x)\nu_x(x) = \mu(x).$$

On en déduit qu'il y a nécessairement égalité au milieu, et donc pour tout z tel qu'il existe n avec $Q_n(z,y) > 0$ on a

$$\mu(z) = \mu(x)\nu_x(z).$$

L'irréducibilité de la chaîne permet alors de conclure que $\mu = \mu(x)\nu_x$ pour x fixé. □

Corollaire 3.3.14. *Considérons une chaîne de Markov récurrente irréductible. Alors on a l'alternative*

- *Ou bien il existe une mesure de probabilité invariante μ, et alors $\forall x \in \mathcal{X}$ on a*
$$\mathbb{E}_x(H_x) = \frac{1}{\mu(x)}.$$
On dit alors que la chaîne est récurrente positive.

- *Ou bien toute mesure invariante a une masse totale infinie, et alors* $\forall x \in \mathcal{X}$ *on a*

$$\mathbb{E}_x(H_x) = +\infty.$$

On dit alors que la chaîne est récurrente nulle.

Si $\mathcal{X}$ *est un ensemble fini, seul le premier cas peut se produire.*

Proof. Tout d'abord, comme les mesures invariantes sont nécessairement de la forme $C\nu_x$ avec $\nu_x(y) := \mathbb{E}_x\left(\sum_{k=0}^{H_x-1} \mathbb{1}_{X_k=y}\right)$. Elles sont donc toutes de masse finie, ou toutes de masse infini, selon la masse de ν_x.

Si μ est une mesure de probabilité invariante, $\mu = C\nu_x$ et $\mu(\mathcal{X}) = 1$ donc $C = \nu_x(\mathcal{X})^{-1}$. En particulier, comme $\nu_x(x) = 1$, $\mu(x) = \nu_x(\mathcal{X})^{-1}$. Or

$$\nu_x(\mathcal{X}) = \sum_y \mathbb{E}_x\left(\sum_{k=0}^{H_x-1} \mathbb{1}_{X_k=y}\right) = \mathbb{E}_x(H_x),$$

et donc on a bien $\mu(x) = \mathbb{E}_x(H_x)^{-1}$.

Dans le second cas, en particulier ν_x est de masse infinie, et donc $\mathbb{E}_x(H_x) = \nu_x(\mathcal{X}) = +\infty$. □

Proposition 3.3.15. *Supposons la chaîne irréductible. S'il existe une mesure invariante de masse totale finie, alors la chaîne est récurrente positive.*

Proof. On a vu que

$$U(y,y) = \frac{U(x,y)}{\mathbb{P}_x(H_y < \infty)} \geqslant U(x,y) = \sum_n Q_n(x,y).$$

Soit γ une mesure invariante finie. On a

$$\begin{aligned}
\gamma(\mathcal{X})U(y,y) &= \sum_x \gamma(x)U(y,y) \\
&\geqslant \sum_x \sum_n \gamma(x)Q_n(x,y) \\
&= \sum_n \gamma(y) \\
&= +\infty.
\end{aligned}$$

Comme $\gamma(\mathcal{X}) < \infty$, on a donc bien $U(y,y) = \infty$ pour tout y, et la chaîne est donc bien récurrente. □

Remarque 3.3.2. *L'existence d'une mesure invariante de masse infinie ne dit rien sur la récurrence. Une marche aléatoire simple sur* $\mathbb{Z}$ *admet toujours la mesure de comptage comme mesure invariante de masse infinie, mais la récurrence dépend de la valeur du paramètre (récurrente si* $p = 1/2$*, transitoire sinon).*

Exercice 3.3.4. *Soit* Q *la matrice de transition*

$$Q := \begin{pmatrix} 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 0 & 0 & 1/2 \end{pmatrix}.$$

Montrer que la chaîne est irréductible, et calculer sa mesure de probabilité invariante.

Solution 3.3.4. *On peut vérifier que les valeurs suivantes sont strictement positives : $Q(1,5), Q_2(1,2), Q_3(1,4), Q_4(1,3)$, et donc on peut accéder à n'importe quel site depuis le site 1. Réciproquement, $Q(5,1), Q(3,1), Q_2(4,1)$ et $Q_3(2,1)$ sont strictement positifs, donc on peut accéder au site 1 depuis n'importe quel site. Tous les sites communiquent, et donc la chaîne est récurrente. La mesure invariante est* $(3/13, 3/13, 1/13, 2/13, 4/13)$.

Exercice 3.3.5. *On définit*

$$\begin{cases} P(0,0) & = 0; \\ P(0,k) & = \frac{1}{k} - \frac{1}{k+1} \ \forall k \in \mathbb{N}^* \\ P(k,\ell) & = 0 \ \forall 1 \leqslant k \leqslant \ell \\ P(\ell,k) & = \frac{1}{\ell} \ \forall 0 \leqslant k < \ell \end{cases}$$

1. *Vérifier que P définit une matrice de transition sur $\mathbb{N}$.*

2. *Montrer qu'une chaîne de Markov de matrice de transition P est irréductible récurrente.*

3. *Montrer qu'elle est récurrente positive en calculant une mesure invariante.*

Solution 3.3.5.

1. *On a bien $\sum_k P(\ell,k) = 1$ pour tout $\ell \in \mathbb{N}$, et les $P(k,\ell)$ sont positifs.*

2. *Montrons d'abord que 0 est récurrent. Comme si $X_n \neq 0$, on a $X_{n+1} < X_n$, de tout point k on va en au plus k pas en 0. Donc p.s. une chaîne issue de 0 retourne en 0 en un certain temps, et donc 0 est récurrent.*

   La chaîne est de plus irréductible, car $P(0,\ell)$ et $P(\ell,0)$ sont strictement positifs pour tout $\ell \geqslant 1$. Donc la chaîne est récurrente, car elle est irréductible et a un état récurrent.

3. *Soit $a_k = \mu(k)$ une mesure invariante. On a alors*

$$a_0 = \sum_{k \geqslant 1} a(k)/k; \quad a_\ell = \frac{a_0}{\ell(\ell+1)} + \sum_{k>\ell} a_k/k.$$

   En posant $b_\ell = \sum_{k>\ell} a_k/k$, on a donc

$$a_0 = b_0; \quad \ell(b_{\ell-1} - b_\ell) = \frac{b_0}{\ell(\ell+1)} + b_\ell \ \forall \ell \geqslant 1.$$

   Si on pose $c_\ell = \ell b_{\ell-1}$ pour $\ell \geqslant 1$, on a alors

$$c_\ell - c_{\ell+1} = \frac{b_0}{\ell(\ell+1)}$$

   donc

$$c_\ell = b_0 \sum_{k \geqslant \ell} \frac{1}{k(k+1)}.$$

D'où

$$b_\ell = \frac{b_0}{\ell} \sum_{k \geqslant \ell+1} \frac{1}{k(k+1)}; \quad a_\ell = b_0 \left(\frac{\ell}{(\ell-1)\ell(\ell+1)} + \frac{1}{\ell+1} \sum_{k \geqslant \ell+1} \frac{1}{k(k+1)} \right).$$

Comme $a_\ell = O(1/\ell^2)$, on a une mesure invariante de masse totale finie, et donc la chaîne est récurrente positive.

3.4 Théorème ergodique

Le but de cette section est l'étude du comportement en temps long des fonctionnelles additives d'une chaîne de Markov.

3.4.1 Théorème principal

Théorème 3.4.1. *On considère une chaîne de Markov irréductible récurrente positive, avec μ son unique mesure invariante, et f une fonction mesurable sur $\mathcal{X}$ avec $||f||_{L^1(\mu)} < \infty$. Alors $\mathbb{P}_x$-p.s. on a*

$$\frac{1}{n} \sum_{k=0}^{n} f(X_k) \underset{n \to \infty}{\longrightarrow} \mathbb{E}_\mu(f).$$

C'est une forme de loi des grands nombres, mais ici les $f(X_k)$ sont des variables corrélées.

Corollaire 3.4.2. *On considère une chaîne de Markov récurrente irréductible. Alors*

- *Si elle est récurrente positive,*

$$\frac{1}{n} \sum_{k=0}^{n} \mathbb{1}_{X_k = x} \xrightarrow{p.s.} \mu(x)$$

  où μ est l'unique mesure de probabilité invariante.

- *Si elle est récurrente nulle,*

$$\frac{1}{n} \sum_{k=0}^{n} \mathbb{1}_{X_k = x} \xrightarrow{p.s.} 0.$$

Preuve du théorème ergodique. On définit par récurrence la suite de temps d'arrêt T_n avec $T_0 = 0$ et

$$T_{n+1} := \inf\{k > T_n; X_k = x\}.$$

Quitte à décomposer $f = f_+ - f_-$, on va supposer f positive.

On pose ensuite

$$Y_n := \sum_{k=T_n}^{T_{n+1}-1} f(X_k).$$

Comme conséquence de la propriété de Markov forte, $(Y_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires i.i.d. De plus, si ν_x est la mesure invariante donnée par le Théorème 3.3.12, comme $\mu = \mu(x)\nu_x$, on a

$$\mathbb{E}_x[Y_n] = \frac{1}{\mu(x)}\mathbb{E}_\mu[f].$$

Par la loi forte des grands nombres appliquées à la suite $(Y_n)_{n\in\mathbb{N}}$, on a donc

$$\frac{1}{n}\sum_{k=0}^{n} Y_k \xrightarrow{p.s.} \frac{1}{\mu(x)}\mathbb{E}_\mu[f].$$

Si on pose alors $N_x(n)$ le nombre de retours en x avant l'instant n, on a pour tout n

$$T_{N_x(n)} \leqslant n < T_{N_x(n)+1} \tag{3.5}$$

et donc

$$\frac{1}{N_x(n)}\sum_{k=0}^{T_{N_x(n)}-1} f(X_k) \leqslant \frac{1}{N_x(n)}\sum_{k=0}^{n} f(X_k) \leqslant \frac{1}{N_x(n)}\sum_{k=0}^{T_{N_x(n)+1}-1} f(X_k)$$

ce qui revient à

$$\frac{1}{N_x(n)}\sum_{k=0}^{N_x(n)-1} Y_k \leqslant \frac{1}{N_x(n)}\sum_{j=0}^{n} f(X_j) \leqslant \frac{1}{N_x(n)}\sum_{k=0}^{N_x(n)} Y_k$$

et donc par encadrement

$$\frac{1}{N_x(n)}\sum_{j=0}^{n} f(X_j) \xrightarrow{p.s.} \frac{1}{\mu(x)}\mathbb{E}_\mu[f].$$

Enfin, comme les $T_{n+1} - T_n$ sont iid, et d'espérance $\frac{1}{\mu(x)}$ d'après le Corollaire 3.3.14,

$$\frac{T_n}{n} \xrightarrow{p.s.} \frac{1}{\mu(x)}, \text{ et donc } \frac{N_x(n)}{n} \xrightarrow{p.s.} \mu(x),$$

où on a utilisé (3.5) ce qui permet de conclure la preuve.

□

Exercice 3.4.1. *Soit Q la matrice de transition sur $\{0,1\}$ définie par*

$$Q = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

avec $0 < \alpha, \beta < 1$ et $\min(\alpha,\beta) < 1$.

1. *Diagonaliser Q, et en déduire*

$$Q^n = \frac{1}{\alpha+\beta}\begin{pmatrix} \beta & \alpha \\ \beta & \alpha \end{pmatrix} + \frac{(1-\alpha-\beta)^n}{\alpha+\beta}\begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix}.$$

2. *Calculer la mesure de probabilité invariante π de la chaîne de Markov associée.*

3. *Calculer* $\mathrm{Cov}_\pi(X_n, X_{n+p}) = \mathbb{E}_\pi[X_n X_{n+p}] - \mathbb{E}[X_n]\mathbb{E}[X_{n+p}]$.

4. *Etudier le comportement asymptotique de* $n^{-1}\sum_{i=1}^n X_i$, *où* X *est un chaîne de Markov de matrice de transition* Q.

Solution 3.4.1. 1. *Les valeurs propres de* Q *sont* 1 *et* $s = 1-\alpha-\beta$, *de vecteurs propres colonnes respectifs* $(1,1)^t$ *et* $(-\alpha,\beta)^t$. *On prend comme matrice de passage*

$$P = \begin{pmatrix} 1 & -\alpha \\ 1 & \beta \end{pmatrix}; \quad P^{-1} = \frac{1}{\alpha+\beta}\begin{pmatrix} \beta & \alpha \\ -1 & 1 \end{pmatrix}.$$

On a alors

$$Q^n = P\begin{pmatrix} 1 & 0 \\ 0 & 1-\alpha-\beta \end{pmatrix}^n P^{-1},$$

et le résultat suit.

2. *Comme une mesure invariante est un vecteur propre (à gauche) de* Q^n, *de valeur propre* 1, *c'est aussi un vecteur propre de* Q^n. *En faisant tendre* n *vers l'infini, on voit que ce vecteur propre (normalisé pour que la somme fasse* 1*) est*

$$\pi = \left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right).$$

3. *On calcule, et on trouve*

$$\mathbb{E}_\pi[X_n] = \frac{\alpha}{\alpha+\beta}; \quad \mathbb{E}_\pi[X_n X_{n+p}] = \frac{\alpha}{\alpha+\beta}Q^p(1,1) = \frac{\alpha(\alpha+\beta(1-\alpha-\beta)^p)}{(\alpha+\beta)^2}.$$

On a donc

$$\mathrm{Cov}_\pi(X_n, X_{n+p}) = \frac{\beta(1-\alpha-\beta)^p}{(\alpha+\beta)^2}.$$

4. *La chaîne de Markov est irréductible, récurrente positive donc en appliquant le théorème ergodique,* $n^{-1}\sum_{i=1}^n X_i$ *converge presque sûrement vers* $\mathbb{E}_\pi[X] = \frac{\alpha}{\alpha+\beta}$.

3.4.2 Une brève introduction aux algorithmes MCMC

Une des applications principales du théorème ergodique est qu'il justifie certaines méthodes probailistes pour le calcul numérique. Par exemple, on peut chercher à calculer des quantités de la forme

$$\sum_{\mathbf{n}=(\mathbf{n_1},\ldots,\mathbf{n_d})\in\mathbb{Z}^\mathbf{d}} f(\mathbf{n})\rho(\mathbf{n})$$

où f est une fonction par exemple bornée, et ρ une mesure de probabilité sur $\mathbb{Z}^d$. Le calcul numérique de cette some par des méthodes déterministes peut être très lent, surtout si d est grand. Certains algorithmes de calcul utilisent à la place des méthodes probabilistes : on construit une chaîne de Markov réversible par rapport à la mesure ρ, on en simule une réalisation $(X_n)_{n\leqslant N}$ pour un temps N grand, et on approxime la somme par la moyenne empirique

$$\frac{1}{N}\sum f(X_n)$$

Le théorème ergodique garantit que dans la limite $N \to \infty$, cette moyenne empirique converge vers la bonne valeur.

Avantage : la complexité de l'algorithme peut être beaucoup plus faible que celle des algorithmes suivant une méthode déterministe.

Inconvénient : contrairement aux méthodes déterministes, la quantité $\frac{1}{N}\sum f(X_n)$ est aléatoire. En particulier, elle a une variance stictement positive, et on a une petite probabilité que l'erreur d'approximation par une réalisation de la chaîne soit très grande. Cette variance décroit typiquement en $1/N$.

Une manière de construire une chaîne de Markov sur $\mathbb{Z}^d$ avec la bonne mesure invarainte est donnée par l'algorithme de Metropolis-Hastings : on consdière une marche aléatoire simple sur Z^d, qu'on déforme pour imposer la bonne mesure invariante. Pour faire ça, à chaque temps, on procède de la manière suivante :

1. Etant donné la position $X_n \in \mathbb{Z}^d$, on génère une variable ε_{n+1} uniforme sur $\{-1, 1\}^d$;

2. On considère les quantité $\rho(X_n)$ et $\rho(X_n + \varepsilon_{n+1})$. Si $\rho(X_n) \leqslant \rho(X_n + \varepsilon_{n+1})$, on pose $X_{n+1} = X_n + \varepsilon_{n+1}$ (acceptation).

3. Si $\rho(X_n) > \rho(X_n + \varepsilon_{n+1})$, on génère une variable U_{n+1} uniforme sur $[0, 1]$, indépendante des autres variables. Si $U_n < \rho(X_n + \varepsilon_{n+1})/\rho(X_n)$, on pose $X_{n+1} = X_n + \varepsilon_{n+1}$ (acceptation), et sinon on pose $X_{n+1} = X_n$ (rejet).

On a donc superposé une procédure d'acceptation/rejet sur la marche aléatoire simple, qui cherche à favoriser les points où $\rho(\mathbf{n})$ est grand. Le processus aléatoire ainsi construit est une chaîne de Markov, et il s'avère que la mesure ρ est réversible pour cette chaîne de Markov. En effet, si x et y sont voisins, et si $\rho(x) > \rho(y)$ (ce qu'on peut supposer par symétrie),

$$\mathbb{P}(X_{n+1} = x | X_n = y) = 2^{-d}; \mathbb{P}(X_{n+1} = y | X_n = x) = \frac{2^{-d}\rho(y)}{\rho(x)}$$

et donc

$$\rho(y)\mathbb{P}(X_{n+1} = x | X_n = y) = \rho(x)\mathbb{P}(X_{n+1} = y | X_n = x).$$

De plus, si $\rho > 0$ en tout point, alors cette chaîne est irréductible. On a donc une manière simple de construire une chaîne de Markov avec les bonnes propriétés. Il existe beaucoup d'autres manières de construire de telles chaînes de Markov, parfois plus efficaces.

Pour citer un exemple, on peut mentionner l'algorithme PageRank de Google, qui cherche à trier les pages internet en fonction du nombre de liens qui y mènent. Pour ce faire, on génère une chaîne de Markov sur l'ensemble des pages web, où à chaque étape on choisit une nouvelle page web uniformément parmis toutes les pages webs auxquelles ont peut accéder depuis la page courante. La mesure invariante de ce processus donne plus de poids aux pages avec beaucoup de liens, et pour estimer cette mesure invariante, on fait tourner la chaîne pour un temps très long.

Un avantage pratique de cet algorithme est qu'il ne faut connaître la densité ρ qu'à une constante multplicative près (puisqu'on utilise seulement les ratios de densité). Ca peut paraître anecdotique, mais il existe de nombreuses distributions calculées en pratique, notamment en physique statistique, pour lesquelles on a une formule explicite qu'à cette constante près.

Chapter 4

Martingales et chaines de Markov

4.1 Fonctions harmoniques

Si $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov, et $f : \mathcal{X} \longrightarrow \mathbb{R}$ est bornée, alors $(f(X_n))_{n\in\mathbb{N}}$ est un processus à valeurs rélles et intégrable. On peut donc s'intéresser à sa décomposition de Doob-Meyer.

La partie prévisible est alors

$$\sum_{k=0}^{n-1} \mathbb{E}[f(X_{k+1}) - f(X_k)|\mathcal{F}_k] = \sum_{k=0}^{n-1} Qf(X_k) - f(X_k).$$

En particulier, si f, vu comme un vecteur colonne de $\mathbb{R}^{|\mathcal{X}|}$, est dans le noyau du générateur $Q - \mathrm{Id}$, alors $(f(X_n))_{n\in\mathbb{N}}$ est une martingale. Cette propriété motive la définition suivante :

Définition 4.1.1. *Une fonction f est dite harmonique en un point $x \in \mathcal{X}$ si $\mathbb{E}_x[f(X_1)] = f(x)$.*

Une fonction f est dite harmonique sur l'ensemble $A \subset \mathcal{X}$ si pour tout $x \in A$ f est harmonique en x.

Si on considère la marche aléatoire simple symétrique sur $\mathbb{Z}^d$, une fonction est harmonique en x si

$$f(x) = \frac{1}{2d} \sum_{y\equiv x} f(y)$$

i.e. si f vérifie une propriété de moyenne discrète, analogue à la propriété de la moyenne des fonctions harmoniques en analyse complexe, ce qui motive la terminologie. Ce lien n'est d'ailleurs pas juste une analogie, les deux notions sont exactement les mêmes si on étend le cadre aux chaînes de Markov en temps continu sur $\mathbb{C}$, et qu'on considère comme processus de Markov le mouvement brownien. Mais ceci va au delà du programme de ce cours.

Exercice 4.1.1. *Quelles sont les fonctions harmoniques sur $\mathbb{Z}$ pour la marche aléatoire simple symétrique?*

Solution 4.1.1. *Soit f une fonction harmonique sur $\mathbb{Z}$. On a pour tout n*

$$f(n+1) = 2f(n) - f(n-1).$$

On reconnaît une relation de récurrence d'ordre 2. Le polynome caractéristique est $(x-1)^2$, donc f est donnée par la formule

$$f(n) = n(f(1) - f(0)) + f(0).$$

Proposition 4.1.2. *Soit f une fonction harmonique sur $A \subset \mathcal{X}$, et T un temps d'arrêt tel que pour tout $n < T$ on a $X_n \in A$. Alors si $X_0 = x \in A$, $(f(X_{n\wedge T}))_{n\in\mathbb{N}}$ est une martingale.*

En particulier, pour tout $n \in \mathbb{N}$, $\mathbb{E}_x[f(X_{n\wedge T})] = f(x)$.

Ce théorème s'applique notamment au temps d'arrêt $T := \inf\{n; X_n \notin A\}$.

Proof. Il nous suffit de montrer que pour tout $n \in \mathbb{N}$ on a

$$\mathbb{E}[f(X_{(n+1)\wedge T})|\mathcal{F}_n] = f(X_{n\wedge T}).$$

Or on a

$$\begin{aligned}\mathbb{E}[f(X_{(n+1)\wedge T})|\mathcal{F}_n] &= \mathbb{E}[f(X_{(n+1)\wedge T})\mathbb{1}_{T\leqslant n}|\mathcal{F}_n] + \mathbb{E}[f(X_{(n+1)\wedge T})\mathbb{1}_{T>n}|\mathcal{F}_n] \\ &= f(X_T)\mathbb{1}_{T\leqslant n} + Qf(X_n)\mathbb{1}_{T>n} \\ &= f(X_T)\mathbb{1}_{T\leqslant n} + f(X_n)\mathbb{1}_{T>n} \\ &= f(X_{n\wedge T}).\end{aligned}$$

□

Définition 4.1.3 (Problème de Dirichlet)**.** *Soit $A \subset \mathcal{X}$, et $g : A \longrightarrow \mathbb{R}$. On dit que f est une solution au problème de Dirichlet associé à A et g si $f = g$ sur A et f est harmonique sur A^c.*

Lemme 4.1.4. *On considère une chaîne de Markov irréductible, de matrice de transition Q. Supposons que A est non-vide, et que A^c est fini et non-vide. Alors la seule solution au problème de Dirichlet avec $g = 0$ sur A est la fonction nulle.*

Proof. Soit f une solution du problème de Dirichlet, et $x_0 \in \operatorname{argmax}_{A^c} f$. Nous allons montrer par l'absurde que $f(x_0) \leqslant 0$. Si on a $f(x_0) > 0$, $f(x_0)$ est le maximum de f sur $\mathcal{X}$, puisque f est nulle sur A. Soit T le temps d'arrêt $\inf\{n; X_n \in A\}$. Comme la chaîne est irréductible, $\mathbb{P}(T = \infty) < 1$. Soit n tel que $\mathbb{P}_{x_0}(T \leqslant n) > 0$.

$$f(x_0) = \mathbb{E}_{x_0}[f(X_{n\wedge T})] \leqslant f(x_0)(1 - \mathbb{P}_{x_0}(T > n)) + 0 \times \mathbb{P}(T \leqslant n) < f(x_0).$$

On a une contradiction, donc $f(x_0) \leqslant 0$, et donc f est négative. Mais comme $-f$ est ausi une solution du même problème de Dirichlet, on en déduit que f est nulle. □

Remarque 4.1.1. *Dans cette preuve, l'hypothèse A^c fini sert uniquement à justifier que le maximum de f sur A^c est atteint.*

La méthode de preuve est une technique souvent utilisée en EDP, connue sous le nom de principe du maximum.

Théorème 4.1.5 (Représentation des solutions du problème de Dirichlet). *Soit Q la matrice de transition d'une chaîne de Markov irréductible récurrente. Soit $A \subset \mathcal{X}$ un ensemble non vide et différent de $\mathcal{X}$, et $g : A \longrightarrow \mathbb{R}$ mesurable bornée. Alors la fonction*

$$f : x \longrightarrow \mathbb{E}_x[g(X_T)]$$

avec $T := \inf\{n; X_n \in A\}$ est une solution du problème de Dirichlet. Si de plus A^c est fini, alors c'est l'unique solution.

Proof. L'unicité est une conséquence du Lemme 4.1.4, puique la différence de deux solutions est une solution au problème de Dirichlet avec condition $g = 0$.

Pour l'existence, tout d'abord comme la chaîne est irréductible récurrente, f est bien définie. De plus, par définition on a bien $f = g$ sur A. Il nous reste donc à montrer que f est harmonique sur A^c.

En utilisant la propriété de Markov, on a pour $x \in A^c$

$$\begin{aligned}\mathbb{E}_x[g(X_T)] &= \mathbb{E}_x[g(X_T)\mathbb{1}_{T=1} + g(X_T)\mathbb{1}_{T>1}] \\ &= \sum_{y\in A} g(y)Q(x,y) + \sum_{y\in A^c} Q(x,y)\mathbb{E}_y[g(X_T)] \\ &= \sum_y Q(x,y)f(y).\end{aligned}$$

Donc f est bien harmonique sur A^c. □

On conclut cette section par une caractérisation des chaînes de Markov en terme de martingales :

Théorème 4.1.6. *Un processus $(X_n)_{n\in\mathbb{N}}$ sur $\mathcal{X}$ est une chaîne de Markov de matrice de transition Q ssi pour toute fonction f mesurable bornée à valeurs réelles le processus $\left(f(X_n) - \sum_{k=0}^{n-1} Qf(X_k) - f(X_k)\right)_{n\in\mathbb{N}}$ est une martingale.*

Proof. On a déjà vu, via la décomposition de Doob-Meyer, que si $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov de matrice de transition Q, alors $\left(f(X_n) - \sum_{k=0}^{n-1} Qf(X_k) - f(X_k)\right)_{n\in\mathbb{N}}$ est une martingale.

Pour l'autre implication, on utilise $f(x) = \mathbb{1}_{X_n=x}$. La propriété de martingale nous donne

$$\mathbb{E}[\mathbb{1}_{X_{n+1}} = x|\mathcal{F}_n] = Q(X_n, x).$$

Or $\mathbb{E}[\mathbb{1}_{X_{n+1}} = x|\mathcal{F}_n] = \mathbb{P}[X_{n+1} = x|X_0, ..., X_{n-1}]$, on reconnait donc la propriété définissant les chaînes de Markov de matrice de transition Q. □

4.2 Applications aux temps de sortie

On peut utiliser les fonctions harmoniques pour établir des propriétés des temps d'atteinte de domaines.

Proposition 4.2.1. *Soient A et B deux ensembles disjoints avec $(A\cup B)^c$ non vide et fini, et f la solution au problème de Dirichlet sur $A \cup B$ avec $g = \mathbb{1}_A$. Alors*

$$\mathbb{P}_x(T_A < T_B) = f(x)$$

où $T_A = \inf\{n; X_n \in A\}$ et $T_B = \inf\{n; X_n \in B\}$.

Donc si on sait calculer la solution du problème de Dirichlet, on peut calculer des probabilités de sortie.

Proof. On a $\mathbb{T}_A < T_B \Leftrightarrow g(X_{T_A \wedge T_B}) = 1$ et $T_B < T_A \Leftrightarrow g(X_{T_A \wedge T_B}) = 0$. On conclut via le Théorème 4.1.5 □

Exercice 4.2.1. *Soit X_n la marche aléatoire simple symétrique sur $\mathbb{Z}$, et $a < x < b$. Montrer que*

$$\mathbb{P}_x(T_a > T_b) = \frac{x-a}{b-a}$$

Solution 4.2.1. *On peut facilement montrer que les fonctions affines sur $\mathbb{Z}$ sont harmoniques pour la marche aléatoire symétrique. On vérifie alors que $\frac{x-a}{b-a}$ sur $[a,b]$ est la solution du problème de Dirichlet sur $\mathbb{Z}\backslash(a,b)$ avec $g = \mathbb{1}_{[b+\infty)}$. La conclusion découle alors de la Proposition précédente.*

Exemple 4.2.1 (Probabilité d'extinction du processus de Galton-Watson)**.** *Soit $(N_n)_{n\in\mathbb{N}}$ un processus de Galton Watson, construit à partir de vaiables aléatoires iid $(X_k^n)_{k,n\in\mathbb{N}}$ à valeurs dans N, et avec $N_0 = 1$. On pose*

$$m := \mathbb{E}[X_1^1], \;\; \varphi(t) := \mathbb{E}\left[t^{X_1^1}\right] = \sum_k t^k \mathbb{P}(X_1^1 = k).$$

Soit ζ la plus petite racine positive de l'équation $\varphi(t) = t$. On suppose que $\mathbb{P}(X_1^1 = 0) > 0$ et $\zeta < 1$. Alors $\mathbb{P}(\exists n \text{ t.q. } N_n = 0) = \zeta$.

Pour montrer ce résultat, on va identifier une fonction harmonique pour la chaîne de Markov sur $\mathbb{N}$ définie par le processus de Galton-Watson. On rappelle que sa matrice de transition est donnée par

$$Q(n,\ell) = \mathbb{P}\left(\sum_{k=1}^{n} X_i^1 = \ell\right).$$

Nous allons montrer que $m \to \zeta^m$ est harmonique. En effet,

$$\begin{aligned}\mathbb{E}[\zeta^{N_1} | N_0 = k] &= \mathbb{E}\left[\zeta^{\sum_{1\leqslant 1\leqslant k} X_i^1}\right] \\ &= \prod_{i=1}^{k} \mathbb{E}[\zeta^{X_i^1}] = \varphi(\zeta)^k = \zeta^k.\end{aligned}$$

Comme $\mathbb{P}(X = 0) > 0$, on peut atteindre 0 avec probabilité positive depuis n'importe quel état k. En particulier, on a l'alternative suivante : soit $N_n = 0$ pour n suffisamment grand, soit $N_n \longrightarrow \infty$. En effet, si un état $k \geqslant 1$ est visité une infinité de fois, on finit par aller en zero, mais alors on n'en repartirait plus, ce qui serait absurde.

On peut alors voir $m \to \zeta^m$ comme la solution au problême de Dirichlet sur $\{0, \infty\}$ avec second membre $\mathbb{1}_0$. On a alors, comme $\zeta < 1$,

$$\mathbb{P}(\exists n \text{ t.q. } N_n = 0) = \lim_{n\to\infty} \mathbb{E}[\zeta^{N_n}] = \zeta^{N_0}.$$

Part II

Théorèmes limites

Chapter 5

Topologie de la convergence en loi

Le but de ce chapitre est de comprendre certains aspects topologiques de la convergence en loi, et notamment de caractériser les ensembles compacts des espaces de mesures de probabilité.

On énoncera les théorèmes, et on fera la plupart des preuves, dans le cadre général des espaces polonais. Toutefois, si ce cadre pose des difficultés, on pourra se placer dans le cadre des espaces $\mathbb{R}^d$.

5.1 Quelques rappels

5.1.1 Convergence étroite

Définition 5.1.1. *Soit E un espace polonais (i.e. métrique, séparable complet). Une suite de mesures positives $(\mu_n)_{n\in\mathbb{N}}$ sur E converge étroitement vers une mesure μ si pour toute fonction $f \in C_b(E,\mathbb{R})$ on a*

$$\int f d\mu_n \longrightarrow \int f d\mu.$$

Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ sur E converge en loi vers X si la suite des lois des X_n converge étroitement vers la loi de X.

Ici, $C_b(E,\mathbb{R})$ est l'ensemble des fonctions continues bornées et à valeurs réelles. Dans ce chapitre, on utilisera plutôt la terminologie de l'analyse, et on parlera de convergence étroite.

Proposition 5.1.2. *On suppose (E,d) est un espace polonais. Soit $(\mu_n)_{n\in\mathbb{N}}$ une suite de mesures de probabilité. Les propriétés suivantes sont équivalentes :*

1. *$(\mu_n)_{n\in\mathbb{N}}$ converge étroitement vers μ;*
2. *Pour toute fonction f uniformément continue bornée, $\int f d\mu_n \longrightarrow \int f d\mu$;*
3. *Pour toute fonction f lispchitz et bornée, on a $\int f d\mu_n \longrightarrow \int f d\mu$;*
4. *Pour tout fermé F, on a $\mu(F) \geqslant \limsup \mu_n(F)$;*
5. *Pour tout ouvert O, on a $\mu(O) \leqslant \liminf \mu_n(O)$;*
6. *Pour tout borélien A avec $\mu(\partial A) = 0$, on a $\lim \mu_n(A) = \mu(A)$;*

7. *Pour toute fonction f mesurable, continue μ-presque partout et bornée, on a $\int f d\mu_n \longrightarrow \int f d\mu$.*

Proof. Les implications $1 \Rightarrow 2 \Rightarrow 3$ et $7 \Rightarrow 1$, et l'équivalence $4 \Leftrightarrow 5$, sont immédiates.

Pour montrer que $3 \Rightarrow 4$, pour F un ensemble fermé, on introduit la fonction $f_K(x) := \max(0, 1 - Kd(x, F))$. Cette fonction est K-lipchitz et bornée, et lorsque $K \longrightarrow +\infty$, f_K converge de manière monotone vers $\mathbb{1}_F$. On a donc $\mu_n(F) \leqslant \int f_K d\mu_n$ pour tout n, et donc

$$\limsup_n \mu_n(F) \leqslant \lim_K \limsup_n \int f_K d\mu_N = \lim_K \int f_K d\mu = \mu(F)$$

où on a utilisé le théorème de convergence dominé appliquée aux fonctions $f_K \leqslant 1$. On a donc bien $3 \Rightarrow 4$.

Montrons maintenant que 4 et 5 (qui sont équivalentes) impliquent 6. Si $\mu(\partial A) = 0$, on a alors $\mu(A) = \mu(A^\circ) = \mu(\bar{A})$, et donc

$$\mu(A) = \mu(A^\circ) \leqslant \liminf \mu_n(A^\circ) \leqslant \liminf \mu_n(A) \leqslant \limsup \mu_n(A)$$

$$\leqslant \limsup \mu_n(\bar{A}) \leqslant \mu(\bar{A}) = \mu(A),$$

ce qui permet de conclure.

Montrons enfin que $6 \Rightarrow 7$. Soit f une fonction mesurable, bornée et continue μ-presque partout. Quitte à décomposer f en $f_+ - f_-$, on peut supposer que f est positive. Grâce au théorème de Fubini, on a

$$\int f d\mu = \int \mu(dx) \int_0^\infty \mathbb{1}_{[O,f(x)]}(y) dy = \int_0^\infty \mu(\{f \geqslant y\}) dy.$$

Soit D l'ensemble des points de discontinuité de f. Pour tout y, si $A_y = \{x; f(x) \geqslant y\}$, alors $\partial A_y \subset D \cup \{f = y\}$. En effet, si $x \in \bar{A}_y$, alors il existe une suite telle que $x_n \to x$ avec $f(x_n) > y$ pour tout n. Alors, si x n'est pas un point de discontinuité de f, et n'est pas dans A_y, on a nécessairement $f(x) = y$. On souhaite montrer que pour Lebesgue-presque tout y, on a $\mu(\partial A_y) = 0$, et pour ce il suffit de montrer que pour Lebesgue-presque tout y, on a $\mu(\{f = y\}) = 0$.

Or $\{y \geqslant 0; \mu(\{f = y\}) > 0\} = \cup_{k \in \mathbb{N}} \{y \geqslant 0; \mu(\{f = y\}) \geqslant 1/k\}$. Or comme les $\{f = y\}$ sont deux à deux disjoints, $\{y \geqslant 0; \mu(\{f = y\}) \geqslant 1/k\}$ est de cardinal inférieur à k. Donc $\{y \geqslant 0; \mu(\{f = y\}) > 0\}$ est une réunion dénombrable d'ensemble finis, donc dénombrable, et donc de mesure de Lebesgue nulle.

Donc, en utilisant 6, pour Lebesgue-presque tout $y \geqslant 0$, on a $\mu_n(f \geqslant y) \longrightarrow \mu(f \geqslant y)$. Alors, par convergence dominée,

$$\int f d\mu_n = \int_0^{||f||_\infty} \mu_n(f \geqslant y) dy \longrightarrow \int_0^{||f||_\infty} \mu(f \geqslant y) dy = \int f d\mu$$

ce qui conclut la preuve.

□

Exercice 5.1.1. *Soit $(\mu_n)_{n \in \mathbb{N}}$ et μ des mesures de probabilités sur $\mathbb{R}^k$, et H un ensemble de fonctions mesurables bornées, dont l'adhérence pour la topologie de la convergence uniforme contient $C_c(\mathbb{R}^k, \mathbb{R})$. Montrer que si $\int f d\mu_n \longrightarrow \int f d\mu$ pour toute fonction f dans H, alors μ_n converge étroitement vers μ*

Ce résultat devient faux si μ n'est pas une mesure de probabilité, ou si on n'est pas en dimension finie (la preuve utilisera la compacité des boules fermées).

Solution 5.1.1. *Montrons d'abord que le résultat est vrai pour $H = C_c(\mathbb{R}^k, \mathbb{R})$. On considère une fonction cutoff ξ_r qui est continue, à valeurs dans $[0,1]$, égale à 1 sur $B(0,r)$ et nulle sur $B(0,r+1)^c$. On a alors*

$$\limsup \mu_n(B(0,r+1)^c) \leqslant 1 - \liminf \int \xi_r d\mu_n \leqslant \mu(B(0,r)^c.$$

Soit f continue bornée, $\varepsilon > 0$ et $r > 1$ tel que $\mu(B(0,r-1)^c) \leqslant \frac{\varepsilon}{4||f||_\infty}$. Par hypothèse, il existe $g \in H$ telle que $||f\xi_r - g||_\infty \leqslant \varepsilon/4$. Alors

$$\begin{aligned}\left|\int f d\mu - \int f d\mu_n\right| &\leqslant \left|\int f\xi_r d\mu_n - \int f d\mu_n\right| + \left|\int f\xi_r d\mu_n - \int g d\mu_n\right| \\ &\quad + \left|\int g d\mu_n - \int g d\mu\right| + \left|\int g d\mu - \int f\xi_r d\mu\right| + \left|\int f\xi_r d\mu - \int f d\mu\right| \\ &\leqslant ||f||_\infty \mu_n(B(0,r-1)^c + ||f||_\infty \mu(B(0,r-1)^c + 2||f\xi_r - g||_\infty + \left|\int g d\mu_n - \int g d\mu\right| \\ &\leqslant \frac{3\varepsilon}{4} + ||f||_\infty \mu_n(B(0,r-1)^c + \left|\int g d\mu_n - \int g d\mu\right|.\end{aligned}$$

On a donc

$$\begin{aligned}\limsup_n \left|\int f d\mu - \int f d\mu_n\right| &\leqslant \limsup_n ||f||_\infty \mu_n(B(0,r-1)^c + \limsup_n \left|\int g d\mu_n - \int g d\mu\right| \\ &\leqslant \varepsilon.\end{aligned}$$

Comme ε était arbitraire, ceci conclut la preuve.

Pour conclure, on rappelle la caractérisation suivante de la convergence en loi sur $\mathbb{R}$:

Proposition 5.1.3. *Une suite de mesures de probabilité (μ_n) sur $\mathbb{R}$ converge étroitement vers une mesure de probabilité μ ssi les fonctions de répartition F_{μ_n} satisfont*

$$F_{\mu_n}(x) \longrightarrow F_\mu(x)$$

en tout point x où F_μ est continue.

A noter que l'ensemble des points de discontinuité d'une fonction de répartition est au plus dénombrable.

NB. *L'hypothèse que toutes les mesures sont des mesures de probabilité est essentielle ici : la suite des fonctions de répartition des mesures δ_n converge simplement vers la fonction nulle, mais il n'y a pas de convergence en loi.*

5.1.2 Théorème de Lévy

On rappelle ici le théorème de continuité de Lévy, qui lie convergence en loi et convergence des fonctions caractéristiques.

Définition 5.1.4. *La fonction caractéristique d'une mesure positive finie μ sur $\mathbb{R}^d$ est*

$$\varphi_\mu(x) := \int_y \exp(i\langle x, y\rangle) d\mu.$$

Théorème 5.1.5 (Théorème de Lévy)**.** *Soit $(\mu_n)_{n\in\mathbb{N}}$ une suite de mesures de probabilités sur $\mathbb{R}^d$. Alors la suite $(\mu_n)_{n\in\mathbb{N}}$ converge étroitement vers une mesure de probabilité μ ssi la suite $(\varphi_{\mu_n})_{n\in\mathbb{N}}$ converge simplement vers φ_μ.*

5.2 Distance de Lévy-Prokhorov

Le but de cette section est de montrer que l'espace des mesures de probabilité, muni de la topologie de la convergence étroite, peut être équippé d'une distance qui en fait un espace polonais.

Définition 5.2.1 (Distance de Lévy-Prokhorov)**.** *Soit (E, d) un espace polonais. La distance de Lévy-Prokhorov (assoicée à d) sur $\mathcal{P}(E)$ est*

$$d_{LP}(\mu, \nu) := \inf\{r > 0; \mu(F) \leqslant \nu(F^r) + r \quad \forall F \subset E \text{ fermé } \},$$

où $F^r := \{x \in E; d(x, F) < r\}$ est le r-voisinage ouvert de F.

Cette distance est utile pour des considérations théoriques, et notamment topologique, mais est rarement utilisée en pratique car elle est difficile à calculer. Nous verrons dans la Section 5.4 d'autres distances, souvent plus pratiques pour l'étude de problèmes concrets.

Montrons que d_{LP} est bien une distance. Elle est trivialement positive, et prend des valeurs finies, car elle est trivialement majorée par 1. Commençons par montrer qu'elle est symétrique. Soit μ et ν donnée, et $r > d_{LP}(\mu, \nu)$. Pour tout F fermé, on a

$$\mu(F) \leqslant \nu(F^r) + r.$$

Soit F un fermé donné. Alors $F' = E\backslash F^r$ est fermé, et $F \subset E\backslash(F')^r$. Donc

$$\nu(F) \leqslant 1 - \nu((F')^r) \leqslant 1 - \mu(F') + r = \mu(F) + r.$$

Comme F est quelconque, on en déduit $d_{LP}(\nu, \mu) \leqslant r$, et on conclut en faisant tendre r vers $d_{LP}(\mu, \nu)$.

Montrons maintenant qu $d_{LP}(\mu, \nu) = 0 \Leftrightarrow \mu = \nu$. L'impliction de la droite vers la gauche est immédiate. Pour l'autre implication, supposons que $d_{LP}(\mu, \nu) = 0$, et soit F un fermé arbitraire. On a $\mu(F) \leqslant \nu(F^\varepsilon) + \varepsilon$ pour tout $\varepsilon > 0$. Mais comme $F = \cap_\varepsilon F^\varepsilon$, en faisant tendre ε vers 0 on obtient $\mu(F) \leqslant \nu(F)$ pour tout ensemble fermé. Par symétrie de d_{LP}, on peut donc conclure que $\mu = \nu$.

Enfin, d_{LP} vérifie l'inégalité triangulaire. En effet, si μ_1, μ_2 et μ_3 sont des mesures de probailité et r, r' sont tels que pour tout fermé F on ait

$$\mu_1(F) \leqslant \mu_2(F^r) + r; \quad \mu_2(F) \leqslant \mu_3(F^{r'}) + r',$$

alors

$$\mu_1(F) \leqslant \mu_2(\bar{F^r}) + r \leqslant \mu_3((\bar{F^r})^{r'}) + r + r' \leqslant \mu_3(F^{r+r'}) + r + r',$$

et donc $d_{LP}(\mu_1, \mu_3) \leqslant r + r'$. On conclut en prenant l'inf sur les r et r' admissibles.

Proposition 5.2.2. *La distance d_{LP} induit la topologie de la convergence étroite.*

La preuve utilisera le lemme suivant :

Lemme 5.2.3. *Soit (E,d) un espace métrique séparable, μ une mesure borelienne finie sur Eet D un sous-ensemble dénombrable dense de E. Pour tout $\varepsilon > 0$, il existe une famille dénombrable de points x_i de D et une famille $(\delta_i)_{i\in\mathbb{N}} \in (0,\delta)^{\mathbb{N}}$ tels que*

$$\bigcup_{i\in\mathbb{N}} B(x_i,\delta_i) = E; \quad \mu(\partial B(x_i,\delta_i)) = 0.$$

Dans ce lemme, ça ne change rien si on considère des boules ouvertes ou fermées. Ce lemme va nous permettre d'approximer des ensembles arbitraires avec des unions de boules arbitrairement petites, dans l'esprit de l'approximation de sous-ensembles de $\mathbb{R}$ par des unions d'intervalles. La condition de frontière de mesure nulle va nous permettre d'utiliser la propriété 6 de la Proposition 5.1.2.

On peut aussi considérer une variante de ce lemme, où le sensembles ne sont plus des boules, mais deviennent disjoints :

Lemme 5.2.4. *Soit (E,d) un espace métrique séparable et μ une mesure borelienne finie sur E, et $\delta > 0$. Il existe une partition de E en des ensembles disjoints A_i avec* $\operatorname{diam}(A_i) < \delta$ *et* $\mu(\partial A) = 0$.

Preuve de la Proposition 5.2.2. Tout d'abord, montrons que si $d_{LP}(\mu_n,\mu) \longrightarrow 0$, alors μ_n converge étroitement vers μ. On a une suite ε_n qui décroît vers 0 telle que pour tout n et tout fermé F on a

$$\mu(F) \leqslant \mu_n(F^{\varepsilon_n}) + \varepsilon_n; \quad \mu_n(F) \leqslant \mu(F^{\varepsilon_n}) + \varepsilon_n.$$

Donc

$$\limsup \mu_n(F) \leqslant \limsup \mu(F^{\varepsilon_n}) + \varepsilon_n = \mu(\bar{F}) = \mu(F).$$

La Proposition 5.1.2 permet de conclure.

Montrons maintenant que si μ_n converge étroitement vers μ, alors $d_{LP}(\mu_n,\mu)$ tend vers 0. Soit $\varepsilon > 0$ et $\delta < \varepsilon/3$. En utilisant le lemme 5.2.3, on se donne une collection de boules $B_i = B(x_i,\delta_i)$ avec $\delta_i < \delta/2$, $\cup_i B_i = E$ et $\mu(\partial B_i) = 0$ pour tout i. Soit k tel que

$$\mu\left(\cup_{i=1}^k B_i\right) \geqslant 1-\delta.$$

On considère la collection finie d'ensembles

$$\mathcal{A} := \{\cup_{j\in J} B_j; \quad J \subset \{1,...,k\}\}.$$

Comme tout élément de $\mathcal{A}$ ets une réunion finie d'ensembles dont les frontières sont de mesure nulle, pour tout $A \in \mathcal{A}$ on a $\mu(\partial A) = 0$. On a donc $\mu_n(A) \longrightarrow \mu(A)$ pour tout $A \in \mathcal{A}$. Soit N suffisemment grand pour que

$$|\mu_n(A) - \mu(A)| < \delta \quad \forall n \geqslant N,\ A \in \mathcal{A}.$$

En particulier, pour tout $n \geqslant N$,

$$\mu_n\left(\cup_{j\leqslant k} B_j\right) \geqslant \mu\left(\cup_{j\leqslant k} B_j\right) - \delta \geqslant 1-2\delta.$$

Soit B un borelien arbitraire, qu'on souhaite approximer avec

$$A = \cup_{j \leqslant k, B_j \cap B \neq \varnothing} B_j.$$

Alors $A \subset B^\delta$ car le diamètre des B_j est inférieur à δ, et $B \subset A \cup (\cup_{j \leqslant k} B_j)^c$. On a donc pou $n \geqslant N$

$$\begin{aligned}
\mu(B) &\leqslant \mu(A) + \mu\left((\cup_{j \leqslant k} B_j)^c\right) \\
&\leqslant \mu(A) + \delta \\
&\leqslant \mu_n(A) + 2\delta \\
&\leqslant \mu_n(B^\delta) + 2\delta \\
&\leqslant \mu_n(B^\varepsilon) + \varepsilon.
\end{aligned}$$

On a donc $d_{LP}(\mu, \mu_n) \leqslant \varepsilon$ pour tout $n \geqslant N$. Donc on a bien $d_{LP}(\mu, \mu_n) \longrightarrow 0$. □

Preuve du Lemme 5.2.3. Soit D un ensemble dénombrable dense, et $x \in D$. On pose $S(x,r) := \{y, d(x,y) = r\}$. On a $\partial B(x,r) \subset S(x,r)$.

Soit $A_k := \{r \in (\delta/2, \delta); \mu(S(x,r)) \geqslant 1/k\}$. Comme les ensembles $S(x,r)$ sont deux à deux disjoints, il y a au plus $k\mu(E)$ éléments dans A_k. Donc $\{r \in (\delta/2, \delta); \mu(S(x,r)) > 0\} = \bigcup_k A_k$ est au plus dénombrable, car réunion dénombrable d'ensembles finis.

Il existe donc toujours un $r \in (\delta/2, \delta)$ tel que $\mu(\partial B(x,r)) = 0$. Comme D est dense, en prenant pour chaque $x \in D$ une telle boule, l'union de ces boules recouvre E, ce qui permet de conclure la preuve. □

Preuve du Lemme 5.2.4. On considère une collection de boules fermées B_i de rayon $< \delta/2$ donnée par le Lemme 5.2.3, via une collection de points dénombrables dense $(x_i)_{i \in \mathbb{N}}$. On construit ensuite itérativement la collection $(A_i)_{i \in \mathbb{N}}$ avec

$$A_0 = B_0; \quad A_{n+1} := B_{n+1} \setminus (\cup_{i \leqslant n} B_i).$$

Les A_j forment bien une partition de E, et ils sont tous de diamètre inférieur à δ. De plus,

$$\partial A_n \subset \cup_{j \leqslant n} \partial B_j$$

et donc on a bien $\mu(\partial A_n) \leqslant \sum_{j \leqslant n} \mu(\partial B_j) = 0$. □

Proposition 5.2.5. *Si (E,d) est un espace polonais, alors l'espace métrique $(\mathbb{P}(E), d_{LP})$ est séparable.*

Preuve de la Proposition 5.2.5. Soit D un ensemble dénombrable dense de E. Nous allons montrer qu'on peut approximer une mesure donnée par des mesures de la forme

$$\sum_j \alpha_j \delta_{x_j}, \quad \sum_j \alpha_j = 1, \quad \alpha_j \in \mathbb{Q}_+ \forall j. \tag{5.1}$$

L'ensemble des mesures de cette forme est bien dénombrable.

Soit A_j la collection d'ensembles donnée par le Lemme 5.2.4 avec $\delta < 1/(2n)$. On a alors $\sum \mu(A_j) = 1$. On peut considérer dans la suite uniquement les A_j d'intérieur non-vide, car sinon $A_j \subset \partial A_j$, qui est alors de mesure nulle. Soit $n \in \mathbb{N}^*$ fixé. On peut trouver une collection de rationnels positifs β_j tels que $\sum \beta_j = 1$ et $\sum |\beta_j - \mu(A_j)| \leqslant \frac{1}{n}$. On peut considérer dans la suite uniquement les A_j d'intérieur

non-vide, car sinon $A_j \subset \partial A_j$, qui est alors de mesure nulle. On choisit alors pour tout j un $x_j \in D$, et on définit

$$\mu_n := \sum_j \beta_j \delta_{x_j}.$$

On note que μ_n est bien une mesure de probabilité de la forme (5.1).

On a ensuite pour tout fermé F l'inclusion $\cup_{x_j \in F} B_j \subset F^{2\delta}$, et donc

$$\mu_n(F) = \sum_{x_j \in F} \beta_j \leqslant \frac{1}{n} + \sum_j \mu(A_j) \leqslant \mu(F^{2\delta}) + \frac{1}{n}$$

et donc, comme $2\delta < 1/n$, on a bien $d_{LP}(\mu_n, \mu) \leqslant 1/n$.

□

Théorème 5.2.6. *Si (E, d) est un espace polonais, alors l'espace métrique $(\mathbb{P}(E), d_{LP})$ est lui aussi polonais.*

La preuve de ce théorème (pour laquelle il suffit de montrer la complétude) sera faite plus tard.

5.3 Tension

L'objectif de cette section est d'étudier plus en détails la topologie de des espaces de mesures de probabilité, et notamment d'en caractériser les sous-ensembles séqentiellement compacts (pour la convergence étroite).

5.3.1 Cas réel

En dimension 1, on peut utiliser comme outil les fonctions de répartition $F_\mu(x) := \mu((-\infty, x])$ et la Proposition 5.1.3.

Proposition 5.3.1 (Lemme de Helly-Braye)**.** *Soit (f_n) une suite de fonctions croissantes de $\mathbb{R}$, à valeurs dans l'intervalle $[0, 1]$. Alors il existe une sous-suite qui converge simplement. En particulier, de toute suites de fonctions de répartition de lois de probailité, on peut extraire une sous-suite convergente.*

Proof. Comme un produit dénombrable d'espaces métriques séquentiellement compacts est séquentiellement compact, il existe une sous suite convergeante en tous points de $\mathbb{Q}$, que nous noterons f_n. Notons f la limite de cette sous-suite convergeante. On peut définir

$$f_-(x) := \sup\{f(y), y \in \mathbb{Q} \cap (-\infty, x]\}, \quad f_+(x) := \inf\{f(y), y \in \mathbb{Q} \cap [x, +\infty)\}.$$

Par monotonie, l'ensemble D des points auxquels $f_-(x) < f_+(x)$ est au plus dénombrable. Notons encore $f(x)$ leurs valeurs communes sur $\mathbb{R}\backslash D$.

Nous allons maintenant montrer que $f_n(x) \longrightarrow f(x)$ sur $\mathbb{R}\backslash D$. Soit $x \in \mathbb{R}\backslash D$. Il existe $y < x$ et $z > x$ rationnels tels que

$$f(y) \geqslant f(x) - \varepsilon; \quad f(z) \leqslant f(x) + \varepsilon$$

. Par convergence simple des suites $(f_n(y))$ et $(f_n(z))$, pour tout n suffisamment grand on a alors

$$f_n(y) \geqslant f(x) - 2\varepsilon; \quad f(z) \leqslant f(x) + 2\varepsilon$$

. Par monotonie des f_n, on a alors pour n suffisamment grand

$$f(x) - 2\varepsilon \leqslant f_n(x) \leqslant f(x) - 2\varepsilon$$

et donc on a bien convergence simple sur $\mathbb{R}\backslash D$.

Enfin, comme D est dénombrable, on peut extraire une sous-suite qui converge aussi simplement sur D, ce qui conclut la preuve. □

Corollaire 5.3.2. *Soit (μ_n) une séquence de mesures de probabilités sur $\mathbb{R}$. Il existe une sous-suite convergeant vaguement vers une mesure positive μ sur $\mathbb{R}$, avec $\mu(\mathbb{R}) \leqslant 1$, c'est à dire que pour toute fonction f continue* à support compact*,*

$$\int f d\mu_n \longrightarrow \int f d\mu.$$

NB. *La limite d'une suite de fonctions de répartition obtenue via le lemme de Helly-Braye n'est pas forcément elle-même une fonction de répartition. Par exemple, la suite des fonctions de répartitions des mesures de Dirac δ_n est $\mathbb{1}_{[n,+\infty)}$, et leur limite est la fonction nulle.*

En revanche, cette limite est toujours une fonction croissante.

En d'autres termes, l'ensemble des mesures positives sur $\mathbb{R}$ de masse inférieure à 1 est séquentiellement compact pour la convergence vague. Par contre, l'ensemble des mesures de probabilité ne l'est pas. De plus, comme la limite peut avoir une masse totale strictement inférieure à 1, il n'y a pas nécessairement convergence étroite.

Pour pouvoir garantir que la limite (après extraction) est une mesure de probabilité, il faut et suffit que la limite des fonctions caractéristiques tende vers 1 en $+\infty$ et 0 en $-\infty$, i.e.

$$\lim_{x\to+\infty} \lim_n F_{\mu_n}(x) = 1; \quad \lim_{x\to-\infty} \lim_n F_{\mu_n}(x) = 0 \tag{5.2}$$

Ceci mène à la définition suivante :

Définition 5.3.3 (Tension, cas réel)**.** *Une famille Γ de mesures de probabilité sur $\mathbb{R}$ est dite tendue si pour tout $\varepsilon > 0$, il existe $r > 0$ tel que*

$$\inf_{\mu\in\Gamma} F_\mu(r) - F_\mu(-r) \geqslant 1 - \varepsilon;$$

ou, de manière équivalente,

$$\sup_{\mu\in\Gamma} \mu([-r, r]^c) \leqslant \varepsilon.$$

Exercice 5.3.1. *Montrer que une suite de mesures de probabilité est tendue ssi on a* (5.2).

Solution 5.3.1. *La notion de tension est équivalente à*

$$\lim_r \inf_n F_{\mu_n}(r) - F_{\mu_n}(-r) = 1.$$

Comme de plus les fonctions de répartition sont croissantes et prennent des valeurs entre 0 *et* 1*, on a*

$$\lim_r \inf_n F_{\mu_n}(r) = 1; \quad \lim_r \sup_n F_{\mu_n}(-r) = 0.$$

On en déduit que

$$1 \geqslant \lim_r \lim_n F_{\mu_n}(r) \geqslant \lim_r \inf_n F_{\mu_n}(r) = 1$$

et la même chose fonctionne pour le comportement en $-\infty$*.*

Réciproquement, si on a (5.2)*, pour tout* $\varepsilon > 0$*, il existe* $R > 0$ *et* $N = N(R)$ *tels que pour tout* $n \geqslant N$ *et* $r \geqslant R$ *on ait*

$$F_{\mu_n}(r) \geqslant 1 - \varepsilon/2.$$

Comme de plus pour n *fixé la fonction de répartition tend vers* 1 *à l'infini, quitte à augmenter* r *on peut avoir la même borne pour* $n < N$*, et donc trouver* r *tel que*

$$\inf_n F_{\mu_n}(r) \geqslant 1 - \varepsilon/2.$$

. De manière symétrique, on a le même comportment en $-\infty$*, et donc il existe* r *suffisemment grand tel que*

$$\inf_n F_{\mu_n}(r) - F_{\mu_n}(-r) \geqslant 1 - \varepsilon.$$

En combinant le lemme de Helly-Braye et ce résultat, on obtient :

Théorème 5.3.4 (Théorème de Prokhorov, cas réel)**.** *Si une suite de mesures de probabilité sur* $\mathbb{R}$ *est tendue, alors elle admet une sous-suite qui converge étroitement.*

La réciproque est aussi vraie en un certain sens, qu'on verra dans la suite.

5.3.2 Cas compact

Dans le cas d'un espace compact, il n'y a pas besoin de parler de tension, et l'espace des mesures de probabilité sera lui-même compact. Pour démontrer cela, on s'appuiera sur des résultats d'analyse fonctionnelle.

Théorème 5.3.5 (Théorème de représentation de Riesz-Markov)**.** *Soit* (E, d) *un espace métrique compact, et* φ *une forme linéaire positive sur* $(C(E, \mathbb{R}), ||\cdot||_\infty)$ *telle que* $\varphi(1) < \infty$*. Alors il existe une unique mesure positive finie borelienne* μ *sur* E *telle que*

$$\varphi(f) = \int f d\mu \quad \forall f \in C(E, \mathbb{R}).$$

De plus, $\mu(E) = \varphi(1) = ||\varphi||$*.*

A noter qu'une telle forme linéaire est nécessairement continue sur $(C(E, \mathbb{R}), ||\cdot||_\infty)$.

Corollaire 5.3.6. *Soit (E,d) un espace métrique compact. Alors la topologie de la convergence étroite de mesures finies coincide avec la topologie faible-* sur l'ensemble des formes linéaires positives.*

Théorème 5.3.7 (Théorème de Banach-Alaoglu-Bourbaki)**.** *Soit E un espace vectoriel normé séparable. La boule unité de son dual (l'ensemble des formes linéaires continues) est compact pour la topologie faible-*.*

Dans le cadre compact, l'ensemble des mesures de probabilité est fermé dans $C(E,\mathbb{R})^*$, et on en déduit le résultat suivant :

Corollaire 5.3.8 (Théorème de Prokhorov, cas des espaces compacts)**.** *Si (E,d) est un espace compact, alors $\mathcal{P}(E)$ est compact pour la topologie de la convergence étroite.*

Corollaire 5.3.9. *Si une suite de variables aléatoires est uniformément bornée (c'est à dire que il existe $M>0$ tel que pour tout n on ait $|X_n| \leqslant M$ p.s.), alors on peut en extraire une sous-suite qui converge en loi.*

5.3.3 Cas général

Le but maintenant sera de voir le théorème de Prokhorov dans un cadre plus général. La généralisation naturelle de la notion de tension pour un espace topologique est :

Définition 5.3.10 (Tension)**.** *Un ensemble $\Gamma \subset \mathcal{P}(E)$ est tendu si pour tout $\varepsilon > 0$ il existe un compact $K_\varepsilon \subset E$ tel que*

$$\sup_{\mu\in\Gamma} \mu(E\backslash K_\varepsilon) \leqslant \varepsilon.$$

Exercice 5.3.2. *1. Soit μ une mesure de probabilité sur un espace polonais. Montrer que pour tout $k \geqslant 1$, il existe un nombre fini N_k de boules de rayon $1/k$ dont l'union est de mesure supérieure à $1-2^{-k}\varepsilon$.*

2. En déduire que le singleton $\{\mu\}$ est tendu.

3. Montrer qu'un ensemble fini de mesures de probabilité sur un espace polonais est tendu.

Solution 5.3.2. *1. Comme E est polonais, on peut considérer une suite dénombrable dense $(z_n)_{n\in\mathbb{N}^*}$. Comme pour tout $k \geqslant 1$, on a $\cup_n \bar{B}(z_n,1/k) = E$, où $\bar{B}$ est une boule fermée de centre et rayon donnés, on peut toujours trouver N_k tel que*

$$\mu\left(\cup_{n\leqslant N_k} \bar{B}(z_n,1/k)\right) \geqslant 1-\frac{\varepsilon}{2^k}.$$

2. On considére $A = \cap_k \left(\cup_{n\leqslant N_k} \bar{B}(z_n,1/k)\right)$. On a

$$\mu(A^c) \leqslant \sum_k \mu\left(\left(\cup_{n\leqslant N_k} \bar{B}(z_n,1/k)\right)^c\right) \leqslant \varepsilon.$$

De plus, A est recouvrable par un nombre fini de boules de rayon arbitrairement petit, donc A est précompact. Comme de plus A est fermé, cet ensemble est compact. On a donc un ensemble compact de masse arbitrairement proche de 1, donc $\{\mu\}$ est tendu.

3. *Soit $\{\mu_1, ..., \mu_n\}$ un ensemble fini de mesures. Pour tout $\varepsilon > 0$, il existe des compacts K_i tels que $\mu_i(K_i) \geqslant 1 - \varepsilon$. Alors $\cup_i K_i$ est un compact, et tous les μ_i lui donnent une masse supérieure à $1-\varepsilon$. L'ensemble $\{\mu_1, ..., \mu_n\}$ est donc tendu.*

Notre but va être maintenant de prouver la caractérisation suivante:

Théorème 5.3.11 (Théorème de Prokhorov)**.** *Soit E un espace polonais. Une famille de mesures de probabilité est relativement compacte dans $\mathbb{P}(E)$ ssi elle est tendue.*

Proof. Montrons d'abord qu'une famille tendue est relativement compacte. Comme on a déjà montré que $\mathcal{P}(E)$ est un espace métrique séparable, on peut utiliser la caractérisation séquentielle de la compacité. Soit $(\mu_n)_{n\in\mathbb{N}}$ une suite d'éléments d'une famille tendue. Par définition de la tension, pour tout $p \geqslant 1$, il existe un compact K_p tel que

$$\sup_n \mu_n((K_p)^c) < \frac{1}{p}.$$

Sans perdre de généralité, on supposera les K_p emboîtés, c'est à dire $K_p \subset K_{p+1}$. On peut considérer des restrictions à ces domaines K_p, c'est à dire les mesures de probabilités μ_n^p définies par

$$\mu_n^p(A) := \frac{\mu_n(A \cap K_p)}{\mu_n(K_p)}.$$

Ces mesures représentent bien sûr des lois conditionnelles, avec contrainte de rester dans K_p.

Comme ces mesures sont à support dans un compact, on sait qu'elles sont séquentiellement compactes : pour tout p, on peut extraire une sous-suite de μ_n^p qui converge étroitement vers une limite ν^p, elle ausi à support dans K_p. Par extraction diagonale, on peut alors trouver une sous suite $\varphi(n)$ telle que

$$\mu^p_{\varphi(n)} \xrightarrow{etr.} \nu^p$$

pour tout p. De plus, comme les $\mu_n(K_p)$ sont à valeurs dans $[1 - p^{-1}, 1]$, on peut aussi sans perdre de généralité (quitte à extraire de nouveau) supposer que pour tout p,

$$\mu_{\varphi(n)}(K_p) \longrightarrow m_p \in [1 - p^{-1}, 1]$$

pour tout $p \geqslant 1$. On pose $\mu^p = m_p \nu^p$. Comme les K_p sont emboîtés, on a $\mu_n((A \cap K_p) \cap K_{p+1}) = \mu_n(A \cap K_p)$, et donc $\mu^{p+1}(A \cap K_p) = \mu^p(A)$. De plus, $\mu^p(A) \leqslant \mu^{p+1}(A)$ pour tout A. On peut donc étendre μ^p à l'espace tout entier par $\mu^p(A) = \mu^p(A \cap K_p)$, et définir

$$\mu(A) = \lim_p \mu^p(A)$$

par convergence monotone. Comme $m_p \longrightarrow 1$, cette mesure μ est une mesure de probabilité. Montrons que $\mu_{\varphi(n)}$ converge étoitement vers μ. On a pour tout ouver O et pour tout $p \geqslant 1$

$$\liminf_n \mu_{\varphi(n)}(O) \geqslant \liminf_n \mu_{\varphi(n)}(O \cap K_p) = \liminf \mu^p_{\varphi(n)}(O)\mu_{\varphi(n)}(K_p).$$

Comme $O \cap K_p$ est un ouvert de K_p, on peut appliquer la convergence étroite des restrictions et avoir

$$\liminf_n \mu_{\varphi(n)}(O) \geqslant \mu^p(O).$$

En faisant tendre p vers l'infini, on en déduit $\liminf_n \mu_{\varphi(n)}(O) \geqslant \mu(O)$, ce qui donne la convergence étroite via la Proposition 5.1.2.

Démontrons maintenant la réciproque. Soit Γ une famille relativement compacte. Quitte à prendre son adhérence, on peut supposer qu'elle est compacte. Ceci se fait sans perte de généralité, puisque si l'adhérence d'une famille est tendue, la famille est aussi tendue. Soit $(x_n)_{n\in\mathbb{N}}$ une famille dénombrable dense et $O_{k,n} := \bigcup_{j\leqslant n} B(x_i, 1/k)$.

Pour tout k et tout $\varepsilon > 0$, il existe N tel que pour toute mesure $\mu \in \Gamma$ on ait $\mu(O_{k,N}) > 1-\varepsilon$. En effet, si ce n'était pas le cas, on pourrait trouver une suite μ_n qui converge vers une mesure $\mu \in \Gamma$ par compacité, et telle que $\mu_n(O_{k,n}) \leqslant 1-\varepsilon$ pour tout n. Mais alors, par monotone de la suite $O_{k,n}$ en n, on aurait pour tout m

$$\mu(O_{k,m}) \leqslant \liminf_n \mu_n(O_{k,m}) \leqslant \liminf_n \mu_n(O_{k,n}) \leqslant 1-\varepsilon.$$

Mais comme $\bigcup_n O_{k,n} = E$, on a aussi $\lim_n \mu(O_{k,m}) = 1$, et donc il y aurait contradiction.

Soit $\varepsilon > 0$, et N_k un entier tel que $\mu(O_{k,N}) > 1-\varepsilon/2^k$ pour tout $\mu \in \Gamma$. Posons $K := \bigcap_{k\geqslant 1} \bar{O}_{k,N_k}$. Alors K est précompact, et fermé, donc compact car E est complet. De plus,

$$\mu(E/K) \leqslant \sum_k \mu(O^c_{k,N_k}) \leqslant \sum_{k\geqslant 1} \frac{\varepsilon}{2^k} = \varepsilon.$$

Comme ε est arbitraire, Γ est donc bien une famille tendue. □

Comme conséquence du théorème de Prokhorov, on peut utiliser le raisonnement suivant pour démontrer une convergence en loi : montrer que la suite des lois est tendue, puis qu'il n'y a qu'une seule limite possible. C'est l'analogue du raisonnement classique en analyse qu'une suite bornée de réels qui n'a qu'une seule valeur d'adhérence converge.

Corollaire 5.3.12. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$. On suppose que pour tout $\varepsilon > 0$ il existe $R > 0$ tel que*

$$\sup_n \mathbb{P}(|X_n| > R) \leqslant \varepsilon.$$

Alors il existe une extraction φ et une variable aléatoire X à valeurs dans $\mathbb{R}^d$ telles que $X_n \xrightarrow{loi} X$.

Démontrons maintenant le Théorème 5.2.6 (qui n'a pas été utilisé dans la démonstration du théorème de Prokhorov).

Preuve du Théorème 5.2.6. Soit (μ_n) une suite de Cauchy pour d_{LP}. Nous allons montrer qu'elle est tendue, et le théorème de Prokhorov nous permettra ensuite de conclure qu'on peut en extraire une sous-suite convergente, ce qui suffira pour déduire la complétude.

Par propriété de Cauchy, il existe une extraction φ telle que

$$d_{LP}(\mu_m, \mu_{\varphi}(n)) \leqslant \frac{\varepsilon}{2^{n+1}} \ \forall m \geqslant \varphi(n).$$

Comme l'ensemble fini $\{\mu_1, ..., \mu_{\varphi(n)}\}$ est tendu, on peut trouver un compact K_n tel que

$$\max_{i \leqslant \varphi(n)} \mu(E/K_n) \leqslant \frac{\varepsilon}{2^{n+1}}.$$

En conséquence, pour tout $m \geqslant \varphi(n)$, on a

$$\mu_m(K_n^{\varepsilon/2^{n+1}}) \geqslant \mu_{\varphi(n)}(K_n) - d_{LP}(\mu_m, \mu_{\varphi(n)}) \geqslant 1 - \frac{\varepsilon}{2^n}.$$

De plus, cette inégalité est aussi valable pour $m \leqslant \varphi(n)$, par choix de K_n. Posons ensuite

$$K = \bigcap_n \overline{K_n^{\varepsilon/2^{n+1}}}.$$

Comme $K_n^{\varepsilon/2^{n+1}}$ peut être recouvert par un nombre fini de boules de rayon $\varepsilon 2^{-n-1}$, K peut être recouvert par un nombre fini de boules de rayon arbitrairement petit, et est donc précompact. De plus, c'est un ensemble fermé dans un espace complet, il est donc compact. De plus, pour tout m, on a

$$\mu_m(E/K) \leqslant \sum_n \mu_m((\overline{K_n^{\varepsilon/2^{n+1}}})^c) \leqslant \sum_n \frac{\varepsilon}{2^{n+1}} = \varepsilon.$$

C'est donc bien une famille tendue, ce qui conclut la preuve. □

5.4 Autres distances sur les espaces des mesures de probabilités

La distance de Lévy-Prokhorov est utile pour des considérations théoriques, mais est difficilement utilisable pour l'analyse quantitative de problèmes concrets, en particulier parce qu'elle est difficile à calculer. On verra dans cette section d'autres distances sur les espaces de mesures, plus faciles à estimer dans des cadres concrets.

NB. *Par abus de notation, lorsque d est une distance sur un espace de mesures de probabilité et X et Y deux variables aléatoires, on notera $d(X,Y)$ pour la distance entre la loi de X et la loi de Y.*

5.4.1 Distance de Kolmogorov

Définition 5.4.1 (Distance de Kolmogorov)**.** *La distance de Kolmogorov entre deux mesures de probabilité sur $\mathbb{R}$ est donnée par*

$$d_{Kol}(\mu, \nu) := \sup_x |\mu((-\infty, x]) - \nu((-\infty, x])| = ||F_\mu - F_\nu||_\infty.$$

La distance de Kolmogorov permet de contrôler la convergence en loi, puisque si on a convergence uniforme des fonctions de répartition, on a a fortiori convergence simple. La réciproque n'est pas vraie : $d_{Kol}(\delta_{1/n}, \delta_0) = 1$ pour tout $n > 0$.

Exercice 5.4.1. *1. Calculer la distance de Kolmogorov entre deux lois de Bernoulli, en fonction de leurs paramètres.*

2. Calculer la distance de Kolmogorov entre deux lois exponentielles, en fonction de leurs paramètres.

Solution 5.4.1. *1. Notons p et q les deux paramètres. La fonction de répartition d'une loi de Bernoulli est constante par morceaux, avec une valeur qui change en 0 et 1. En comparant les valeurs en ces deux points, on voit que la distance de Kolmogorov est donnée par $|p-q|$.*

2. On considère deux lois exponentielles de paramètres respetcifs α et β, et on suppose sans perdre de généralité que $\alpha > \beta$. La différence des fonctions de répartition en x est donnée par $\exp(-\beta x) - \exp(-\alpha x)$. Si on dérive, on voit que le maximum est atteint en

$$x = \frac{\log(\alpha) - \log(\beta)}{\alpha - \beta},$$

et donc leur distance de Kolmogorov est $\exp\left(-\beta\left(\frac{\log(\alpha)-\log(\beta)}{\alpha-\beta}\right)\right) - \exp\left(-\alpha\left(\frac{\log(\alpha)-\log(\beta)}{\alpha-\beta}\right)\right)$.

Exercice 5.4.2. *Soit γ_t la mesure gaussienne centrée de variance t^2. Montrer que si μ a une densité bornée par une constante K, alors pour tout $h > 0$ on a*

$$F_{\gamma_t * \mu}(x) - F_\mu(x) \leqslant Kh + \frac{t}{h}$$

et en déduire que

$$d_{Kol}(\gamma_t * \mu, \mu) \leqslant 2\sqrt{tK}.$$

Solution 5.4.2. *Soit X une variable aléatoire de loi μ, et Z de loi γ_t, indépendante de X. Pour tout $h > 0$ on a*

$$\{X + Z \leqslant x\} \subset \{X \leqslant x + h\} \cup \{Z \leqslant -h\}$$

et donc

$$F_{\gamma_t * \mu}(x) \leqslant F_\mu(x+h) + \mathbb{P}(Z \leqslant -h) \leqslant F_\mu(x+h) + \frac{t}{h}.$$

De plus, comme la densité de μ est bornée, $F_\mu(x+h) - F_\mu(x) \leqslant Kh$. La première inégalité suit.

De même,

$$F_\mu(x) \leqslant F_{\gamma_t * \mu}(x-h) + \mathbb{P}(Z \geqslant h) \leqslant F_{\gamma * \mu}(x) + \frac{t}{h}.$$

On a donc bien

$$\sup_x |F_{\gamma_t * \mu}(x) - F_\mu(x)| \leqslant Kh + \frac{t}{h}.$$

On optimise sur h pour conclure.

On peut utiliser ce résultat pour relier la convergence en distance de Kolmogorov au comportement des fonctions caractéristiques :

Proposition 5.4.2. *Soit μ et ν deux lois de probailité sur $\mathbb{R}$ absolument continues par rapport à la mesure de Lebesgue, et de densités bornées par K. Il existe une constante $C > 0$, qui ne dépend pas de K, telle que pour tout $T > 0$ on ait*

$$d_{Kol}(\mu, \nu) \leqslant \frac{1}{2\pi} \int_{-T}^{T} \left| \frac{\varphi_\mu(t) - \varphi_\nu(t)}{t} \right| dt + CK^{2/5}T^{-2/5}.$$

Ce résultat est une variante d'une inégalité classique de Esseen, qui utilise une convolution avec une variable de densité $(1-\cos(Tx))/(Tx^2)$, dont la fonction caractéristique a l'avantage d'être à support compact, plutôt qu'avec une gaussienne. Le résultat optimal a un second terme en T^{-1}, et ne demande une borne sur la densité que pour une seule des deux variables.

Proof. Soit X (resp. Y) une variable aléatoire de loi μ (resp. ν). Soit Z une variable aléatoire gausienne de variance σ^2, indépendante de X et Y. On note f_σ (resp. g_σ) la densité de la loi de $X + Z$ (resp. $Y + Z$). De même, on note F_σ (resp. G_σ) la fonction de répartition de la loi de $X + Z$ (resp. $Y + Z$).Ces fonctions de répartition sont continues, donc en utilisant la formule d'inversion de Fourier, on a

$$F_\sigma(x) - G_\sigma(x) = \frac{1}{2\pi} \int_y \frac{(\varphi_X(y) - \varphi_Y(y))}{iy} \varphi_Z(y) e^{ix\cdot y} dy.$$

La fonction caractéristique de Z est $\exp(-\sigma^2 x^2/2)$. On a donc

$$F_\sigma(x) - G_\sigma(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{(\varphi_X(y) - \varphi_Y(y))}{iy} e^{ixy} e^{-\sigma^2 y^2/2} dy.$$

D'où

$$\sup_x |F_\sigma(x) - G_\sigma(x)| \leqslant \frac{1}{2\pi} \int_{-T}^{T} \frac{|\varphi_X(y) - \varphi_Y(y)|}{|y|} dy + \frac{2}{\pi} \int_T^{\infty} \frac{e^{-\sigma^2 y^2/2}}{|y|} dy. \quad (5.3)$$

Pour le deuxième terme, on prouve aisément une borne de la forme $C/(\sigma^2 T^2)$ avec C une constante positive qui ne dépend d'aucun des paramètres du problème. On a donc

$$d_{Kol}(X + Z, Y + Z) \leqslant \frac{1}{2\pi} \int_{-T}^{T} \frac{|\varphi_X(y) - \varphi_Y(y)|}{|y|} dy + \frac{C}{\sigma^2 T^2}.$$

Avec l'inégalité triangulaire, on en déduit que

$$d_{Kol}(X, Y) \leqslant \frac{1}{2\pi} \int_{-T}^{T} \frac{|\varphi_X(y) - \varphi_Y(y)|}{|y|} dy + \frac{C}{\sigma^2 T^2} + 4\sqrt{\sigma K}.$$

On conclut en prenant $\sigma = K^{-1/5} T^{-4/5}$. □

On a une distinction d'ordre topologique entre la distance de Kolmogorov et la distance de Lévy-Prokhorov :

Proposition 5.4.3. *L'espace des mesures de probabilité sur $\mathbb{R}$ muni de la distance d_{Kol} n'est* pas *séparable.*

Proof. On a $d_{Kol}(\delta_x, \delta_y) = \mathbb{1}_{x \neq y}$. On a donc une famile indénombrable d'éléments dont les distances deux à deux sont uniformément minorées par 1, ce qui n'est pas possible dans un espace séparable. □

5.4.2 Distance en variation totale

Définition 5.4.4 (Variation totale). *La distance en variation totale entre deux mesures de probabilité sur un espace E est*

$$d_{TV}(\mu,\nu) := \sup_{A \text{ mesurable}} |\mu(A) - \nu(A)|.$$

Sur $\mathbb{R}$, on a trivialement $d_{TV} \geqslant d_{Kol}$.

Proposition 5.4.5. *Si μ et ν sont deux mesures de probabilité sur $\mathbb{R}^k$ avec des densités f et g par rapport à la mesure de Lebesgue, alors*

$$d_{TV}(\mu,\nu) = \frac{1}{2}||f-g||_{L^1(dx)}.$$

Proof. Soit A un ensemble mesurable. Comme $\int_{\mathbb{R}^k} f dx = \int_{\mathbb{R}^k} g dx = 1$, on a

$$\left|\int_A f - g dx\right| = \left|\int_{A^c} f - g dx\right|$$

Et donc

$$|\mu(A) - \nu(A)| = \frac{1}{2}\left(\left|\int_A f - g dx\right| + \left|\int_{A^c} f - g dx\right|\right) \leqslant \frac{1}{2}\int |f-g| dx.$$

Réciproquement, si on considère l'ensemble mesurable $A = \{x; f(x) > g(x)\}$, on a

$$\int |f-g|dx = \int_A f - g dx + \int_{A^c} g - f dx = |\mu(A) - \nu(A)| + |\mu(A^c) - \nu(A^c)| = 2|\mu(A) - \nu(A)|.$$

□

Tout comme la distance de Kolmogorov, la distance en variation totale ne rend pas l'espace des mesures de probabilité séparable en général (même si ça peut être le cas, par exemple sur un espace fini).

5.4.3 Distance de Wasserstein

Définition 5.4.6 (Distance de Wasserstein L^1).

$$W_1(\mu,\nu) := \inf\{\mathbb{E}[|X-Y|]; \mathcal{L}(X) = \mu, \mathcal{L}(Y) = \nu\}.$$

Cette distance est un cas particulier de distance de transport optimal, dont l'étude remonte à Monge au 18ème siècle, et qui ont été introduites dans leur formulation moderne par Kantorovitch dans les années 40.

Exercice 5.4.3. *Montrer que si μ est une mesure de probabilité à support dans $[a,b]$ et ν une mesure de probabilité à support dans $[c,d]$ avec $c > b$, alors*

$$W_1(\mu,\nu) = \int y d\nu - \int x d\mu.$$

Donner un contre-exemple lorsque la condition sur les supports n'est pas vérifiée.

Solution 5.4.3. *Pour tout couplage (X,Y) de ces deux mesures, à cause de la comparaison entre les supports on a $X<Y$, et donc*

$$\mathbb{E}[|X-Y|]=\mathbb{E}[Y]-\mathbb{E}[X].$$

Comme contre-exemple, on a par exemple $\mu=\delta_0$ et $\nu=\frac{1}{2}(\delta_{-1}+\delta_1)$.

Proposition 5.4.7. *Soit μ et ν deux mesures de probabilté avec des moments d'ordre 1 finis. Il existe un couplage π de μ et ν tel que*

$$\int |x-y|d\pi = W_1(\mu,\nu).$$

Ce couplage est appelé couplage (ou transport) optimal pour le coût L^1.

On démontrera ce résultat d'existence par un argument de compacité :

Lemme 5.4.8. *Pour toutes mesures de probabilité μ et ν sur $\mathbb{R}^d$ données, l'ensemble des couplages de μ et ν est compact.*

Proof. Commençons par montrer que l'ensemble des couplages est relativement compact. Soit $\varepsilon>0$ et K_1 et K_2 des compacts tels que

$$\mu(K_1^c)\leqslant\varepsilon/2;\quad \nu(K_2^c)\leqslant\varepsilon/2.$$

Alors pour tout couplage π on a

$$\pi((K_1\times K_2)^c)\leqslant\pi(K_1^c\times\mathbb{R}^d)+\pi(\mathbb{R}^d\times K_2^c)=\mu(K_1^c)+\nu(K_2^c)\leqslant\varepsilon.$$

Comme $K_1\times K_2$ est compact, on en déduit que l'ensemble des couplages est tendu, et donc relativement compact par le Théorème de Prokhorov.

Il nous reste à montrer que l'ensemble des couplages est fermé pour la convergence étroite pour conclure la preuve. Soit π_n une suite de couplages, qui converge étoitement vers une mesure de probabilité π sur $\mathbb{R}^d\times\mathbb{R}^d$. Pour toute fonction continue bornée

$$\int f(x)d\pi(x,y)=\lim\int f(x)d\pi_n(x,y)=\lim_n\int f(x)d\mu(x)=\int f(x)d\mu(x).$$

Donc la première marginale de π est μ. Le même raisonnement appliqué àla seconde marginale montre qu'elle est égale à ν, et donc π est aussi un couplage de μ et ν. L'ensemble des couplages est donc bien fermé, et donc compact. □

Exercice 5.4.4. *Démontrer la Proposition 5.4.7*

Solution 5.4.4 (Preuve de la Proposition 5.4.7)**.** *Tout d'abord, comme les mesures ont des moments d'ordre 1 finis, $W_1(\mu,\nu)$ est fini.*

Soit π_n une suite de couplages de μ et ν tels que

$$\lim\int |x-y|d\pi_n = W_1(\mu,\nu)$$

Comme l'ensemble des couplages est compact, quitte à extraire, on peut supposer que π_n converge étroitement vers un couplage π. Comme π est un couplage, par définition on a

$$\int |x-y|d\pi\geqslant W_1(\mu,\nu).$$

De plus, pour tout $R > 0$

$$\begin{aligned}\int (|x-y| \wedge R) d\pi &= \lim_n \int (|x-y| \wedge R) d\pi_n \\ &\leqslant \lim_n \int |x-y| d\pi_n = W_1(\mu, \nu)\end{aligned}$$

Et donc, par limite monotone en R, on a

$$\int |x-y| d\pi \leqslant W_1(\mu, \nu).$$

Comme on avait déjà l'inégalité inverse, on en déduit qu'il y a égalité, ce qui conclut la preuve.

On a aussi une formulation duale de W_1 :

Proposition 5.4.9 (Formule de Kantorovitch-Rubinstein)**.**

$$W_1(\mu, \nu) = \sup \left\{ \int f d\mu - \int f d\nu;\ f \textit{1-Lipschitz} \right\}.$$

Esquisse de preuve. On tout d'abord

$$\sup_{f,g} - \int f(x) - g(y) d\gamma(x,y) + \int f d\mu - \int g d\nu = \begin{cases} 0 & \text{si } \gamma \in \Pi(\mu, \nu) \\ +\infty & \text{sinon} \end{cases}$$

où le sup est sur l'ensemble des mesures positives sur $\mathbb{R}^d \times \mathbb{R}^d$ (et pas juste les mesures de probabilité). On peut donc relacher la contrainte de couplage en

$$\inf_{\pi \in \Pi(\mu,\nu)} \int |x-y| d\pi = \inf \gamma \geqslant 0 \sup_{f,g} \int |x-y| d\gamma - \int f(x) - g(y) d\gamma(x,y) + \int f d\mu + \int g d\nu.$$

On admet qu'on peut échanger l'inf et le sup (théorème du minmax de Rockafellar). Comme de plus

$$\inf_{\gamma \geqslant 0} \sup_{f,g} \int |x-y| d\gamma - \int f(x) - g(y) d\gamma(x,y) = \begin{cases} 0 & \text{si } f(x) - g(y) \leqslant |x-y| \ \ \forall x, y \\ +\infty & \text{sinon,} \end{cases}$$

on a alors

$$\inf_{\pi \in \Pi(\mu,\nu)} \int |x-y| d\pi = \sup_{f(x) - g(y) \leqslant |x-y|} \int f d\mu - \int g d\nu.$$

Si f est 1-lipschitz, on peut prendre $g = f$. Sinon, on peut remplacer g par $\hat{g}(y) := \sup_x f(x) - |x-y|$. Cette fonction est 1-lipschitz, et pour toute fonction f telle que $f(x) - \tilde{g}(y) \leqslant |x-y|$ pour tout x et y, on a $f \leqslant \hat{g}$, donc on peut remplacer f par $\hat{g}$. On peut donc se restreindre aux fonctions 1-lipschitz. □

Exercice 5.4.5. *Quelle est la distance W_1 entre deux lois de Bernoulli de paramètres respectifs p et q? Et entre deux lois gaussiennes de variance 1 et de moyennes différentes?*

Solution 5.4.5. *Sans perdre de généralité, on suppose $p > q$. Si on considère le couplage où*

$$\pi(1,1) = q; \quad \pi(1,0) = p - q; \quad \pi(0,0) = q,$$

alors on voit que $W_1(\mathrm{Ber}(p), \mathrm{Ber}(q)) \leqslant |p - q|$.

Réciproquement, avec la formule de dualité de Kantorovitch-Rubinstein, pour tout fonction f 1 lipschitz on a

$$W_1(\mathrm{Ber}(p), \mathrm{Ber}(q)) \geqslant (p - q)(f(1) - f(0)).$$

En prenant comme fonction $f(x) = x$ on en déduit que $W_1 \geqslant |p - q|$.

Pour des gaussiennes de moyenne $a > b$, en testant la formule de dualité avec $f(x) = x$ on a $W_1(\mathcal{N}(a,1), \mathcal{N}(b,1)) \geqslant a - b$. Reciproquement, en considérant le couplage $(X, X + b - a)$ où $X \equiv \mathcal{N}(a,1)$, on a $W_1 \leqslant a - b$. On conclut que la distance W_1 entre deux gaussiennes de mêmes variances est donnée par la distance entre les moyennes. A noter que ce résultat se généralise en dimension supérieure.

On peut ensuite regarder quelle est la topologie induite par W_1.

Théorème 5.4.10. *Soit $(\mu_n)_{n\in\mathbb{N}}$ et μ des mesures de probabilité sur $\mathbb{R}^d$. Les propositions suivantes sont équivalentes :*

1. *$(\mu_n)_{n\in\mathbb{N}}$ converge étroitement vers μ et $\int |x| d\mu_n \longrightarrow \int |x| d\mu$;*
2. *$W_1(\mu_n, \mu) \longrightarrow 0$.*

Pour démontrer ce résultat, on commencera par le cas compact, donné par le lemme suivant :

Lemme 5.4.11. *Soit $(\mu_n)_{n\in\mathbb{N}}$ et μ des mesures de probabilité sur la boules fermé $B(0,R)$. Alors (μ_n) converge étroitement vers μ ssi $W_1(\mu_n, \mu) \longrightarrow 0$.*

Proof. Le sens réciproque est immédiat via la Proposition 5.1.2 (puisque les fonctions lispchitz sur un compact sont bornées).

Pour le sens direct, on va raisonner par contradiction. Supposons que $W_1(\mu_n, \mu)$ ne tend pas vers 0. Par la formule de dualité de Kantorovitch-Rubinstein, il existe alors $\varepsilon > 0$ et une suite de fonctions 1-lispchitz sur $B(0,R)$ telles que

$$\inf_n \int f_n d\mu_n - \int f_n d\mu > \varepsilon.$$

Sans perdre de généralité, on peut supposer que $f_n(0) = 0$ pour tout n. On peut alors utiliser le théorème d'Arzela-Ascoli pour en extraire une suite convergent vers une fonction f pour la norme uniforme. Cette fonction f est encore 1 lipschitz et vérifie $f(0) = 0$. Par convergence étroite vers μ, pour tout N assez grand

$$\left| \int f d\mu_N - \int f d\mu \right| \leqslant \varepsilon/4.$$

Mais par convergence uniforme de f_n vers f, il existe des N arbitrairement grands tels que

$$\left| \int f_N d\mu_N - \int f d\mu_N \right| \leqslant \varepsilon/4; \quad \left| \int f_N d\mu - \int f d\mu \right| \leqslant \varepsilon/4.$$

On en déduit qu'on peut trouver N tel que

$$\left| \int f_N d\mu_N - \int f_N d\mu \right| \leqslant 3\varepsilon/4,$$

ce qui aboutit à une contradiction. □

Preuve du Théorème 5.4.10. Pour démontrer que (2) $\Rightarrow$ (1), on utilise la convergence des espérances de fonction lispchitz, et la Proposition 5.1.2 pour obtenir la convergence en loi. Pour la convergence du moment, on peut utiliser la borne

$$|\mathbb{E}[|X|] - \mathbb{E}[Y]| \leqslant \mathbb{E}[|X - Y|]$$

pour tout couplage (X, Y) de μ_n et ν_n, et donc en particulier

$$\left| \int |x| d\mu_n - \int |x| d\mu \right| \leqslant W_1(\mu_n, \mu).$$

Démontrons maintenant que (1) $\Rightarrow$ (2). Pour chaque mesure on définit sa projection sur la boule de rayon R par

$$\mu^R := \frac{\mu \mathbb{1}_{B(0,R)}}{\mu(B(0,R))}.$$

Comme ces mesures vivent sur un compact, et que $\mu_n(B(0,R)) \longrightarrow \mu(B(0,R))$ pour presque tout $R > 0$, on peut trouver R arbitrairement grand tel que

$$W_1(\mu_n^R, \mu^R) \longrightarrow 0.$$

De plus, en prenant un couplage (X, Y) de μ et μ^R où conditionnellement à $|X| \leqslant R$ on a $Y = X$ (ce qui est possible car $\mu(B(0,R)) \leqslant 1 = \mu^R(B(0,R))$), on a

$$W_1(\mu^R, \mu) \leqslant \int (|x| + R) \mathbb{1}_{|x| \geqslant R} d\mu \leqslant 2 \int |x| \mathbb{1}_{|x| \geqslant R} d\mu \underset{R \to \infty}{\longrightarrow} 0.$$

De plus, comme $\int |x| d\mu_n \longrightarrow \int |x| d\mu$, on a aussi pour tout R

$$\int |x| \mathbb{1}_{|x| \geqslant R} d\mu_n \longrightarrow \int |x| \mathbb{1}_{|x| \geqslant R} d\mu.$$

Donc pour tout $\varepsilon > 0$ on peut trouver R arbitrairement grand et N tels que pour tout $n \geqslant N$

$$W_1(\mu_n^R, \mu_n) \leqslant \varepsilon.$$

Alors en prenant R assez grand,

$$\limsup_n W_1(\mu_n, \mu) \leqslant \limsup_n (W_1(\mu_n^R, \mu_n) + W_1(\mu_n^R, \mu^R)) + W_1(\mu^R, \mu) \leqslant 2\varepsilon.$$

Comme ε est arbitraire, ceci conclut la preuve. □

Proposition 5.4.12. *Soit μ et ν deux mesures de probabilité sur $\mathbb{R}$. On suppose que ν a une densité bornée par rapport à la mesure de Lebesgue, et soit C un majorant de la densité. Alors*

$$d_{Kol}(\mu, \nu) \leqslant 2\sqrt{C W_1(\mu, \nu)}.$$

Proof. Soit $x \in \mathbb{R}$, $\varepsilon > 0$ et g_1 la fonction définie par

$$g_1(t) := \begin{cases} 1 \text{ si } t \leqslant x; \\ 0 \text{ si } t \geqslant x + \varepsilon; \\ \frac{t-x}{\varepsilon} \text{ si } t \in (x, x + \varepsilon). \end{cases}$$

Alors g_1 est ε^{-1}-lipschitz, et $\mathbb{1}_{(-\infty,x]} \leqslant g_1$. Donc

$$\mu((-\infty,x]) - \nu((-\infty,x]) \leqslant \mu(g_1) - \nu(g_1) + \nu(g_1) - \nu((-\infty,x]).$$

Or $\mu(g_1) - \nu(g_1) \leqslant \varepsilon^{-1} W_1(\mu,\nu)$, et $\nu(g_1) - \nu((-\infty,x]) \leqslant \nu((x,x+\varepsilon)) \leqslant C\varepsilon$ car la densité de ν est bornée par C. On en déduit

$$d_{Kol}(\mu,\nu) \leqslant \varepsilon^{-1} W_1(\mu,\nu) + C\varepsilon.$$

On peut alors prendre $\varepsilon = \sqrt{W_1/C}$ pour conclure. □

5.5 Théorème de représentation de Skorokhod

Théorème 5.5.1. *Soit $(\mu_n)_{n\in\mathbb{N}}$ une suite de lois de probabilité sur un espace polonais (E,d), qui converge étroitement vers une mesure de probabilité μ. Alors il existe un espace de probailité $(\Omega,\mathcal{F},\mathbb{P})$ et des variables aléatoires $(Z_n)_{n\in\mathbb{N}}$ et Z sur cet espace telles que*

1. *Pour tout $n \in \mathbb{N}$ la loi de Z_n est μ_n, et la loi de Z est μ;*
2. *$(Z_n)_{n\in\mathbb{N}}$ converge p.s. vers Z.*

Cas $E = \mathbb{R}$. Si $F_\mu : x \longrightarrow \mu((-\infty,x]$ est la fonction de répartition de la mesure de probabilité μ sur $\mathbb{R}$, alors la fonction

$$F_\mu^{-1}(t) := \inf\{x \in \mathbb{R}, F_\mu(x) \geqslant t\}$$

est une fonction de $(0,1)$ dans $\mathbb{R}$, telle que si U est uniforme sur $(0,1)$, alors $F_\mu^{-1}(U)$ a pour loi μ. Cette construction nous permet de constuire des suites de variables de lois données, mais fortement corrélées.

Si μ_n converge étroitement vers μ, alors on a pour tout $t \in (0,1)$

$$F_\mu^{-1}(t) \leqslant \liminf F_{\mu_n}^{-1}(t) \leqslant \limsup F_{\mu_n}^{-1}(t) \leqslant F_\mu^{-1}(t+)$$

où $F_\mu^{-1}(t+)$ est la limite à droite de F_μ^{-1} en t. Comme F_μ^{-1} est monotone, elle est continue presque partout, et on en déduit que $F_{\mu_n}^{-1}(U)$ converge p.s. vers $F_\mu^{-1}(U)$, ce qui conclut la preuve. □

Preuve du cas général pas faite en cours. □

Exercice 5.5.1. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoire réelles uniformément intégrable, et qui converge en loi vers X. Alors $\mathbb{E}[X_n] \longrightarrow \mathbb{E}[X]$.*

Solution 5.5.1. *D'après le théorème de Skorokhod, il existe une suite de variables Y_n constuites sur un même espace de probabilité, telles que pour tout n X_n et Y_n ont la même loi, et Y_n converge p.s. vers un variable Y de même loi que X. L'uniforme intégrabilité étant une propriété qui ne dépend que de la suite des lois, les Y_n sont aussi uniformément intégrables.*

D'après le Théorème 2.4.7, une suite de variables aléatoire uniformément intégrables qui converge p.s. converge aussi dans L^1, la conclusion suit.

5.6 Appendice : convergence faible-* et convergence en loi

Cette section ne sera pas traitée en cours.

Définition 5.6.1 (Convergence vague). *Une suite (μ_n) de mesures de probabilité sur un espace métrique (E,d) converge vaguement vers une mesure μ si pour toute fonction continue à support compact,*

$$\int f d\mu_n \longrightarrow \int f d\mu.$$

La convegrence étroite implique la convergence vague, mais la réciproque est fausse en général. Toutefois, sur $\mathbb{R}^d$ on a le résultat suivant :

Proposition 5.6.2. *Une suite de mesures de probabilité (μ_n) sur $\mathbb{R}^d$ converge étroitement vers une mesure de probabilité μ ssi elle converge vaguement et si $\mu_n(\mathbb{R}^d) \longrightarrow 1$.*

Plus généralement, ce résultat est vrai si l'espace métrique (E,d) est localement compact.

Proof. L'implication réciproque est immédiate. Prouvons le sens direct.

Soit f une fonction continue bornée et (φ_ℓ) une suite croissante de fonctions continues à support compact dont la limite est la fonciton constante égale à 1. Les fonctions $f\varphi_\ell$ sont continues à support compact, donc pour tout ℓ on a

$$\lim \int f\varphi_\ell d\mu_n \longrightarrow \int f\varphi_\ell d\mu.$$

On a aussi

$$\lim_\ell \int \varphi_\ell d\mu = 1; \quad \lim_n \int \varphi_\ell d\mu_n = \int \varphi_\ell d\mu.$$

Or

$$\begin{aligned}\left|\int f d\mu_n - \int f d\mu\right| &\leqslant \left|\int f d\mu_n - \int f\varphi_\ell d\mu_n\right| \\ &\quad + \left|\int f\varphi_\ell d\mu_n - \int f\varphi_\ell d\mu\right| + \left|\int f\varphi_\ell d\mu - \int f d\mu\right| \\ &\leqslant \left|\int f\varphi_\ell d\mu_n - \int f\varphi_\ell d\mu\right| + ||f||_\infty \left(2 - \int \varphi_\ell d\mu_n - \int \varphi_\ell d\mu\right)\end{aligned}$$

et on conclut en prenant la limsup lorsque n, puis ℓ, tendent vers l'infini.

□

Dans le cas d'un espace compact, la convergence étoite et la convergence vague coïncident. Dans le cadre non-compact, la topologie étroite ne coïncide plus avec la topologie vague, et c'est la convergence vague dont la topologie coïncide avec la topologie faible-* :

Théorème 5.6.3 (Théorème de représentation de Riesz-Markov-Kakutani). *Soit (E,d) un espace métrique localement compact. On munit l'espace des fonctions continues à support compact $C_c(E)$ de la topologie de la convergence uniforme.*

Alors pour toute forme linéaire ψ continue positive sur $C_c(E)$ il existe une mesure borélienne positive μ telle que

$$\psi(f) = \int f d\mu \quad \forall f \in C_C(E).$$

Si de plus on demande que la mesure μ soit régulière, alors elle est unique.

On peut donc identifier l'espace des mesures de probailité boréliennes régulières avec le sous-espace des formes linéaires continues positives et de norme égale à 1.

Ce résultat n'est pas vrai pour la convergence étroite : même si les mesures de probabilité représentent des formes linéaires continues sur l'espace des fonctions continues bornées, il existe des formes linéaires continues qui ne sont pas représentables par des mesures.

A partir du théorème de Riesz-Markov-Kakutani et du théorme de Banach-Alaoglu-Bourbaki, on en déduite le théorème suivant :

Proposition 5.6.4. *L'espace des mesures positives de masse totale inférieure à* 1 *est compact pour la convergence vague.*

Toute suite de mesures de probailité a donc une valeur d'adhérencepour la convergence vague, mais elle peut avoir une masse totale strictement inférieure à 1. En effet, l'espace des mesures de probabilité n'ets pas fermé pour la convergence vague si l'espace sous-jacent n'ets pas compact. Ceci explique le théorème de Prokhorov : la seule obstruction à la compacité de l'espace des mesures de probabilité sur un espace localement compact est le phénomène de perte de masse.

Chapter 6

Autour du théorème central limite

6.1 Rappels sur le TCL

Si $X = (X_1, ..., X_d)$ est un vecteur aléatoire de dimension d dont les coordonnées sont L^2, on notera $\mathrm{Cov}(X)$ sa matrice de covariance, définie par

$$\mathrm{Cov}(X)_{i,j} := \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j].$$

Cette notion est invariante par translations, et une matrice de covariance est toujours symétrique et positive. Si on applique la transformation affine

$$X \longrightarrow \mathrm{Cov}(X)^{-1/2}(X - \mathbb{E}[X])$$

on obtient un vecteur centré, dont la matrice de covariance est égale à l'identité.

Définition 6.1.1 (Vecteurs gaussiens). *Une variable aléatoire $X = (X_1, ..., X_d)$ à valeurs dans $\mathbb{R}^d$ est un vecteur gaussien si pour tout $y \in \mathbb{R}^d$ la variables $X \cdot y$ est une variable gaussienne réelle.*

Proposition 6.1.2. *La loi d'un vecteur gaussien est caractérisée par son espérance et sa matrice de covariance. On notera $\mathcal{N}(x, A)$ la loi d'un vecteur gaussien d'espérance x et de matrice de covariance A.*

On peut aussi caractériser les vecteurs gaussiens par leurs fonctions caractéristiques :

$$\varphi_{\mathcal{N}(x,A)}(z) := \exp(ix \cdot z + \frac{1}{2}\langle Ax, x\rangle).$$

Théorème 6.1.3 (Théorème central limite). *Soit $(X_i)_{i\in\mathbb{N}}$ une suite de v.a. i.i.d., centrées et dont les matrices de covariance sont égales à l'identité. Alors*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \xrightarrow{loi} \mathcal{N}(0, \mathrm{Id}).$$

Exercice 6.1.1. *Utiliser le théorème de Prokhorov et le TCL unidimensionnel pour démontrer le TCL multidimensionnel.*

Solution 6.1.1. *Soit $Y_n := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$. Comme on a une borne uniforme sur le second moment de la norme des vecteurs aléatoires Y_i, la suite de leurs loi est tendue, et donc on peut en extraire une sous suite qui converge en loi vers une loi μ sur $\mathbb{R}^d$.*

Comme pour tout vecteur p fixé les variables aléatoires $Y_n \cdot p$ vérifient les hypothèses du TCL unidimensionnel, on en déduit qu'elles converent en loi vers une gaussienne unidimensionnelle centrée de variance $|p|^2$. Par continuité, la mesure image de μ par l'application $x \longrightarrow x \cdot p$ est alors une mesure gaussienne centrée de variance $|p|^2$. Par définition, μ est donc une mesure gaussienne multidimensionnelle centrée, et sa matrice de covariance est alors Id.

6.2 Théorème de Lindeberg

Le but de cette section est de relâcher l'hypothèse que toutes les variables aléatoires ont la même loi dans le TCL.

Théorème 6.2.1 (TCL de Lindeberg)**.** *Soit $(X_i^n)_{n\in\mathbb{N},i\leqslant n}$ une famille de variable indépendantes centrées, et $\sigma_{i,n}^2 := \mathbb{E}[(X_i^n)^2]$. On pose*

$$S_n := \sum_{i=1}^n X_i^n; \quad s_n^2 := \sum_{i=1}^n \sigma_{i,n}^2.$$

Si pour tout $\varepsilon > 0$ on a

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E}[(X_i^n)^2 \mathbb{1}_{|X_i^n|>\varepsilon s_n}] = 0$$

alors S_n/s_n converge en loi vers une gaussienne centrée réduite.

Nous allons remettre la preuve de ce théorme, et d'abord voir quelques corollaires, plus facilement applicables en pratique.

Corollaire 6.2.2 (TCL de Lyapunov)**.** *Supposons qu'il existe $\delta > 0$ tel que*

$$\lim_n \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}[|X_i^n|^{2+\delta}] = 0.$$

Alors S_n/s_n converge en loi vers une gaussienne centrée réduite.

Proof. En utilisant l'inégalité de Holder avec exposants $(2+\delta)/2$ et $(2+\delta)/\delta$, on a

$$\mathbb{E}[X_i^2 \mathbb{1}_{|X_i|>\varepsilon s_n}] \leqslant \mathbb{E}[X_i^{2+\delta}]^{2/(2+\delta)} \mathbb{P}(|X_i| > \varepsilon s_n)^{\delta/(2+\delta)}.$$

Or, en utilisant l'inégalité de Markov

$$\mathbb{P}(|X_i| > \varepsilon s_n) \leqslant \frac{\mathbb{E}[|X_i|^{2+\delta}]}{(\varepsilon s_n)^{2+\delta}}$$

et donc

$$\frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E}[X_i^2 \mathbb{1}_{|X_i|>\varepsilon s_n}] \leqslant \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}[|X_i|^{2+\delta}] \varepsilon^{-\delta}.$$

On en déduit que l'hypothèse du TCL de Lyapunov est effectivement plus forte que l'hypothèse du TCL de Lindeberg. □

Exercice 6.2.1. *Montrer que si $\inf_i \sigma_{i,n} > \alpha > 0$ et $\sup_i \mathbb{E}[|X_i^n|^{2+\delta}] \leqslant \beta < \infty$ alors on peut appliquer le TCL de Lyapunov.*

Solution 6.2.1. *On a*

$$s_n^{2+\delta} = \left(\sum \sigma_{i,n}^2\right)^{1+\delta/2} \geqslant \alpha^{2+\delta} n^{1+\delta/2}$$

et

$$\sum_{i=1}^{n} \mathbb{E}[|X_i^n|^{2+\delta}] \leqslant \beta n.$$

Donc

$$\frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}[|X_i^n|^{2+\delta}] \leqslant \frac{\beta}{n^{\delta/2}\alpha^{2+\delta}} \longrightarrow 0.$$

Le TCL de Lyapunov est donc applicable.

Exercice 6.2.2. *Montrer que la condition de Lindeberg implique que*

$$\lim_n \max_{i \leqslant n} \frac{\sigma_{i,n}^2}{s_n^2} \longrightarrow 0.$$

Solution 6.2.2. *Pour tout $\varepsilon > 0$, on a*

$$\begin{aligned} \max_{i \leqslant n} \sigma_{i,n}^2 &\leqslant s_n^2 \varepsilon^2 + \max_{i \leqslant n} \mathbb{E}[(X_i^n)^2 \mathbb{1}_{|X_i^n| > \varepsilon s_n}] \\ &\leqslant s_n^2 \varepsilon^2 + \sum_{i=1}^{n} \mathbb{E}[(X_i^n)^2 \mathbb{1}_{|X_i^n| > \varepsilon s_n}] \end{aligned}$$

Et donc, en utilisant la condition de Lindeberg,

$$\limsup_n \max_{i \leqslant n} \frac{\sigma_{i,n}^2}{s_n^2} \leqslant \varepsilon^2.$$

Comme ε était arbitraire, on en déduit la convergence vers 0.

Corollaire 6.2.3. *Dans le cadre du TCL de Lindeberg, si il existe $\beta > 0$ tel que pour tout i, n $|X_i^n| \leqslant \beta$ et $s_n \longrightarrow \infty$, alors S_n/s_n converge en loi vers une gaussienne centrée réduite.*

Proof. Comme les X_i^n sont uniformément bornés, pour tout $\varepsilon > 0$, dès que n suffisamment grand et pour tout $i \leqslant n$ on a $\mathbb{P}(|X_i^n| \geqslant \varepsilon s_n) = 0$. L'hypothèse du TCL de Lindeberg est donc vérifiée. □

Passons maintnant à la preuve du TCL de Lindeberg. Elle reposera sur les deux lemmes élémentaires suivants :

Lemme 6.2.4 (Développements de Taylor)**.** *Pour tout $x \in \mathbb{R}$, on a*

$$|e^{ix} - 1| \leqslant \min(2, |x|); \quad |e^{ix} - 1 - ix| \leqslant \min(2|x|, x^2/2); \tag{6.1}$$

$$|e^{ix} - 1 - ix + x^2/2| \leqslant \min(x^2, |x|^3/6). \tag{6.2}$$

Lemme 6.2.5. *Pour toutes collections $z_1, ..., z_n$ et $w_1, ..., w_n$ de nombres complexes de modules tous inférieurs à 1, on a*

$$\left| \prod_{k=1}^{n} z_k - \prod_{k=1}^{n} w_k \right| \leqslant \sum_{k=1}^{n} |z_k - w_k|.$$

Preuve du TCL de Lindeberg. Le but va être d'appliquer le théorème de Lévy, en faisant un développement asymptotique de la fonction caractéristique. Sans perdre de généralité, on supposera $s_n = 1$ pour tout n.

On pose $S_n := \frac{1}{s_n}\sum_{k=1}^n X_k^n$, et on notera φ_U la fonction caractéristique de la loi de la variable aléatoire U. Par indépendance, on a

$$\varphi_{S_n}(x) = \prod_{k=1}^n \varphi_{X_k^n}(x).$$

De plus, en utilisant le Lemme 6.2.4, on a

$$|\varphi(x) - 1 + t^2\sigma_{k,n}^2/2| \leqslant t^2\mathbb{E}[\min((X_k^n)^2, |t||X_k^n|^3)];$$

et comme, pour tout $\varepsilon > 0$,

$$\begin{aligned}\mathbb{E}[\min((X_k^n)^2, |t||X_k^n|^3)] &\leqslant |t|\mathbb{E}[|X_k^n|^3\mathbb{1}_{|X_k^n|\leqslant\varepsilon}] + \mathbb{E}[|X_k^n|^2\mathbb{1}_{|X_k^n|>\varepsilon}] \\ &\leqslant |t|\varepsilon\sigma_{k,n}^2 + \mathbb{E}[|X_k^n|^2\mathbb{1}_{|X_k^n|>\varepsilon}].\end{aligned}$$

Donc, en sommant sur $k \leqslant n$, comme $s_n = 1$ on obtient

$$\limsup_n \sum_{k\leqslant n} |\varphi(x) - 1 + t^2\sigma_{k,n}^2/2| \leqslant |t|\varepsilon.$$

Comme ε était arbitraire, on a

$$\lim_n \sum_{k\leqslant n} |\varphi(x) - 1 + t^2\sigma_{k,n}^2/2| = 0.$$

De plus, comme $\max_{k\leqslant n}\sigma_{k,n} \longrightarrow 0$ quand n tend vers l'infini (cf. Exercice 6.2.2), pour t fixé et n suffisamment grand les $1 - \sigma_{k,n}^2 t^2$ sont tous plus petits que 1 en valeur absole, et donc on peut appliquer le Lemme 6.2.5, et on obtient

$$\prod_{k=1}^n \varphi_{X_k^n}(t) = \prod_{k=1}^n (1 - \sigma_{k,n}^2 t^2) + o(1) = \exp(-t^2/2) + o(1),$$

où on a une nouvelle fois utilisé que $\max_{k\leqslant n}\sigma_{k,n} \longrightarrow 0$. Ceci conclut la preuve.

□

6.2.1 Une application en combinatoire

Nous allons maintenant voir une conséquence du TCL de Lindeberg à l'étude des permutations aléatoires : les fluctuations du nombre de cycles d'une permutation aléatoire uniforme sont asymptotiquement gaussiennes.

Notre but va être de démontrer le résultat suivant :

Théorème 6.2.6. *Soit π_n une permutation aléatoire uniforme de $\mathcal{S}_n$, et K_n : $\mathcal{S}_n \longrightarrow \mathbb{N}$ la fonction qui, a une permutation de $\mathcal{S}_n$ associe le nombre de cycles dans sa décomposiiton minimale en produit de cycles. Alors*

$$\frac{K_n(\pi_n) - \mathbb{E}[K_n(\pi_n)]}{\mathrm{Var}(K_n(\pi_n))^{1/2}} \xrightarrow{loi} \mathcal{N}(0,1).$$

Pour se ramener au TCL de Lindeberg, il nous faut représenter $K_n(\pi_n)$ comme une somme de variable indépendantes. Pour cela, on va utiliser le couplage de Feller, qui construit une permutation aléatoire uniforme de $\mathcal{S}_n$ à partir d'une famille de variables $X_1^n, ..., X_n^n$ indépendantes, de lois

$$\mathbb{P}(X_i^n = 0) = 1 - \frac{1}{i}; \;\; \mathbb{P}(X_i^n = 1) = \frac{1}{i}.$$

Pour cela, on procède en n étapes. A la première étape, on part avec l'élément 1, et on lui choisit une image uniformément au hasard, en choisissant comme image 1 si $X_n^n = 1$ (ce qui clot un cycle). Ensuit, à l'étape i, on clot le cycle courant si $X_{n-i+1}^n = 1$, et on redémarre un nouveau cycle avec le plus petit élème pour lequel on n'a pas encore choisit d'image. Sinon, si $X_{n-i+1} = 0$, on chosiit une image au denrier élément du cycle uniformément parmis les éléments qui ne sont pas encore dans un cyle, et on démarre l'étape suivante avec ce même cycle inachevé comme cycle courant.

Cette procédure définit bien une permutation aléatoire uniforme. Et avec cette représentation, s π_n est la permutation aléatoire qu'on a construit, comme à chaque étape on ferme un cycle ssi $X_{n-i+1}^n = 1$, on a

$$K_n(\pi_n) = \sum_{i=1}^{n} X_i^n.$$

Donc K_n se représente bien comme une some de variables indépendantes, mais pas identiquement distribuées. Donc on ne peut pas appliquer le TCL classique, mais on peut espérer appliquer le TCL de Lindeberg. Et en fait, ce sera le Corollaire 6.2.3 qu'on va pouvoir appliquer. En effet, pour tout i, n $|X_i^n - \mathbb{E}[X_i^n]| \leqslant 1$, alors que

$$s_n^2 = \sum_{i=1}^{n} \frac{i-1}{i^2} \;\equiv \log n \;\longrightarrow \infty.$$

Exercice 6.2.3. *A partir du couplage de Feller, calculer* $\mathbb{E}[K_n(\pi_n)]$ *et* $\mathrm{Var}(K_n(\pi_n))$.

Solution 6.2.3. *On a*

$$\mathbb{E}[K_n(\pi_n)] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{1}{i}$$

et

$$\mathrm{Var}(K_n(\pi_n)) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = \sum_{i=1}^{n} \frac{i-1}{i^2}.$$

6.3 Vitesse de convergence dans le TCL

6.3.1 Lemme de Stein

Le but de cette section est de donner une borne sur la distance $W_1(\mu, \gamma)$ entre une mesure de probabilité arbitraire sur $\mathbb{R}$ et la loi gaussienne centrée réduite, via des formules d'intégration par partie. Le point de départ est la caractérisation suivante de la mesure gaussienne :

Lemme 6.3.1 (Lemme de Stein). *La loi normale centrée réduite $\mathcal{N}(0,1)$ est la seule loi de probabilité sur $\mathbb{R}$ telle que pour toute fonction C^1, à croissance au plus polynomiale et dont la dérivée est à croissance au plus polynomiale, on ait*

$$\mathbb{E}[Xf(X)] = \mathbb{E}[f'(X)].$$

Le fait que la mesure gaussienne vérifie cette formule se prouve par IPP, car

$$\int xf(x)e^{-x^2/2}dx = -\int f(e^{-x^2/2})'dx = \int f'e^{-x^2/2}dx.$$

On omet la preuve du sens réciproque, qui sera impliquée par la proposition ci-dessous.

Le résultat principal de cette section est que ce lemme peut être renforcé en une version quantitative :

Proposition 6.3.2. *Soit γ la loi $\mathcal{N}(0,1)$. Pour toute mesure de probailité μ sur $\mathbb{R}$, on a*

$$W_1(\mu,\gamma) \leqslant \sup_{||f||_\infty, ||f'||_\infty \leqslant 1, ||f''||_\infty \leqslant 4} \mathbb{E}_\mu[f'(X) - Xf(X)].$$

Il est possible d'améliorer les constantes 1 et 4 dans ce résultat, mais la preuve serait un peu plus longue et technique que celle qu'on donnera.

La preuve repose sur le lemme de régularisation suivant

Lemme 6.3.3. *Soit h une fonction 1-lipschitz. Il existe une solution f de l'équation différentielle*

$$f' - xf = h - \mathbb{E}_\gamma[h]$$

telle que $|f|_\infty \leqslant 1$, $|f'|_\infty \leqslant 1$ et $|f''|_\infty \leqslant 4$.

Proof. Pour cette preuve, on rappelle le Théorème 2.4.9 : toute fonction lipschitz est dérivable presque partout. On peut alors définir la fonction h' comme étant la dérivée de h aux points où elle existe, et 0 partout ailleurs. On a de plus $h(x) = h(y) + \int_y^x h'(t)dt$.

On peut alors directement vérifier que la fonction suivante est une solution de l'équation différentielle :

$$f(x) := -\int_0^1 \frac{1}{2\sqrt{t}\sqrt{1-t}} \mathbb{E}_\gamma[Xh(\sqrt{t}x + \sqrt{1-t}X)]dt.$$

En effet, en dérivant sous l'intégrale (ce qui se justifie par le théorème de convergence dominée), on a

$$f'(x) = -\int_0^1 \frac{1}{2\sqrt{1-t}} \mathbb{E}_\gamma[Xh'(\sqrt{t}x + \sqrt{1-t}X)]dt. \tag{6.3}$$

Mais par intégration par partie gaussienne

$$\mathbb{E}_\gamma[Xh(\sqrt{t}x + \sqrt{1-t}X)] = \sqrt{1-t}\mathbb{E}_\gamma[h'(\sqrt{t}x + \sqrt{1-t}X)]. \tag{6.4}$$

Donc

$$\begin{aligned} f'(x) - xf(x) &= \int_0^1 \mathbb{E}_\gamma\left[\left(-\frac{X}{2\sqrt{1-t}} + \frac{x}{2\sqrt{t}}\right) h'(\sqrt{t}x + \sqrt{1-t}X)\right] dt \\ &= \int_0^1 \mathbb{E}_\gamma\left[\frac{d}{dt} h(\sqrt{t}x + \sqrt{1-t}X)\right] dt \\ &= h(x) - \mathbb{E}_\gamma[h(Z)]. \end{aligned}$$

Ce n'est pas la seule solution de l'équation différentielle, mais on peut montrer que c'est la seule solution bornée, car la différence de deux solutions est nécessairement de la forme $Ce^{x^2/2}$.

A partir de cette formule, on peut explicitement calculer les bornes voulues. Tout d'abord, comme $|h'| \leqslant 1$, en utilisant (6.4), on a

$$||f||_\infty \leqslant \int_0^1 \frac{1}{2\sqrt{t}} dt = 1.$$

Ensuite, en utilisant (6.3), on a

$$||f'||_\infty \leqslant \int_0^1 \frac{1}{2\sqrt{1-t}} \mathbb{E}_\gamma[|X|] dt \leqslant 1.$$

Pour borner la dérivée seconde, en dérivant l'équation différentielle, on a

$$f'' - xf' = f + h'$$

Donc on peut voir f' comme la solution bornée de l'équation différentielle en prenant comme second membre $\tilde{h} = f + h'$, qui vérifie la borne $|h|_\infty \leqslant 2$. On peut vérifier que les solutions bornées de l'équation différentielle $f' - xf = g - \mathbb{E}_\gamma[g]$ peuvent aussi s'écrire sous la forme

$$f(x) = e^{x^2/2} \int_{-\infty}^{x} (g(y) - \mathbb{E}_\gamma[g]) e^{-y^2/2} dy.$$

En effet, c'est bien une solution, et elle n'explose pas exponentiellement à l'infini, donc c'est la même solution que celle définie par la formule de représentation stochastique ci-dessus, puisque deux solutions diffèrent nécessairement par un terme de la forme $Ce^{x^2/2}$. Alors on a

$$f'(x) = g(x) - \mathbb{E}_\gamma[g] + xe^{x^2/2} \int_{-\infty}^{x} (g(y) - \mathbb{E}_\gamma[g]) e^{-y^2/2} dy,$$

et donc

$$|f'(x)| \leqslant ||g - \mathbb{E}_\gamma(g)||_\infty \left(1 + xe^{x^2/2} \left(\sqrt{2\pi} - \int_{-\infty}^{x} e^{-y^2/2} dy\right)\right).$$

On a ensuite l'inégalité classique sur la fonction de répartition gaussienne suivante (inégalité de Mills)

$$\int_x^{+\infty} e^{-y^2/2} dy \leqslant \frac{e^{-x^2/2}}{x}$$

qui permet de conclure. Pour démontrer cette dernière, on a

$$\left(-\exp(-x^2/2)/x\right)' = \exp(-x^2/2) + \exp(-x^2/2)/x^2 \geqslant \exp(-x^2/2)$$

et on conclut en intégrant. □

Preuve de la Proposition 6.3.2. Par définition de la distance W_1, on a

$$W_1(\mu, \gamma) = \sup_{h1-lip} \mathbb{E}_\mu[h] - \mathbb{E}_\gamma[h]$$

Pour toute fonction h 1-lipschitz, on substitue ensuite à $h - \mathbb{E}_\gamma[h]$ la fonction $f' - xf$ avec f la solution de l'équation différentielle donée par le lemme de régularisation. Comme cette solution vérifie les bornes $|f|, |f'| \leqslant 1$, et $|f''| \leqslant 4$, le résultat suit immédiatement. □

6.3.2 Méthode des paires échangeables et application au TCL

Dans cette section, on va voir comment utiliser le lemme de Stein pour donner une borne sur la vitesse de convergence dans le TCL.

Théorème 6.3.4. *Soit (W, W') une paire de variables aléatoires aléatoires réelles, de même loi, centrées et de variance égale à 1. Supposons de plus que il existe $\lambda \in (0,1)$ tel que*

$$\mathbb{E}[W' - W|W] = -\lambda W.$$

Alors si Z est une gaussienne centrée réduite,

$$W_1(W, Z) \leqslant \sqrt{\mathrm{Var}\left(\mathbb{E}\left[\frac{1}{2\lambda}(W'-W)^2|W\right]\right)} + \frac{2}{3\lambda}\mathbb{E}[|W'-W|^3]$$

Proof. Le but va être d'appliquer la Proposition 6.3.2. Soit f une fonction avec $|f|, |f'| \leqslant 1$, $|f''| \leqslant 4$. Posons

$$F(x) := \int_0^x f(y)dy.$$

On a alors, via un développement de Taylor

$$\begin{aligned}
0 &= \mathbb{E}[F(W') - F(W)] \\
&= \mathbb{E}\left[(W'-W)f(W) + \frac{1}{2}(W'-W)^2 f'(W) + R\right] \\
&= -\lambda\mathbb{E}[Wf(W)] + \frac{1}{2}\mathbb{E}[(W'-W)^2 f'(W)] + \mathbb{E}[R]
\end{aligned}$$

avec

$$|R| \leqslant \frac{1}{6}|f''|_\infty |W'-W|^3 \leqslant \frac{2}{3}|W'-W|^3.$$

On en déduit que

$$\begin{aligned}
\left|\mathbb{E}[f'(W) - Wf(W)]\right| &\leqslant \frac{1}{2\lambda}\mathbb{E}[((W'-W)^2 - 1)f'(W)] + \frac{2}{3\lambda}\mathbb{E}[|W'-W|^3] \\
&\leqslant \sqrt{\mathrm{Var}\left(\mathbb{E}\left[\frac{1}{2\lambda}(W'-W)^2|W\right]\right)} + \frac{2}{3\lambda}\mathbb{E}[|W'-W|^3]
\end{aligned}$$

où on a utilisé $\mathbb{E}[(W'-W)^2] = 2 - 2\mathbb{E}[W'W] = 2\lambda$. La Proposition 6.3.2 permet de conclure. □

A partir de ce résultat, on va pouvoir prouver le théorème suivant, qui est une variante du théorème de Berry-Essen sur la vitesse de convergence dans le TCL :

Théorème 6.3.5. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variable aléatoire i.i.d. centrées réduites, et $S_n := n^{-1/2}\sum_{i=1}^n X_i$. Alors*

$$W_1(S_n, Z) \leqslant \frac{16}{3\sqrt{n}}\mathbb{E}[|X_1|^3] + \frac{1}{2\sqrt{n}}\sqrt{\mathbb{E}[|X_1|^4]}.$$

où Z est une variable gaussienne centrée réduite.

Il est possible de prouver une version plus forte de ce théorème, où il faut seulement une borne sur le moment d'ordre 3. Si le moment d'ordre 3 est infini, la vitesse de convergence est en général plus lente qu'en $1/\sqrt{n}$.

Proof. Le but va être d'appliquer le Théorème 6.3.4, il nous faut donc construire une paire pour S_n qui vérifie les hypothèses du théorème. Nous allons la construire en choisissant uniformément au hasard l'un des éléments de la somme, et le remplacer par une copie indépendante.

Soit X' une copie indépendante des X_i, et I un élément de $\{1, ..., n\}$ choisi uniformément au hasard, indépendamment des X_i et de X'. On pose

$$S'_n := S_n + \frac{X'}{\sqrt{n}} - \frac{X_I}{\sqrt{n}}.$$

Par symétrie, (S_n, S'_n) est une paire échangeable, et de plus

$$\mathbb{E}[S'_n - S_n | S_n] = \mathbb{E}\left[\frac{X'}{\sqrt{n}} - \frac{X_I}{\sqrt{n}} \mid S_n\right] = -\frac{1}{n} S_n.$$

Donc les hypothèses du Théorème 6.3.4 sont vérifiées avec $\lambda = 1/n$. De plus

$$\mathbb{E}[|S_n - S'_n|^3] = n^{-3/2}\mathbb{E}[|X_1 - X'_1|^3] \leqslant 8\mathbb{E}[|X_1|^3]$$

et

$$\mathrm{Var}(\mathbb{E}[n(S'_n - S_n)^2 | S_n]) = \mathrm{Var}\left(\frac{1}{n}\mathbb{E}[\sum X_i^2 | S_n]\right) \leqslant \mathrm{Var}\left(\frac{1}{n}\sum X_i^2\right) \leqslant \frac{\mathbb{E}[X_1^4]}{n}.$$

La conclusion suit immédiatement.

Dans cette dernière étape, on a utilisé le fait que pour toute variable aléatoire rélle Y et sous-tribu $\mathcal{F}$, on a

$$\mathrm{Var}(Y) = \inf_c \mathbb{E}[(Y - c)^2] \geqslant \inf_c \mathbb{E}[(\mathbb{E}[Y - \mathcal{F}] - c)^2] = \mathrm{Var}(\mathbb{E}[Y|\mathcal{F}]).$$

□

Exercice 6.3.1. *Quelles bornes obtient-t-on pour la convergnce d'une somme de variables de Bernoulli de paramètre p?*

6.4 TCL martingale

Dans cette section, nous allons voir un TCL dans le cas où les variables sont des incréments de martingales (et en particulir, peuvent ne pas être indépendantes).

6.4.1 Théorème principal

Théorème 6.4.1 (TCL Martingale)**.** *Soit $(k_n)_{n\in\mathbb{N}}$ une suite croissante d'entiers naturels, $(\mathcal{F}_j^n)_{j\leqslant k_n, n\in\mathbb{N}}$ une famille de sous-tribus de $\mathcal{F}$, telles que pour tout n fixé la suite $(\mathcal{F}_j^n)_{j\leqslant k_n}$ soit une filtration par rapport à la variable j. Pour tout n, on se donne une martingale $(M_j^n)_{j\leqslant k_n}$ par rapport à la filtration $(\mathcal{F}_j^n)$ (toujours par rapport à la variable j), et avec $M_0^n = 0$ pour tout n. On suppose que*

1. $\mathbb{E}[\max_{j\leqslant k_n - 1} |M_{j+1}^n - M_j^n|] \underset{n\to\infty}{\longrightarrow} 0$;

2. $\sum_{j=1}^{k_n}(M_j^n - M_{j-1}^n)^2$ *converge en probabilité vers la constante* 1.

Alors $M_{k_n}^n$ *converge en loi vers une gaussienne centrée réduite.*

La preuve utilisera le lemme suivant :

Lemme 6.4.2. *Soit* (U_n) *et* (V_n) *deux suites de variables aléatoires. Supposons que*

1. $(U_n)_{n\in\mathbb{N}}$ *converge en probabilité vers une constante* a*;*
2. *Les suites* $(V_n)_{n\in\mathbb{N}}$ *et* $(U_nV_n)_{n\in\mathbb{N}}$ *sont uniformément intégrables;*
3. $\mathbb{E}[V_n] \longrightarrow 1$.

Alors $\mathbb{E}[U_nV_n] \longrightarrow a$.

Proof. On a $\mathbb{E}[U_nV_n] = a\mathbb{E}[V_n] + \mathbb{E}[V_n(U_n - a)]$, donc il suffit de montrer que $\mathbb{E}[V_n(U_n - a)]$ tend vers 0. Comme $(V_n(U_n - a))_{n\in\mathbb{N}}$ est uniformément intégrable, car somme de variables uniformément intégrables, il suffit de montrer la convergence en probabilité vers 0.

Or pour tout $K > 0$, on a

$$\mathbb{P}(|V_n(U_n - a)| > \varepsilon) \leqslant \mathbb{P}(|U_n - a| > \varepsilon/K) + \mathbb{P}(|V_n| > K).$$

Comme $(V_n)_{n\in\mathbb{N}}$ est uniformément intégrable, on peut choisir K pour rendre le second terme arbitrairment petit. Et comme $(U_n)_{n\in\mathbb{N}}$ converge converge en probabilité vers a, le premier terme tend vers 0, et donc on a bien

$$\limsup_n \mathbb{P}(|V_n(U_n - a)| > \varepsilon) = 0 \ \ \forall \varepsilon > 0,$$

ce qui conclut la preuve. □

Preuve du Théorème 6.4.1. Nous allons faire la preuve sous l'hypothèse supplémentaire

$$\forall n,\ \sum_{j=1}^{k_n}(M_j^n - M_{j-1}^n)^2 \leqslant 2. \tag{6.5}$$

Le cas général se traite à partir de celui là, avec une procédure de cutoff supplémentaire, qu'on omet ici. Le but va être de montrer la convergence de la fonction caractéristique de $M_{k_n}^n$ vers $\exp(-t^2/2)$, et d'appliquer le théorème de Lévy. Tout d'abord, via un développement de Taylor, on peut écrire

$$\exp(ix) = (1 + ix)\exp(-x^2/2 + r(x)); \quad |r(x)| < |x|^3.$$

Alors, en posant $Z_j^n = M_j^n - M_{j-1}^n$, on a

$$\mathbb{E}[\exp(itM_{k_n}^n)] = \left(\prod_{j\leqslant k_n}(1 + itZ_j^n)\right)\exp\left(-t^2\sum_{j=1}^{k_n}(Z_j^n)^2 + \sum_{j=1}^{k_n} r(tZ_j^n)\right).$$

On souhaite appliquer le Lemme 6.4.2 avec

$$V_n := \prod_{j\leqslant k_n}(1 + itZ_j^n);\ \ U_n := \exp\left(-t^2\sum_{j=1}^{k_n}(Z_j^n)^2 + \sum_{j=1}^{k_n} r(tZ_j^n)\right).$$

Pour cela, montrons que les hypothèses du Lemme 6.4.2 sont vérifiées.

1. Par hypothèse, $\sum_{j=1}^{k_n}(Z_j^n)^2$ converge en probabilité vers 1, donc il suffit de montrer que $\sum_{j=1}^{k_n} r(tZ_j^n)$ converge en probabilité vers 0. On a

$$\begin{aligned}\left|\sum_{j=1}^{k_n} r(tZ_j^n)\right| &\leqslant |t|^3 \sum_{j=1}^{k_n} |Z_j^n|^3 \\ &\leqslant |t|^3 \max_{j\leqslant k_n} |Z_j^n| \sum_{j=1}^{k_n} |Z_j^n|^2 \\ &\leqslant 2|t|^3 \max_{j\leqslant k_n} |Z_j^n|.\end{aligned}$$

   Cette variable converge vers 0 dans L^1, et donc en probabilité. On en déduit que U_n converge en probabilité vers $\exp(-t^2/2)$.

2. Comme la suite $(U_nV_n)_{n\in\mathbb{N}}$ est bornée (en module) par 1, elle est uniformément intégrable. Pour l'uniforme intégrabilité de V_n, grâce à l'inégalité $|1+ix|^2 \leqslant \exp(x^2)$, on a

$$|V_n| \leqslant \exp\left(\frac{1}{2}t^2 \sum_{j=1}^{k_n}(Z_j^n)^2\right) \leqslant \exp(t^2),$$

   ce qui nous suffit.

3. Enfin, par la propriété de martingale,

$$\begin{aligned}\mathbb{E}[V_n] &= \mathbb{E}\left[\prod_{j\leqslant k_n}(1+itZ_j^n)\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{j\leqslant k_n}(1+itZ_j^n) \mid \mathcal{F}_{k_n-1}\right]\right] \\ &= \mathbb{E}\left[\prod_{j\leqslant k_n-1}(1+itZ_j^n)\mathbb{E}[(1+itZ_{k_n}^n) \mid \mathcal{F}_{k_n-1}]\right] = \mathbb{E}\left[\prod_{j\leqslant k_n-1}(1+itZ_j^n)\right].\end{aligned}$$

   Par récurrence, on obtient bien $\mathbb{E}[V_n] = 1$.

L'application du lemme et du théorème de Lévy permet donc de déduire le résultat sous l'hypothèse supplémentaire (6.5).

□

6.4.2 Application aux chaînes de Markov

Théorème 6.4.3. *Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov irréductible sur un espace fini $\mathcal{X}$, de mesure invariante π. Soit $f : \mathcal{X} \longrightarrow \mathbb{R}$ une fonction (qui n'est pas identiquement égale à 0) vérifiant $\mathbb{E}_\pi[f] = 0$. Alors sous $\mathbb{P}_\pi$ on a*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f(X_i) \xrightarrow{loi} \mathcal{N}(0,\sigma^2)$$

avec une variance $\sigma^2 > 0$ qui ne dépend que de f.

La preuve va utiliser le lemme suivant :

Lemme 6.4.4. *Soit Q la matrice de transition d'une chaîne de Markov irréductible sur un espace fini $\mathcal{X}$, et π sa mesure de probabilité invariante. Soit $f : \mathcal{X} \longrightarrow \mathbb{R}$ telle que $\mathbb{E}_\pi[f] = 0$. Alors il existe une fonction $h : \mathcal{X} \longrightarrow \mathbb{R}$ telle que $f = h - Qh$.*

Proof. Tout d'abord, on a

$$\mathbb{R}^{|\mathcal{X}|} = \mathrm{Ker}(I - Q^*) \oplus \mathrm{Im}(I - Q).$$

L'espace $\mathrm{Ker}(I - Q^*)$ est l'espace des mesures (signées) invariantes. On sait que comme la chaîne de Markov est irréductible sur un espace fini, les mesures invariantes forment un espace de dimension 1, et donc $\dim(\mathrm{Ker}(Id - Q^*)) = 1$. Comme $\mathbb{E}_\pi[f] = 0$, f est orthogonale à cet espace, et donc $f \in \mathrm{Im}(I - Q)$. □

Preuve du Théorème 6.4.3. Le but va être d'appliquer le TCL pour les martingales, via la décompositon de Doob-Meyer de $S_n := \sum_{i=1}^n f(X_i)$, qui va pouvoir être rendue plus explicite via la fonction h telle que $f = h - Qh$, donnée par le Lemme 6.4.4, et où Q est la matrice de transition de la chaîne de Markov.

On a

$$\frac{S_n}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n u(X_i) - Qu(X_i) = \frac{Qu(X_0) - Qu(X_n)}{\sqrt{n}} + \frac{1}{\sqrt{n}} \sum_{i=1}^n u(X_i) - Qu(X_{i-1}).$$

Or comme l'espace est fini, u est bornée, et donc $\frac{Qu(X_0)-Qu(X_n)}{\sqrt{n}}$ converge p.s. vers 0, donc la limite en loi de $S_/\sqrt{n}$ est la même que celle de $M_n/\sqrt{n}$, où

$$M_n = \sum_{i=1}^n u(X_i) - Qu(X_{i-1}).$$

Or M_n est une martingale, et on va pouvoir appliquer le TCL martingale. A cause de la normalisation en $\sqrt{n}$, comme u est bornée, le maximum des incréments converge vers 0, et donc il nous suffit de calculer la limite de $n^{-1} \sum (u(X_i) - Qu(X_{i-1})^2$. On va chercher à appliuer le théorème ergodique pour les chaînes de Markov. Tout d'abord, $Y_i = (X_i, X_{i+1})$ est une chaîne de Markov sur (un sous-ensemble de) $E \times E$, dont la matrice de transiton est

$$\tilde{Q}((a,b),(c,d)) = \mathbb{1}_{b=c} Q(b,d).$$

On vérifie que $\tilde{\pi}((a,b)) = \pi(a)Q(a,b)$ est une mesure de probabilité invariante. On applique le théorème ergodique avec la fonction $f(a,b) = (u(b) - Qu(a))^2$ à cette chaîne de Markov pour obtenir

$$\frac{1}{n} \sum (u(X_i) - Qu(X_{i-1})^2 \xrightarrow{p.s.} \mathbb{E}_\pi[(u(X_1) - Qu(X_0))^2.$$

On a, via un conditionnement,

$$\mathbb{E}_\pi[(u(X_1) - Qu(X_0))^2] = \mathbb{E}_\pi[u^2 - (Qu)^2] > 0.$$

On a donc convergence en loi vers une gaussienne centrée de variance $\sigma^2 = \mathbb{E}_\pi[u^2 - (Qu)^2]$. □

Exercice 6.4.1. *On considère la chaîne de Markov sur* $\{0,1\}$ *dont la matrice de transition est donnée par*

$$Q = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

avec $\alpha, \beta \in]0,1[$. *étudier le comportement asymptotique des fluctuations de la proportion de temps passé en* 0.

Solution 6.4.1. *La chaîne de Markov est irréductible sur un espace fini, donc récurrente. Un calcul montre que la mesure de probabilité invariante est donnée par* $(\beta/(\alpha+\beta), \alpha/(\alpha+\beta))$.

Si $f(x) = \mathbb{1}_{x=0}$, *alors* $\mathbb{E}_\pi[f] = \beta/(\alpha+\beta)$. *La résolution de l'équation* $f - \mathbb{E}_\pi[f] = u - Qu$ *donne* $u = (0, -(\alpha+\beta)^{-1})$. *La variance limite dans le TCL pour les fluctuations de* $n^{-1/2}\sum f(X_i)$ *est donnée par*

$$\mathbb{E}_\pi[u^2 - (Qu)^2] = \frac{\alpha}{(\alpha+\beta)^3} - \frac{\beta\alpha^2 + \alpha(1-\beta)^2}{(\alpha+\beta)^3} = \frac{\alpha\beta(\alpha-\beta+2)}{(\alpha+\beta)^3}.$$

Chapter 7

Introduction aux inégalités de concentration et aux grandes déviations

7.1 Premiers exemples

On s'intéresse à la probabilité observer une déviation significative dans la loi des grandes nombres pour une somme de variables iid. Le TCL régit déjà la probabilité de voir des déviations d'ordre $n^{1/2}$. Ici, on s'intéressera à la probabilité de voir des déviations d'ordre n, i.e. à des évènements de la forme

$$\mathbb{P}\left[\frac{1}{n}\sum X_i \geqslant \mathbb{E}[X_1] + r\right]; \;\; r > 0.$$

Commençons par regarder le cas le plus simple, celui de variables gaussiennes. Si les X_i sont des gaussiennes centrées réduites iid, alors $\frac{1}{n}\sum X_i$ suit une loi $\mathcal{N}(0, n^{-1})$, et donc

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum X_i \geqslant \mathbb{E}[X_1] + r\right] &= \int_r^\infty \frac{\sqrt{n}\exp(-nt^2/2}{\sqrt{2\pi}}dt \\ &= \int_{\sqrt{n}r}^\infty \frac{\exp(-t^2/2)}{\sqrt{2\pi}}dt \\ &\approx \frac{1}{\sqrt{n}r\sqrt{2\pi}}\exp(-nr^2/2).\end{aligned}$$

En particulier

$$\frac{1}{n}\log\mathbb{P}\left[\frac{1}{n}\sum X_i \geqslant r\right] \longrightarrow -\frac{r^2}{2}$$

pour tout $r > 0$. On peut montrer que cette asymptotique est encore vraie pour les déviations par en dessous par symétrie, i.e.

$$\frac{1}{n}\log\mathbb{P}\left[\frac{1}{n}\sum X_i \leqslant -r\right] \longrightarrow -\frac{r^2}{2}$$

L'échelle logarithmique s'avère être la bonne échelle pour étudier l'asymptotique de cette probabilité pour des variables générales. Toutefois, contrairement au TCL, la limite n'est pas universelle, mais dépend des détails de la loi des X_i.

Commençons par regarder un deuxième exemple, celui des variables de Bernoulli symétriques. On a

$$\mathbb{P}(S_n \geqslant xn) = \frac{1}{2^n} \sum_{k \geqslant xn} \binom{n}{k}.$$

Pour $x > 1$, on a trivialement

$$\lim \frac{1}{n} \log \mathbb{P}(S_n \geqslant xn) = -\infty.$$

Prenons maintenant $x \in [1/2, 1]$. Comme les coefficients binomiaux sont décroissants pour $k \geqslant n/2$, pour $\ell = \lfloor xn \rfloor$ on a

$$\frac{1}{2^n} \binom{n}{\ell} \leqslant \mathbb{P}(S_n \geqslant xn) \leqslant \frac{n+1}{2^n} \binom{n}{\ell}.$$

Or, par la formule de Stirling,

$$\begin{aligned}
\lim_n \frac{1}{n} \log \binom{n}{\ell} &= \lim_n \frac{1}{n} (\log n! - \log \ell! - \log(n-\ell)!) \\
&= \lim_n \log n - 1 - \frac{\ell}{n} \log \ell + \frac{\ell}{n} - \frac{n-\ell}{n} \log(n-\ell) + \frac{n-\ell}{n} \\
&= \lim_n -\frac{\ell}{n} \log \left(\frac{\ell}{n} \right) - \frac{n-\ell}{n} \log \left(\frac{n-\ell}{n} \right) \\
&= -x \log x - (1-x) \log(1-x).
\end{aligned}$$

On en déduit que

$$\lim_n \frac{1}{n} \log \mathbb{P}(S_n \geqslant xn) = -x \log x - (1-x) \log(1-x) - \log 2.$$

Par symétrie, on a pour $x \in (0, 1/2)$

$$\lim_n \frac{1}{n} \log \mathbb{P}(S_n \leqslant xn) = -x \log x - (1-x) \log(1-x) - \log 2.$$

On voit donc que cette limite est différente pour les variables gaussiennes et pour les variables de Bernoulli. La question est donc de comprendre comment calculer cette limite pour des variables plus générales. Mais pour commencer, on va regarder de soutils pour donner de sbornes exponentiellement petites sur des probabilités d'évènements dans différentes situations.

7.2 Inégalité de Chernoff et conséquences

7.2.1 Inégalité de Chernoff

Proposition 7.2.1 (Inégalité de Chernoff)**.**

$$\mathbb{P}[f(X) \geqslant r] \leqslant \inf_{\lambda > 0} e^{-\lambda r} \mathbb{E}[e^{\lambda f(X)}].$$

Exercice 7.2.1. *Utiliser l'inégalité de Markov pour démontrer cette inégalité.*

Regardons déjà ce que donne cette inégalité pour une variable gaussienne. Soit X de loi $\mathcal{N}(0,\sigma^2)$. Alors

$$\mathbb{E}[\exp(\lambda X)] = \exp(\lambda^2\sigma^2/2).$$

Appliquer l'inégalité de Chernoff nous donne alors

$$\mathcal{P}[X \geqslant r] \leqslant \inf_{\lambda} \exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda r\right) = \exp(-r^2/(2\sigma^2)).$$

Si on reprend le calcul de la section précédente, on ale développement asymptotique

$$\mathcal{P}(X \geqslant r) \equiv \frac{1}{r\sigma\sqrt{2\pi}} \exp(-r^2/(2\sigma^2)).$$

La borne donnée par l'ingalité de Chernoff est donc déjà très précise, pusiqu'elle ne peut pas être améliorée en $\exp(-(1+\varepsilon)r^2/(2\sigma^2))$.

7.2.2 Inégalité de Hoeffding

Théorème 7.2.2 (Inégalité de Hoeffding)**.** *Soit (X_i) une suite de v.a. i.i.d. centrée telles que $|X_i| \leqslant A$ et $S_n = \sum_{i=1}^n X_i$. Alors*

$$\mathbb{P}\left(S_n \geqslant r\right) \leqslant \exp\left(-\frac{r^2}{2nA^2}\right).$$

Si on regarde des écarts dans la loi des grands nombres, c'est à dire $\mathbb{P}(S_n/n \geqslant r)$, ca nous donne une borne en $\exp(-cnr^2)$. En particulier, les fluctations observables sont au plus d'ordre $1/\sqrt{n}$, en accord avec le TCL.

Lemme 7.2.3. *Soit X une variable aléatoire à valeurs dans $[0,1]$ et d'espérance $p \in [0,1]$. Soit $\varpi : [0,1] \longrightarrow \mathbb{R}$ une fonction convexe. Alors*

$$\mathbb{E}[\varphi(X)] \leqslant p\varphi(1) + (1-p)\varphi(0) = \mathbb{E}[\varphi(U)]$$

où U est une variable de Bernoulli de paramètre p

Ce lemme nous dit que parmi les variables aléatoires de moyenne donnée et à support compact, celle dont les fluctuations (par exemple la variance) sont les plus grandes est celle qui prend uiquement les valeurs les plus éloignées possibles.

Proof. Par convexité, on a $\varphi(X) \leqslant X\varphi(1) + (1-X)\varphi(0)$. On obtient le résultat en prenant l'espérance de cette inégalit. □

Exercice 7.2.2. *Utiliser ce lemme pour montrer que si X est une variable de moyenne nulle et à valeurs dans $[-A,B]$, alors pour toute fonction convexe φ, on a*

$$\mathbb{E}[\varphi(X)] \leqslant \frac{B}{A+B}\varphi(-A) + \frac{A}{A+B}\varphi(B).$$

Preuve de l'inégalité de Hoeffding. Tout d'abord, en appliquant le lemme 7.2.3, on a

$$\mathbb{E}[\exp(\lambda X_1)] \leqslant \cosh(\lambda A) \leqslant \exp(\lambda^2 A^2/2).$$

Par indépendance, on a alors

$$\mathbb{E}[\exp(\lambda S_n)] = \mathbb{E}[\exp(\lambda X_1)]^2 \leqslant \exp(n\lambda^2 A^2/2).$$

On utilise ensuite l'inégalité de Chernoff pour conclure. □

On peut généraliser cette inégalité de la manière suivante :

Théorème 7.2.4 (Inégalité de Azuma-Hoeffding). *Soit M_n une martingale issue de 0. On suppose que pour tout k on a $|M_{k+1} - M_k| \leqslant \sigma_k$. Alors*

$$\mathbb{P}\left(M_n \geqslant r\right) \leqslant \exp\left(-\frac{r^2}{2\sum_{k=1}^n \sigma_k^2}\right).$$

Cette inégalité généralise celle de Hoeffding, car une somme de variables aléatoires centrées est une martingale.

Proof. On a toujours

$$\mathbb{E}[\exp(\lambda(M_{k+1} - M_n))] \leqslant \exp(\lambda^2\sigma_k^2/2)$$

car les incréments sont bornés. On a aussi

$$\begin{aligned}\mathbb{E}[\exp(\lambda M_n)] &= \mathbb{E}\left[\mathbb{E}\left[\exp(\lambda M_{n-1} + \lambda(M_n - M_{n-1})) \mid \mathcal{F}_n\right]\right] \\ &= \mathbb{E}\left[\exp(\lambda M_{n-1})\mathbb{E}\left[\exp(\lambda(M_n - M_{n-1})) \mid \mathcal{F}_n\right]\right].\end{aligned}$$

On souhaite majorer $\mathbb{E}\left[\exp(\lambda(M_n - M_{n-1})) \mid \mathcal{F}_n\right]$. Comme

$$M_n - M_{n-1} = \left(\frac{M_n - M_{n-1}}{2\sigma_k} + \frac{1}{2}\right)\sigma_k + \left(1 - \frac{M_n - M_{n-1}}{2\sigma_k} - \frac{1}{2}\right)(-\sigma_k),$$

on a par convexité

$$\exp(\lambda(M_n - M_{n-1})) \leqslant \left(\frac{M_n - M_{n-1}}{2\sigma_k} + \frac{1}{2}\right)\exp(\lambda\sigma_k) + \left(1 - \frac{M_n - M_{n-1}}{2\sigma_k} - \frac{1}{2}\right)\exp(-\lambda\sigma_k).$$

En prenant l'espérance conditonnelle, on obtient

$$\mathbb{E}\left[\exp(\lambda(M_n - M_{n-1})) \mid \mathcal{F}_n\right] \leqslant \cosh(\lambda\sigma_k) \leqslant \exp(\lambda^2\sigma_k^2/2).$$

On en déduit par récurrence que

$$\mathbb{E}[\exp(\lambda M_n)] \leqslant \exp\left(\frac{\lambda^2\sum_1^n \sigma_k^2}{2}\right)$$

et on conclut en appliquant l'inégalité de Chernoff. □

Exercice 7.2.3. *On reprend l'exercice 6.4.1. Borner de manière non-asymptotique la probabilité que la chaîne de Markov (avec condition initiale de loi π) passe au moins $n\frac{\beta+\varepsilon}{\alpha+\beta}$ fois par 0 parmis les n premiers pas.*

Solution 7.2.1.

7.3 Concentration gaussienne

7.3.1 Inégalité de concentration gaussienne

Théorème 7.3.1. *Soit $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ une fonction 1-lipschitz et X un vecteur gaussien centré réduit dans $\mathbb{R}^d$ (i.e. dont la matrice de covariance est l'identité). Alors*

$$\mathbb{E}[\exp(\lambda f(X))] \leqslant \exp(\lambda\mathbb{E}[f(X)] + \lambda^2/2). \tag{7.1}$$

En conséquence,

$$\mathbb{P}(f(X) \geqslant \mathbb{E}[f(X)] + r) \leqslant \exp(-r^2/2).$$

La borne (7.1), est optimale, pusiqu'on a égalité pour $f(x) = x_1$. Le fait que la constante soit indépendante de la dimension est à l'origine de nombreuses applications en statistique, mais aussi en physique mathématique, en géométrie et en sciences des données.

Proof. On va seulement prouver une version non-optimale, à savoir

$$\mathbb{E}[\exp(\lambda f(X))] \leqslant \exp(\lambda \mathbb{E}[f(X)] + \lambda^2 \pi^2/8),$$

et on suppose pour simplifier que f est C^1. Soit X et Y deux variables indépendantes de loi $\mathcal{N}(0, I_d)$. Sans perdre de généralité, on peut supposer $\mathbb{E}[f(X)] = 0$. Les variables

$$X_t = cos(t\pi/2)X + \sin(t\pi/2)Y; \quad Y_t = -\sin(t\pi/2)X + \cos(t\pi/2)Y$$

suivent aussi une loi $\mathcal{N}(0, I_d)$ et sont indépendantes. On a

$$f(Y) - f(X) = \int_0^1 \frac{d}{dt} f(X_t) dt = \frac{\pi}{2} \int_0^1 Y_t \cdot \nabla f(X_t) dt.$$

D'où

$$\begin{aligned} \mathbb{E}\left[\exp(\lambda f(Y) - \lambda f(X))\right] &= \mathbb{E}\left[\exp\left(\lambda \frac{\pi}{2} \int_0^1 Y_t \cdot \nabla f(X_t) dt\right)\right] \\ &\leqslant \int_0^1 \mathbb{E}\left[\exp\left(\lambda \frac{\pi}{2} Y_t \cdot \nabla f(X_t)\right)\right] dt \end{aligned}$$

Comme X_t et Y_t sont indépendantes, conditionnellement à X_t la variable $Y_t \cdot \nabla f(X_t)$ suit une loi gaussienne de variance $|\nabla f(X_t)|^2 \leqslant 1$. Donc

$$\int_0^1 \mathbb{E}\left[\exp\left(\lambda \frac{\pi}{2} Y_t \cdot \nabla f(X_t)\right)\right] dt \leqslant \exp\left(\frac{\lambda^2 \pi^2}{8}\right).$$

Donc $\mathbb{E}\left[\exp(\lambda f(Y) - \lambda f(X))\right]$. Par indépendance, ca nous donne

$$\mathbb{E}\left[\exp(\lambda f(Y))\right] \leqslant \exp\left(\frac{\lambda^2 \pi^2}{8}\right) \mathbb{E}[\exp(-\lambda f(X))]^{-1}.$$

On conclut en utilisant que $\mathbb{E}[f(X)] = 0$ et l'inégalité de Jensen pour avoir $\mathbb{E}[\exp(-\lambda f(X))] \geqslant 1$.

□

Exercice 7.3.1. *Soit f une fonction 1-lipschitz, et (X_i) une suite de gaussiennes centrées réduites indépendantes. Montrer que*

$$\mathbb{P}\left[\frac{1}{n} \sum_1^n f(X_i) \geqslant \mathbb{E}[f(X_1)] + r\right] \leqslant \exp(-nr^2/2).$$

Comparer avec le TCL.

7.3.2 Maximum de gaussiennes indépendantes

On considère des gaussiennes centrées réduites réelles indépendantes, qu'on notera X_i.

La fonction $(x_1, ..., x_n) \longrightarrow \max x_i$ est 1-lipschitz pour tout n. On a donc

$$\mathbb{P}(\max(X_1..., X_n) \geqslant \mathbb{E}[\max(X_i)] + r) \leqslant \exp(-r^2/2)$$

pour tout n. Or $\mathbb{E}[\max(X_1, ..., X_n)] \longrightarrow +\infty$. Le maximum d'un grand nombre de gaussiennes indépendantes va donc être fortement concentré autour de sa moyenne, c'est à dire

$$\mathbb{P}(|\max(X_1, ..., X_n) - \mathbb{E}[\max(X_1, ..., X_n)]| \geqslant \varepsilon \mathbb{E}[\max(X_1, ..., X_n)]) \longrightarrow 0$$

pour tout $\varepsilon > 0$.

Pour estimer $\mathbb{E}[\max(X_1, ..., X_n)]$, on peut encore utiliser des controles sur les moments exponentiels. Soit $\lambda > 0$. On a

$$\begin{aligned}
\mathbb{E}[\max(X_1, ..., X_n)] &\leqslant \lambda^{-1} \log \mathbb{E}[\exp(\lambda \max(X_1, ..., X_n))] \\
&\leqslant \lambda^{-1} \log \left(\sum_1^n \mathbb{E}[\exp(\lambda X_i)] \right) \\
&\leqslant \frac{2 \log n + \lambda^2}{2\lambda}.
\end{aligned}$$

La valeur optimale est atteinte pour $\lambda = \sqrt{2 \log n}$, ce qui donne

$$\mathbb{E}[\max(X_1, ..., X_n)] \leqslant \sqrt{2 \log n}.$$

Cette borne est en fait asymptotiquement optimale (i.e. lorsque n tend vers l'infini).

A noter que cette borne n'est pas spécifique au cas gaussien, elle est vraie pour toute famille de variables (y compris non-indépendantes) telle que $\mathbb{E}[\exp(\lambda X)] \leqslant \exp(\lambda^2/2)$. On appelle de telles variables des variables sous-gaussiennes. Par contre, la borne sur le maximum n'ets plus optimale en générale (par exemple, elle est très sous-optimale pour des variables bornées).

7.3.3 Lemme de Johnson-Lindenstrauss et compression

Le probleme de la compression est de trouver une manière efficace de plonger N points dans un espace de grande dimension (qu'on peut toujours prendre égale à N) dans un espace de dimension plus petite, sans trop déformer les distances entre les points. Le résultat principal que nous allons démontrer ici est

Théorème 7.3.2 (Lemme d'applatissement de Johnson-Lindenstrauss). *Soit $N \in \mathbb{N}$, $\varepsilon \in (0, 1)$ et T un ensemble de N points de $\mathbb{R}^N$. Alors pour tout $n > \frac{6 \log(2N^2)}{\varepsilon^2}$ il existe une application linéaire $A : \mathbb{R}^N \longrightarrow \mathbb{R}^n$ telle que*

$$\forall x, y \in T, \ \ (1 - \varepsilon)||x - y||_2 \leqslant ||Ax - Ay||_2 \leqslant (1 + \varepsilon)||x - y||_2. \tag{7.2}$$

Proof. La méthode de preuve qu'on va utiliser, parfois connue sous le nom de méthode probailiste (terminologie utilisée en théorie des graphes), sera de choisir A au

hasard suivant une loi donnée, et montrer qu'on a une probabilité positive de vérifier (7.2).

Soit $n \in \mathbb{N}$ qu'on choisira plus tard. Soit B une matrice $n \times N$ avc $B_{i,j} = \frac{1}{\sqrt{n}} g_{i,j}$, où les $g_{i,j}$ sont des gaussiennes i.i.d. centrées réduites. Alors pour tout $u \in \mathbb{R}^N$ tel que $||u|| = 1$, Bu est un vecteur gaussien de $\mathbb{R}^n$ centré, dont la matrice de covariance est l'identité. Comme la norme euclidienne est 1-lipschitzienne, on a

$$\mathbb{P}(|||Bu||_2 - \mathbb{E}[||Bu||]| \geqslant r) = \mathbb{P}(||Bu||_2 - \mathbb{E}[||Bu||] \geqslant r) + \mathbb{P}(||Bu||_2 - \mathbb{E}[||Bu||] \leqslant -r) \leqslant 2\exp(-r^2/2).$$

Soit $m = \mathbb{E}[||X||]$ l'espérance de la norme d'un vecteur gaussien standard en dimension n et $A = \frac{1}{m} B$. On a pour $||u|| = 1$ et en prenant $r = \varepsilon m$

$$\mathbb{P}\left(|||Au|| - 1| \geqslant \varepsilon\right) \leqslant 2\exp(-\varepsilon^2 m^2/2).$$

En particulier, pour tout $x, y \in T$, on a

$$\mathbb{P}\left(\left|\frac{||A(x-y)||}{||x-y||} - 1\right| \geqslant \varepsilon\right) \leqslant 2\exp(-\varepsilon^2 m^2/2).$$

On en déduit que

$$\mathbb{P}\left(\exists x, y \in T \text{ t.q. } \left|\frac{||A(x-y)||}{||x-y||} - 1\right| \geqslant \varepsilon\right) \leqslant 2N^2 \exp(-\varepsilon^2 m^2/2).$$

Il suffit donc de prendre n tel que

$$m^2 > \frac{2\log(2N^2)}{\varepsilon^2}$$

pour qu'un A vérifiant (7.2) existe.

Comme on a pour X vecteur gaussien standard

$$n = \mathbb{E}[||X||^2] \leqslant \mathbb{E}[||X||]^{2/3} \mathbb{E}[||X||^4]^{1/3} \text{ et } \mathbb{E}[||X||^4] \leqslant 3n^2,$$

on a $m^2 \geqslant n/3$, donc il suffit de prendre $n > \frac{6\log(2N^2)}{\varepsilon^2}$ pour que ça marche. □

7.4 Théorème de Cramer sur $\mathbb{R}$

Le but de cette section va être de généraliser les deux exemples de la Section 7.1 à des variables générales.

Définition 7.4.1. *La transformée de Legendre d'une fonction $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ est*

$$\varphi^*(y) := \sup_x xy - \varphi(x).$$

Exercice 7.4.1. *1. Montrer que la transformée de Legendre de $x \to x^2/2$ est encore $x \to x^2/2$.*

2. Montrer que la transformée de Legendre de la fonction $t \longrightarrow \log \mathbb{E}[\exp(tX)]$, avec X une variable de Bernoulli de paramètre p, est $x\log(x/p) + (1-x)\log((1-x)/(1-p))$ si $x \in (0,1)$, et $+\infty$ sinon.

Solution 7.4.1. *1. On peut vérifier que, pour y fixé, $\sup_x xy - x^2/2$ est atteint en $x = y$, et donc cette valeur est bien $y^2/2$.*

2. *Si X est une variable de Bernoulli de paramètre p, on a $\varphi(t) := \log \mathbb{E}[\exp(tX)] = \log(pe^t + (1-p))$. Si pour x fixé on pose $f(t) = tx - \log(pe^t + (1-p))$, on a $f'(t) = x - pe^t/(pe^t + 1 - p)$, et donc, pour $x \in (0,1)$,*

$$f'(t) = 0 \leftrightarrow pe^t(x-1) + (1-p)x = 0 \leftrightarrow t = \log\left(\frac{(1-p)x}{p(1-x)}\right).$$

On en déduit que

$$\begin{aligned}\sup_t tx - \log(pe^t + (1-p)) &= x\log\left(\frac{(1-p)x}{p(1-x)}\right) - \log\left(\frac{(1-p)x}{1-x} + 1 - p\right) \\ &= x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right).\end{aligned}$$

Au vu de ces calculs, les deux exemples de la Section 7.1 sont des cas particuliers du résultat suivant :

Théorème 7.4.2 (Théorème de Cramèr sur $\mathbb{R}$)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires i.i.d. On suppose que la fonction génératrice*

$$\varphi_X(t) := \log \mathbb{E}\left[e^{tX}\right]$$

est finie sur un intervalle ouvert contenant 0. Alors pour tout $x \geqslant \mathbb{E}[X]$, on a

$$\lim \frac{1}{n}\log \mathbb{P}\left[\sum_{i=1}^n X_i \geqslant nx\right] = -\varphi_X^*(x)$$

où φ_X^* est la transformée de Legendre de φ_X.

Pour pouvoir prouver ce théorème, on aura besoin des propriétés suivantes de la transformée de Legendre :

Exercice 7.4.2. 1. *Montrer que une transformée de Legendre est toujours une fonction convexe.*

2. *Montrer que si φ est C^2, strictement convexe minorée sur son domaine, t a pou limites $+\infty$ aux bords de son domaine, alors $\varphi^{**} = \varphi$.*

Solution 7.4.2. 1. *C'est un sup de fonctions affines, donc convexe.*

2. *Dans ce cas, pour tout $y \in \mathbb{R}$, la fonction $x \longrightarrow xy - \varphi(x)$ est strictement concave et majorée. Supposons qu'elle atteint son supremum en un point x. Alors en dérivant, on a $y = f'(x)$, et comme f' est inversible, $x = (f')^{-1}(y)$. Donc*

$$f^*(y) = y(f')^{-1}(y) - f((f')^{-1}(y)).$$

En particulier, en dérivant on voit que $(f^*)' = (f')^{-1}$. Ceci implique que la transformée de Legendre est une involution sur les fonctions strictement convexes C^2.

Ces résultats restent vrais si on suppose juste φ convexe, et rien de plus.

Théorème 7.4.3 (Théorème de Cramèr sur $\mathbb{R}$). *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires i.i.d. Soit*

$$\varphi_X(t) := \log \mathbb{E}\left[e^{tX}\right],$$

qu'on suppose finie dans un voisinage de 0. Alors pour tout $x \geqslant \mathbb{E}[X]$, on a

$$\lim \frac{1}{n} \log \mathbb{P}\left[\sum_{i=1}^{n} X_i \geqslant nx\right] = -\varphi_X^*(x)$$

où φ_X^* est la transformée de Legendre de φ_X.

Par symétrie, pour $x < \mathbb{E}[X]$, on a

$$\lim \frac{1}{n} \log \mathbb{P}\left[\sum_{i=1}^{n} X_i \leqslant nx\right] = -\varphi_X^*(x)$$

Sans perdre de généralité, on supposera $\mathbb{E}[X] = 0$. La preuve reposera sur les trois faits suivants :

1. Si on pose $s(x) = \sup_m \frac{1}{m} \log \mathbb{P}(\bar{X}_m \geqslant x)$, on a $\varphi_X = (-s)^*$ sur $\mathbb{R}_+$.
2. On a $\lim_n \frac{1}{n} \log \mathbb{P}(\bar{X}_n \geqslant x) = s(x)$;
3. s est une fonction concave.

Une fois ces trois résultats démontrées, on en déduit que $s = -\varphi^*$ en appliquant la transformée de Legendre à la 1ere assertion, modulo une subtilité sur le fait que la première identité est sur $\mathbb{R}_+$ et pas $\mathbb{R}$.

La preuve de la première partie utilisera l'inégalité de Chernoff (Proposition 7.2.1).

Exercice 7.4.3. *Utiliser l'inégalité de Chernoff pour montrer que $\varphi_X \geqslant (-s)^*$ sur $\mathbb{R}_+$.*

Solution 7.4.3. *Soit $t \geqslant 0$. En prenant le log de l'inégalité de Chernoff*

$$\begin{aligned}
\varphi_X(t) &= \frac{1}{n} \log \mathbb{E}[\exp(nt\bar{X}_n)] \\
&\geqslant \frac{1}{n} \log\left(\exp(nt\lambda)\mathbb{P}(\bar{X}_n \geqslant \lambda)\right) \\
&= t\lambda + \frac{1}{n} \log \mathbb{P}(\bar{X}_n \geqslant \lambda).
\end{aligned}$$

En optimsant sur λ, on obtient bien $\varphi_X(t) \geqslant (-s)^*(t)$ pour $t \geqslant 0$.

Prouvons maintenant l'inégalité inverse. L'identité $\varphi_X(0) = (-s)^*(0) = 0$ est triviale, on considèrera seulement les $t > 0$. Soit $K > 0$ fixé. On a

$$\begin{aligned}
\log \mathbb{E}\left[\exp(tX_1)\mathbb{1}_{|X_1|\leqslant K}\right] &= \frac{1}{n} \log \mathbb{E}\left[\exp(tn\bar{X}_n)\mathbb{1}_{|X_1|\leqslant K} \cdots \mathbb{1}_{|X_n|\leqslant K}\right] \\
&\leqslant \frac{1}{n} \log \mathbb{E}\left[\exp(tn\bar{X}_n)\mathbb{1}_{|\bar{X}_n|\leqslant K}\right] \\
&= \frac{1}{n} \log \mathbb{E}\left[\left(\exp(-nK) + \int_{-K}^{\bar{X}_n} nt\exp(ntu)du\right)\mathbb{1}_{|\bar{X}_n|\leqslant K}\right] \\
&= \frac{1}{n} \log\left(\exp(-nK) + \int_{-\infty}^{+\infty} \mathbb{E}[\mathbb{1}_{-K\leqslant u\leqslant \bar{X}_n}\mathbb{1}_{|\bar{X}_n|\leqslant K}]nt\exp(ntu)du\right)
\end{aligned}$$

où on a utilisé le théorème de Fubini à la dernière étape. Comme de plus

$$\mathbb{E}[\mathbb{1}_{-K\leqslant u\leqslant \bar{X}_n}\mathbb{1}_{|\bar{X}_n|\leqslant K}] \leqslant \mathbb{E}[\mathbb{1}_{\leqslant u\leqslant \bar{X}_n}\mathbb{1}_{|u|\leqslant K}] = \mathbb{P}[\bar{X}_n \leqslant u]\mathbb{1}_{|u|\leqslant K} \leqslant \exp(ns(u))\mathbb{1}_{|u|\leqslant K}$$

d'après l'inégalité de Chernoff, on a alors

$$\begin{aligned}\log \mathbb{E}\left[\exp(tX_1)\mathbb{1}_{|X_1|\leqslant K}\right] &\leqslant \frac{1}{n}\log\left(e^{-nK} + \int_{-K}^{K} ntu\exp(ntu+ns(u))du\right)\\ &\leqslant \frac{1}{n}\log\left(e^{-nK} + 2Knt\exp(n\sup_u tu + s(u))\right)\\ &\leqslant \frac{\log(1+2Kn)}{n} + \sup_u tu + s(u).\end{aligned}$$

On conclut en faisant tendre n vers l'infini, puis K vers l'infini. Ceci conclut la preuve de la première asserstion.

Pour la deuxième assertion, on procède avec un argument de sous-additivité. Il nous suffira de montrer que

$$\liminf_n \frac{1}{n}\log\mathbb{P}(\bar{X}_n \geqslant x) \geqslant \sup_m \frac{1}{m}\log\mathbb{P}(\bar{X}_m \geqslant x).$$

Fixons un entier m, et on décompose $n = q_n m + r_n$ avec $0 \leqslant r_m < m$. Si pour tout $i = 1, ...q_n$ on a

$$\frac{1}{m}\sum_{j=(i-1)m+1}^{im} X_j \geqslant x$$

et $X_j \geqslant x$ pour $j = q_n m+, ..., q_n m + r_n$, alors $\bar{X}_n \geqslant x$. Donc

$$\mathbb{P}(\bar{X}_n \geqslant x) \geqslant \mathbb{P}(\bar{X}_m \geqslant x)^{q_n}\mathbb{P}(X_{\geqslant}x)^{r_n}.$$

En prenant le log et en divisant par n, comme $q_n/n \longrightarrow 1/m$, on a

$$\liminf \frac{1}{n}\log\mathbb{P}(\bar{X}_n \geqslant x) \geqslant \frac{1}{m}\mathbb{P}(\bar{X}_m \geqslant x).$$

Comme m est arbitraire, la deuxième assertion suit.

Pour la troisème assertion, avec les mêmes arguments qu'avant, on a

$$\mathbb{P}(\bar{X}_{2n} \geqslant \frac{x+y}{2}) \geqslant \mathbb{P}(\bar{X}_n \geqslant x)\mathbb{P}(\bar{X}_n \geqslant y)$$

et donc

$$s((x+y)/2) \geqslant (s(x)+s(y))/2.$$

En itérant cet argument, on a alors $s(\alpha x + (1-\alpha)y) \geqslant \alpha s(x) + (1-\alpha)s(y)$ pour α de la forme $k/2^p$. On en déduit la même inégalité pour tout $\alpha \in [0,1]$ par approximation, en utilisant la monotonie de s.

On a alors montré que pour $t \geqslant 0 = \mathbb{E}[X]$, $\varphi_X(t) = (-s)^*(t)$. Pour les $x < 0$, par la loi des grands nombres, on a $s(x) = 0$. Donc il faut prolonger s par 0 sur $\mathbb{R}_-$. Ce prolongement est toujours concave, et n'empêche pas de retrouver s par application de la transformée de Legendre pour les valeurs positives, en utilisant le fait que $\varphi'_X(0) = \mathbb{E}[X] = 0$ (ce qui utilise l'hypothèse que φ_X est finie sur un voisinage de 0).

Exercice 7.4.4. *Calculer* $\lim n^{-1}\log\mathbb{P}(\bar{X}_n \geqslant x)$ *lorsque* $\bar{X}_n$ *est la moyenne empirique de variables iid de loi exponentielle, de paramètre* λ.

7.5 Principes de grandes déviations et théorème de Cramér multidimensionnel

Dans la formulation du théorème de Cramér sur $\mathbb{R}$, on utilise le caractère bien ordonné de la droite réelle. Dans un contexte plus général, il nous faut une autre formulation.

Définition 7.5.1. *Soit (E,d) un espace polonais. Une fonction $I : E \longrightarrow [0,+\infty]$ est une bonne fonction de taux si elle est semi-continue inférieurement et si ses ensembles de niveau sont compacts.*

Soit (a_n) une suite croissante de réels strictement positifs, et I une bonne fonction de taux sur E. Une suite de mesures de probabilité (μ_n) sur E vérifie un principe de grande déviations de vitesse (a_n) et de fonction de taux I si :

1. *pour tout fermé $F \subset E$, on a $\limsup a_n^{-1} \log \mu_n(F) \leqslant -\inf_F I$;*
2. *pour tout ouvert $O \subset E$, on a $\liminf a_n^{-1} \log \mu_n(O) \geqslant -\inf_O I$.*

Dans le cas du théorème de Cramér sur $\mathbb{R}$, la fonction de taux était convexe, donc continue, et donc le sup sur un intervalle $[x,+\infty[$ coincidait avec celui sur $]x,+\infty[$.

On va maintenant énoncer le théorème de Cramér multidimensionnel (qu'on ne démontrera pas ici). Pour cela, définissons a version multidiemsnionnel de la transormée de Legendre :

$$\varphi^*(y) = \sup_x x \cdot y - \varphi(x).$$

C'est encore une involution sur l'espace des fonctions convexes.

Théorème 7.5.2 (Théorème de Cramér multidimensionnel)**.** *Soit $\bar{X_n}$ la moyenne empirique de n variables iid à valeurs dans $\mathbb{R}^d$, et*

$$\varphi_X(y) := \log \mathbb{E}[\exp(y \cdot X_1)],$$

qu'on suppose finie sur un voisinage de l'origine. Alors la suites de lois des $\bar{X_n}$ vérifie un principe de grandes déviations, de vitesse n et de fonction de taux φ_X^*.

7.6 Théorème de Sanov

Définition 7.6.1 (Entropie)**.** *Soit μ une mesure positive sur un espace E. L'entropie relative par rapport à μ est la fonctionnelle sur $\mathcal{P}(E)$ définie par*

$$\mathrm{Ent}_\mu(\nu) := \begin{cases} \int \rho \log \rho d\mu \text{ si } \nu = \rho\mu; \\ +\infty \text{ si } \nu \text{ n'est pas absolutment continue par rapport à } \mu. \end{cases}$$

NB. *Les physicens ont plutôt l'habitude de définir l'entropie avec la convention de signe opposée, i.e. $-\int \rho \log \rho$. Cette notion est aussi connue sous le nom de divergence de Kullback-Leibler.*

Proposition 7.6.2 (Formules de dualité pour l'entropie). *Si μ et ν sont deux mesures de probailité, on a*

$$\mathrm{Ent}_\mu(\nu) = \sup_{f \text{ bornee}} \int f d\nu - \log \int e^f d\mu.$$

Conversement, pour toute fonction f, on a

$$\log \int e^f d\mu = \sup_\nu \int f d\nu - \mathrm{Ent}_\mu(\nu).$$

On peut interpréter ces formules comme des formules de dualité.

Proof. On démontrera seulement la première égalité, dans le cas où l'entropie est finie. C'est une conséquence de la dualité de Legendre pour la fonction exponentielle, dont la transformée de Legendre est $y \log y - y$ pour $y > 0$. Alors, en posant $\rho = d\nu/d\mu$, et en supposant sans perdre de généralité que $\int e^f d\mu = 1$, on a

$$\begin{aligned}\int f d\nu &\leqslant \int e^f d\mu + \int \rho \log \rho d\mu - \int \rho d\mu \\ &= 1 + \mathrm{Ent}_\mu(\nu) - 1.\end{aligned}$$

On a l'inégalité inverse en approcant $\log \rho$ avec une suite de fonctions bornées. □

Corollaire 7.6.3. *Soit f une fonction mesurable. Alors* $\mathrm{Ent}_{\mu \circ f^{-1}}(\nu \circ f^{-1}) \leqslant \mathrm{Ent}_\mu(\nu)$.

Proposition 7.6.4 (Inégalité de Pinsker). *Soit μ et ν deux mesures de probabilité sur un même espace. Alors*

$$d_{TV}(\mu, \nu) \leqslant \sqrt{\frac{\mathrm{Ent}_\mu(\nu)}{2}}.$$

Proof. On commence par le cas où l'espace est $\{0, 1\}$, et on peut voir μ et ν comme des lois de Bernoulli, de paramètres q et p. Alors

$$\mathrm{Ent}_\mu(\nu) = p \log(p/q) + (1-p) \log((1-p)/(1-q)) \geqslant 2(p-q)^2 = 2 d_{TV}(\mu, \nu)^2/2.$$

Donc l'ingálité de Pinsker est vraie sur un espace à deux points.

Pour le cas général, on suppose que ν a une densité f par rapport à μ, et on pose $A = \{x, f(x) > 1\}$. Soit $p = \nu(A)$ et $q = \mu(A)$. On a $d_{TV}(\mu, \nu) = 2|p - q|$ (comme vu dans la section 5.4.2), ce qui est aussi la distance en variation totale entre des lois de Bernoulli de paramètre p et q. Comme ces lois de Bernoulli sont les pushforward de μ et ν par la fonction $x \to \mathbb{1}_A(x)$, on a alors l'inégalité générale en combinant l'inégalitpour les lois de Bernoulli et le corollaire 7.6.3

□

Nous allons maintenant voir le théorème de Sanov, un renforcement fonctionnel du théorème de Cramér. On considère des variables aléatoires X_i dans un espace polonais (E, d), qu'on suppose indépendantes et identiquement distribuées, de loi μ. On va s'intéresser au comportement de la *mesure empirique* du système, c'est à dire la mesure de probabilité aléatoire

$$\mu_N := \frac{1}{N} \sum_{i=1}^N \delta_{X_i}.$$

C'est une variable aléatoire à valeurs dans $\mathcal{P}(E)$, qu'on peut aussi voir comme un espace polonais lorsqu'on le munit de la topologie de la convergence étroite. D'après la loi forte des grands nombres, elle converge p.s. vers la mesure μ. Le résultat suivant nous dit que la probabilité d'observer une mesure empirique anormale proche de $\nu \neq \mu$ se comporte comme $\exp(-N \operatorname{Ent}_\mu(\nu))$:

Théorème 7.6.5 (Théorème de Sanov)**.** *Soit (E,d) un espace polonais, $\mu \in \mathcal{P}(E)$ et (X_i) un suite de variables aléatoires iid de loi μ. La suite des lois des mesures empiriques associées à (X_i) vérifie un principe de grandes déviations de vitesse n et de fonction de taux* Ent_μ

On donne maintenant une ébauche de la preuve dans le cas particulier où E est un espace fini $\{a_1, ..., a_d\}$. Les valeurs possibles pour μ_N sont alors les mesures de la forme

$$\nu_{k,N} = \frac{k_1}{N}\delta_{a_1} + \frac{k_2}{N}\delta_{a_2} \cdots + \frac{k_1}{N}\delta_{a_d}$$

avec $k_1, ..., k_d$ des entiers naturels dont la somme vaut N. Il s'avère qu'elles ont toutes la même probabilité, égale à

$$\prod_{j=1}^{d} \mu(a_j)^{k_j} = \exp\left(N \sum \mu_N(a_j) \log \mu(a_j)\right). \tag{7.3}$$

Pour calculer la probabilité que μ_N soit égale à $\nu_{k,N}$, il faut donc dénombrer les configurations possibles des X_i qui donnent lieu à cette mesure empirique. Le lemme combinatoire suivant répond à cette question :

Lemme 7.6.6. *On a*

$$\frac{1}{(N+1)^d} \exp\left(-N \sum \left(\frac{k_j}{N}\right) \log\left(\frac{k_j}{N}\right)\right) \leqslant \left|\{(x_1, ..., x_N) \ t.q. \ N^{-1} \sum \delta_{x_j} = \nu_{k,N}\}\right| \leqslant \exp\left(-N \sum \left(\frac{k_j}{N}\right.\right.$$

En combinant cette borne et (7.3), on voit que

$$\frac{1}{(N+1)^d} \exp(-N \operatorname{Ent}_\mu(\nu_{k,N})) \leqslant \mathbb{P}\left(\mu_N = \nu_{k,N}\right) \leqslant \exp(-N \operatorname{Ent}_\mu(\nu_{k,N})),$$

et donc si $\nu_{k,N}$ approxime ν lorsque N tend vers l'infini, on a

$$\frac{1}{N} \log \mathbb{P}\left(\mu_N = \nu_{k,N}\right) \longrightarrow - \operatorname{Ent}_\mu(\nu).$$

Si on considère ensuite un ouvert O de $\mathcal{P}(E)$, pour tout $\nu \in O$ et pour N assez grand on peut trouver des $\nu_{k,N}$ dans O qui approximent ν. Alors

$$\frac{1}{N} \log \mathbb{P}(\mu_n \in O) \geqslant \frac{1}{N} \log \mathbb{P}(\mu_n = \nu_{k,N}) \longrightarrow - \operatorname{Ent}_\mu(\nu).$$

Si ensuite on prend le sup sur $\nu \in O$, on obtient la borne inf sur les ouverts.

Ensuite, pour F fermé de $\mathcal{P}(E)$, en notant P_N l'ensemble des valeurs possibles de μ_N, comme cet ensemble est de cardinal inférieur à $(N+1)^d$, on a

$$\frac{1}{N} \log \mathbb{P}(\mu_N \in F) \leqslant \frac{1}{N} \log \left(\sum_{\nu_{k,N} \in P_N \cap F} \mathbb{P}(\mu_N = \nu_{k,N}) \right) \leqslant \frac{1}{N} \log \left((N+1)^d \sup_{\nu \in F} \exp(-N \operatorname{Ent}_\mu(\nu)) \right)$$

et le résultat suit.

7.6.1 Application aux tests d'hypothèse

On a un échantillon $x = (x_1, .., x_N)$ obtenu à partir de variables iid de loi μ inconnue sur un alphabet fini. On cherche à tester l'hypothèse $\mu = \nu$. Un test possible est d'utiliser la fonction $T(x) = 1$ si $\mathrm{Ent}_\nu(\mu_x) < \delta$, et 0 sinon. On accepte l'hypothèse si $T(x) = 1$, et on la rejete sinon.

D'après le théorème de Sanov, si a_n est la probabilité d'erreur de type 1 (rejeter l'hypothèse à tort), on a l'asymptotique $\limsup n^{-1} \log a_n = \delta$. Donc le choix de δ nous permet d'ajuster cette erreur.

Il est encore plus naturel de faire dépendre δ de la taille de l'échantillon.

Exercice 7.6.1. *1. Montrer que de manière non-asymptotique et ppour δ fixé, la probabilité d'erreur de type* 1 *est majorée par* $(n+1)^k \exp(-n\delta)$. *En déduire une valeur de δ_n pour laquelle cette probabilité est inférieure à* $(n+1)^{-1}$.

2. Montrer que pour ce choix de valeur de δ_n, la probabilité b_n d'erreur de type 2 *(accepter l'hypothèse à tort) vérifie* $\lim n^{-1} \log b_n = \mathrm{Ent}_\mu(\nu)$.

Chapter 8

Bibliographie

Les chapitres 1 à 3 suivent largement les notes de cours de Jean-François Le Gall : https://www.imo.universite-paris-saclay.fr/ jflegall/IPPA2.pdf

Le traitement des théorèmes de convergence de martingales du chapitre 2 suit le poste de blog de Djalil Chafaï : https://djalil.chafai.net/blog/2020/10/02/convergence-of-discrete-time-martingales/

Le chapitre 5 est en partie inspiré par les notes de cours de Grégory Miermont : https://perso.ens-lyon.fr/gregory.miermont/thlim.pdf

Le traitement du TCL martingale du chapitre 6 est tiré des notes de cours de Sunder Sethuraman : https://www.math.arizona.edu/ sethuram/notes/wi_mart1.pdf

La preuve du théorème de Cramér au chapitre 7 est issue de l'article *A short proof of Cramér's theorem on* $\mathbb{R}$ de Raphaël Cerf et Pierre Petit.