Échauffement : deux algorithmes probabilistes

Les meilleurs algorithmes probabilistes (connus) sont souvent plus simples et/ou plus efficaces que les meilleurs algorithmes déterministes (connus). On va illustrer ce principe sur deux exemples, en utilisant le langage probabiliste («indépendance», «probabilité conditionnelle») qui sera introduit rigoureusement dans le prochain chapitre.

0.1 Vérifier la multiplication matricielle

Soient A, B, C trois matrices $n \times n$ à coefficients dans le corps $\mathbf{F}_2 = \{0, 1\}$. Le problème est de déterminer si l'équation $AB = C$ est vraie ou fausse.

Une première idée est de calculer le produit $A \cdot B$ et de vérifier si les coefficients sont les mêmes que ceux de C. L'algorithme naïf qui utilise la formule

$$(AB)_{ij} = \sum_k A_{ik} B_{kj}$$

a une complexité $\Theta(n^3)$. Des algorithmes plus sophistiqués basés sur une idée de Strassen améliorent la complexité en $\Theta(n^\alpha)$ pour $2 < \alpha < 3$ (le record actuel est $\alpha \approx 2,37$ et on conjecture que la valeur optimale est $\alpha = 2$).

Une autre idée est de vérifier la formule à travers le prisme probabiliste, c'est-à-dire de vérifier si l'équation

$$ABx = Cx$$

est satisfaite pour un vecteur $x \in \mathbf{F}_2^n$ choisi au hasard. Une telle vérification s'effectue en $\Theta(n^2)$, qui est clairement la complexité optimale de la multiplication matrice $\times$ vecteur. La clé est le lemme suivant.

Lemme. *Soit $D \in \mathsf{M}_n(\mathbf{F}_2)$ une matrice non nulle et $x \in \mathbf{F}_2^n$ choisi uniformément au hasard. Alors*

$$\mathbf{P}(Dx \neq 0) \geqslant 1/2.$$

Démonstration. Il existe un coefficient non nul dans la matrice D ; sans perte de généralité supposons que c'est le coefficient D_{1n}. On a alors

$$(Dx)_1 = \sum_{j=1}^{n} d_{1j} x_j = \sum_{j=1}^{n-1} d_{1j} x_j + x_n.$$

On remarque alors que quels que soient $(x_1, \ldots, x_{n-1})$ fixés, il y a probabilité $\frac{1}{2}$ (sur le choix de x_n) que $(Dx)_1 \neq 0$. □

Comme conséquence du lemme, on a le résultat suivant : si $AB \neq C$ et si $x \in \mathbf{F}_2^n$ est choisi au hasard, alors

$$\mathbf{P}(ABx = Cx) \leqslant \frac{1}{2}.$$

Si on répète 100 fois cette vérification pour des vecteurs $x_1, \ldots, x_{100}$ choisis indépendemment, on a

$$\mathbf{P}(ABx_i = Cx_i \text{ pour tout } i) \leqslant 2^{-100} = 0 \text{ en pratique}$$

et on obtient donc un algorithme probabiliste qui permet de vérifier la multiplication matricielle en temps $\Theta(n^2)$.

Cet argument repose implicitement sur le concept *d'indépendance* que l'on étudiera formellement plus tard.

0.2 Coupe minimale dans un graphe

Soit $G = (V, E)$ un graphe non orienté sans boucle, ayant possiblement des arêtes multiples. On pose $n = |V|$.

Une *coupe* de G est un sous-ensemble $C \subset E$ tel que $(V, E \setminus C)$ n'est pas connexe. Le problème est de déterminer le cardinal minimal d'une coupe de G, que l'on note $\text{mincut}(G)$. Autrement dit, on cherche une partition $V = V_1 \cup V_2$ (avec V_1 et V_2 non vides) qui minimise le nombre d'arêtes joignant un sommet de V_1 à un sommet de V_2. Il existe des algorithmes déterministes efficaces pour résoudre ce problème. Mais il y a plus simple : l'algorithme probabiliste de Karger (1993).

L'algorithme de Karger repose sur la notion de contraction d'un graphe selon une arête. Étant donnée une arête $e = \{x, y\} \in E$, la contraction de G selon e, notée G/e, est le graphe obtenu en identifiant les sommets x et y (pour obtenir un nouveau sommet noté xy), en remplaçant les arêtes $\{x, z\}$ ou $\{y, z\}$ par $\{xy, z\}$ et en effaçant les boucles éventuellement créées. Une contraction d'un graphe à n sommets peut être implémentée en temps $O(n)$, par exemple en représentant le graphe par sa matrice d'adjacence.

Figure 1 – Un graphe G (à gauche) et sa contraction G/e pour $e = \{a, b\}$ (à droite)

Pour toute arête e de G, on a $\text{mincut}(G) \leqslant \text{mincut}(G/e)$ puisque les coupes de G/e correspondent aux coupes de G qui n'utilisent pas l'arête e. L'algorithme de Karger consiste à effectuer des contractions au hasard.

Algorithme (Algorithme de Karger)**.** Tant que G contient > 2 sommets, répéter la procédure suivante : choisir uniformément au hasard une arête e de G et remplacer G par G/e. On obtient ainsi un graphe à 2 sommets qui correspond à une partition $V = V_1 \cup V_2$ et donc à une coupe du graphe initial.

Il est important de conserver les arêtes multiples : par exemple, si on l'applique l'algorithme au graphe qui est à droite de la figure 1, l'arête $\{ab, d\}$ est contractée avec probabilité $\frac{1}{5}$ puisque le graphe comprend 5 arêtes.

Dans la description de l'algorithme donnée ci-dessus, on considère implicitement que les différents choix aléatoires effectués par l'algorithme sont indépendants. Cette remarque vaut pour tous les algorithmes probabilistes étudiés dans ce cours.

Il est clair que l'algorithme termine puisque le nombre de sommets diminue de 1 à chaque étape. Le lemme-clé est le suivant.

Lemme. *La coupe C produite par l'algorithme de* Karger *vérifie*

$$\mathbf{P}(|C| = \text{mincut}(G)) \geqslant \frac{2}{n^2}.$$

Si on répète $N = 50n^2$ fois cet algorithme (tous les choix étant indépendants), et si on note k_i la coupe obtenue à la $i$ème exécution de l'algorithme, alors

$$\begin{aligned}\mathbf{P}\left(\min_{1\leqslant i\leqslant N} k_i \neq \text{mincut}(G)\right) &\leqslant \left(1-\frac{2}{n^2}\right)^N \\ &\leqslant \exp\left(-\frac{2N}{n^2}\right) \\ &= \exp(-100) \approx 0.\end{aligned}$$

On a donc un algorithme probabiliste de complexité $O(n^2T)$ pour trouver la coupe minimale d'un graphe, où $T = O(n^2)$ est la complexité d'une itération.

Preuve du lemme. Soit $k = \text{mincut}(G)$ et C une coupe de taille k. Pour $1 \leqslant i \leqslant n-2$, considérons les événements

$$A_i = \text{« l'arête choisie à la } i\text{ème étape est dans } C \text{ »}$$

et soit B_i l'événement complémentaire de A_i. On a $\mathbf{P}(A_1) = \frac{k}{|E|}$. Mais tout sommet a degré $\geqslant k$ et donc $|E| \geqslant \frac{kn}{2}$; on a donc $\mathbf{P}(A_1) \leqslant \frac{2}{n}$.

Conditionnellement à B_1, le graphe obtenu après contraction de la première arête a aussi une coupe minimale égale à k. Ce graphe a $n-1$ sommets et on a donc par le même argument

$$\mathbf{P}(A_2|B_1) \leqslant \frac{2}{n-1}.$$

De la même manière, on a

$$\mathbf{P}(A_3|B_1 \cap B_2) \leqslant \frac{2}{n-2}$$

$$\vdots$$

$$\mathbf{P}(A_{n-2}|B_1 \cap B_2 \cap \ldots B_{n-3}) \leqslant \frac{2}{3}$$

On a donc

$$\begin{aligned}\mathbf{P}(B_1 \cap B_2 \cap \cdots \cap B_{n-2}) &= \mathbf{P}(B_1)\mathbf{P}(B_2|B_1)\mathbf{P}(B_3|B_1 \cap B_2)\ldots\mathbf{P}(B_{n-2}|B_1 \cap \cdots \cap B_{n-3}) \\ &\geqslant \left(1-\frac{2}{n}\right)\left(1-\frac{2}{n-1}\right)\cdots\left(1-\frac{2}{3}\right) \\ &= \frac{2}{n(n-1)} \\ &\geqslant \frac{2}{n^2}\end{aligned}$$

Lorsque les événements A_1, A_2, …, A_n sont réalisés, la coupe produite par l'algorithme de Karger est la coupe C. Ceci conclut la preuve du lemme. □

Chapitre 1

Événements, probabilités, variables aléatoires

1.1 Espaces de probabilité

Définition. Un *espace de probabilité* est la donnée de
- un ensemble Ω,
- une famille $\mathcal{F}$ de parties de Ω (c'est-à-dire $\mathcal{F} \subset \mathcal{P}(\Omega)$), l'ensemble des *événements*,
- une fonction $\mathbf{P} : \mathcal{F} \to [0,1]$ qui a un événement associe sa *probabilité*,

qui vérifie les axiomes suivants :

1. La famille $\mathcal{F}$ est une *tribu* (en anglais : σ-algebra), c'est-à-dire telle que
   - Ω est un événement,
   - Si A est un événement, alors $\Omega \setminus A$ est événement,
   - si $(A_n)_{n\in\mathbf{N}}$ est une suite d'événements, alors $\bigcup A_n$ est un événement.
2. $\mathbf{P}$ est une *mesure de probabilité*, c'est-à-dire que
   - on a $\mathbf{P}(\Omega) = 1$ et $\mathbf{P}(\emptyset) = 0$,
   - si $(A_n)_{n\in\mathbf{N}}$ est une suite d'événements deux à deux disjoints (c'est à dire que $A_m \cap A_n = \emptyset$ si $m \neq n$), alors

$$\mathbf{P}\left(\bigcup_{n\in\mathbf{N}} A_n\right) = \sum_{n\in\mathbf{N}} \mathbf{P}(A_n).$$

   Cette propriété s'appelle la *σ-additivité*.

Dans tout le cours, on suppose donné un espace de probabilité $(\Omega, \mathcal{F}, \mathbf{P})$.

Exemple. Si Ω est un ensemble fini, on peut prendre $\mathcal{F} = \mathcal{P}(\Omega)$ et définir pour $A \subset \Omega$

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|}.$$

On dit que $\mathbf{P}$ est la probabilité uniforme sur Ω.

Exemple (généralise le précédent). Si Ω est un ensemble fini ou dénombrable et si $(p_\omega)_{\omega\in\Omega}$ est une famille de réels $\geqslant 0$ vérifiant $\sum p_\omega = 1$, on peut prendre $\mathcal{F} = \mathcal{P}(\Omega)$ et définir pour $A \subset \Omega$

$$\mathbf{P}(A) = \sum_{\omega\in A} p_\omega.$$

Un espace de probabilité de ce type est appelé un espace de probabilité discret.

Remarquons que si A et B sont des événements tels que $A \subset B$, alors $\mathbf{P}(A) \leqslant \mathbf{P}(B)$. En effet, par σ-additivité (appliquée à une suite d'événements dont tous sauf deux sont vides) on a $\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B \setminus A) \geqslant \mathbf{P}(A)$. Le lemme suivant est à la fois trivial et fondamental.

Lemme (Borne de l'union)**.** *Si (A_n) est une suite finie ou dénombrable d'événements, alors*

$$\mathbf{P}\left(\bigcup_n A_n\right) \leqslant \sum_n \mathbf{P}(A_n).$$

Démonstration. On définit $B_n = A_n \setminus \bigcup_{k<n} A_k$. On a alors $B_n \subset A_n$ et $\bigcup B_n = \bigcup A_n$. Puisque les événements B_n sont deux à deux disjoints, on a par σ-additivité,

$$\mathbf{P}\left(\bigcup_n A_n\right) = \mathbf{P}\left(\bigcup_n B_n\right) = \sum_n \mathbf{P}(B_n) \leqslant \sum_n \mathbf{P}(A_n)$$

d'où le résultat. □

Une question naturelle : pourquoi ne pas toujours prendre $\mathcal{F} = \mathcal{P}(\Omega)$? Quel intérêt y a-t-il a exclure des parties de l'ensemble des événements ? Il y a deux raisons sur lesquelles on reviendra

— il y a des cas où on ne peut pas, pour des raisons liées à l'infini.
— même dans le cas discret, il y a parfois intérêt à considérer plusieurs tribus différentes.

1.2 Événements

Définition. Deux événements A et B sont *indépendants* ($A \perp\!\!\!\perp B$) si

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

Si $\mathbf{P}(B) > 0$, la probabilité conditionnelle de A sachant B est définie par $\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$. On a donc

$$A \perp\!\!\!\perp B \iff \mathbf{P}(A|B) = \mathbf{P}(A)$$

et donc la probabilité de A «ne dépend pas» de B. Voila un autre lemme trivial.

Lemme. *Soit (A_n) une partition finie ou dénombrable de Ω en événements telle que $\mathbf{P}(A_n) > 0$ pour tout n. Alors pour tout événement B*

$$\mathbf{P}(B) = \sum_n \mathbf{P}(B \cap A_n) = \sum_n \mathbf{P}(B|A_n)\mathbf{P}(A_n).$$

Définition. Soit (A_n) une famille finie ou infinie d'événements. On dit que les événements (A_n) sont *indépendants* si pour tout ensemble I fini, on a

$$\mathbf{P}\left(\bigcap_{i\in I} A_i\right) = \prod_{i\in I} \mathbf{P}(A_i).$$

Attention : soient trois événements A_1, A_2 , A_3. On a l'implication

$$A_1, A_2, A_3 \text{ indépendants } \implies \mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3)$$

mais la réciproque est fausse en général, comme on s'en convainc en considérant par exemple $A_3 = \emptyset$. De même, si (A_n) sont des événements, alors

$$(A_n) \text{ indépendants } \implies (A_n) \text{ 2 à 2 indépendants}$$

et la réciproque est fausse en général.

Exercice. Montrer que des événements (A_n) sont indépendants si et seulement si les événements $(\Omega \setminus A_n)$ sont indépendants.

Exercice. L'indépendance de n événements requiert de vérifier 2^n équations. Donner, pour tout n, un exemple où toutes ces équations sont vérifiées sauf une.

Fin cours # 1 du 12 septembre

1.3 Théorèmes d'existence

Le théorème suivant justifie l'existence de suites finies ou infinies de «bits aléatoires indépendants», qui sont utilisées dans beaucoup d'algorithmes probabilistes, comme celui de la multiplication matricielle.

Théorème (Existence de bits aléatoires)**.**

1. *Pour tout n, il existe un espace de probabilité $(\Omega_n, \mathcal{F}_n, \mathbf{P}_n)$ et n événements $A_1, \ldots, A_n$ indépendants de probabilité $1/2$.*
2. *Il existe un espace de probabilité $(\Omega, \mathcal{F}, \mathbf{P})$ et une suite infinie $(A_n)_{n\in\mathbf{N}}$ d'événements indépendants de probabilité $1/2$.*

Démonstration. Pour le premier point, on pose $\Omega_n = \{0,1\}^n$, $\mathcal{F}_n = \mathcal{P}(\Omega_n)$ et $\mathbf{P}_n$ la probabilité uniforme. On considère pour $k \in [n]$

$$A_k = \{\omega \in \{0,1\}^n \ : \ \omega_k = 1\}.$$

On a alors $\mathbf{P}_n(A_k) = \frac{1}{2}$, et pour tout $I \subset [n]$

$$\mathbf{P}_n\left(\bigcap_{i\in I} A_i\right) = \frac{2^{n-|I|}}{2^n} = \frac{1}{2^{|I|}} = \prod_{i\in I} \mathbf{P}_n(A_i).$$

Le second point est un résultat difficile que l'on admet. □

Le second point du théorème est équivalent à l'existence d'une probabilité $\mathbf{P}$ sur l'ensemble $\Omega = \{0,1\}^{\mathbf{N}}$ des suites infinies de bits ayant la propriété suivante : pour tout événement $A \subset \{0,1\}^{\mathbf{N}}$ et pour tout $\omega \in \{0,1\}^{\mathbf{N}}$, on a la propriété d'*invariance par translation*

$$\mathbf{P}(A \oplus \omega) = \mathbf{P}(A),$$

où $A \oplus \omega = \{a \oplus \omega \ : \ a \in A\}$, le symbole $\oplus$ désignant l'addition modulo 2 (ou XOR) coordonnée par coordonnée.

Supposant construite une telle probabilité, les événements $(A_n)_{n\in\mathbf{N}}$ définis par

$$A_n = \{\omega \in \{0,1\}^{\mathbf{N}} \ : \ \omega_n = 1\},$$

forment une suite d'événements indépendants de probabilité 1/2 (en effet, si $I \subset \mathbf{N}$ est une partie finie de cardinal k, on peut partitionner $\{0,1\}^{\mathbf{N}}$ en 2^k translatés de $B := \bigcap_{i\in I} A_i$, ce qui implique $\mathbf{P}(B) = 2^{-k}$).

Une difficulté est que la mesure $\mathbf{P}$ ne peut pas être définie sur $\{0,1\}^{\mathbf{N}}$. Supposons par l'absurde qu'elle le soit et considérons la relation d'équivalence sur $\{0,1\}^{\mathbf{N}}$ donnée par

$$(u_n) \sim (v_n) \Longleftrightarrow \{n \ : \ u_n \neq v_n\} \text{ est fini.}$$

Formons un ensemble B en choisissant un représentant dans chaque classe d'équivalence. Notons $Q \subset \{0,1\}^{\mathbf{N}}$ l'ensemble (dénombrable) des suites ayant un nombre fini de 1. On a alors la partition dénombrable

$$\{0,1\}^{\mathbf{N}} = \bigcup_{\omega\in Q} B \oplus \omega$$

et donc, par σ-additivité

$$1 = \mathbf{P}(\{0,1\}^{\mathbf{N}}) = \sum_{\omega\in Q} \mathbf{P}(B \oplus \omega) = \sum_{\omega\in Q} \mathbf{P}(B),$$

ce qui est absurde car la somme d'une infinité de nombres tous égaux ne peut pas valoir 1. La définition de l'ensemble B n'est pas constructive car elle utilise l'axiome du choix. La tribu sur laquelle la probabilité $\mathbf{P}$ est définie est la plus petite tribu contenant les événements A_n ; l'ensemble B n'en fait pas partie.

L'existence de la probabilité $\mathbf{P}$ est équivalente à l'existence de la *mesure de* LEBESGUE λ, qui est l'unique mesure de probabilité sur $[0,1[= \mathbf{R}/\mathbf{Z}$ qui est invariante par translation (modulo 1) et qui a la propriété que $\lambda([a,b]) = b - a$ pour tous $a < b$ dans $[0,1[$. Le lien avec l'ensemble $\{0,1\}^{\mathbf{N}}$ s'obtient en identifiant un réel $x \in [0,1[$ avec la suite de $\{0,1\}^{\mathbf{N}}$ donnée par son développement binaire.

En pratique, l'ensemble des algorithmes probabilistes utilisés par l'humanité n'utilisera qu'un nombre fini de bits aléatoires, donc la version facile du théorème d'existence suffit.

Fin cours # 1 du 29 janvier

1.4 Variables aléatoires

On note $\mathcal{B}_{\mathbf{R}}$ la plus petite tribu de $\mathbf{R}$ qui contient les intervalles ; la tribu $\mathcal{B}_{\mathbf{R}}$ s'appelle la tribu des boréliens de $\mathbf{R}$. Dans la suite on emploiera assez librement les concepts d'ensemble borélien ou de fonction borélienne. L'existence d'ensembles non boréliens ou de fonctions non boréliennes ne s'obtient qu'en utilisant l'axiome du choix ou un axiome de nature similaire ; tout ce qui s'écrit explicitement est borélien.

Définition. Une *variable aléatoire (réelle)* est une fonction $X : \Omega \to \mathbf{R}$ telle que, pour tous $a < b$ réels l'ensemble $\{a \leqslant X \leqslant b\} = X^{-1}([a,b])$ est un événement (c'est-à-dire est dans $\mathcal{F}$). On dit aussi que X est $\mathcal{F}$-mesurable.

Si X est une variable aléatoire réelle, on peut montrer que $X^{-1}(B)$ est un événement pour tout $B \in \mathcal{B}_{\mathbf{R}}$.

Quand $\mathcal{F}$ est la tribu $\mathcal{P}(\Omega)$, toute fonction de Ω dans $\mathbf{R}$ est $\mathcal{F}$-mesurable. Quand $\mathcal{F}$ est la tribu triviale $\{\emptyset, \Omega\}$, seules les fonctions constantes sont $\mathcal{F}$-mesurables. Toute fonction continue (ou même continue par morceaux ou plus généralement «borélienne») d'une v.a. est une v.a.

On définit une variable aléatoire à valeurs dans $\mathbf{R}^n$ (ou *vecteur aléatoire*) comme un n-uplet de variables aléatoires. Si E est un ensemble fini, on définit une variable aléatoire à valeurs dans E comme une fonction $X : \Omega \to E$ telle que $X^{-1}(A)$ est un événement pour tout $A \subset E$.

Exemple. Si A est un événement, la *fonction indicatrice* de A définie pour $\omega \in \Omega$ par

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A \\ 0 & \text{sinon} \end{cases}$$

est une variable aléatoire.

Définition. Soit X une variable aléatoire. La *loi* ou *distribution* de X est la mesure de probabilité $\mathbf{P}_X$ définie sur $(\mathbf{R}, \mathcal{B}_{\mathbf{R}})$ par

$$\mathbf{P}_X(B) = \mathbf{P}(X \in B)$$

pour tout borélien B.

Si X et Y sont des v.a., on note $X \sim Y$ si X et Y ont même loi, c'est-à-dire si $\mathbf{P}_X = \mathbf{P}_Y$. Une idée fondamentale dans l'axiomatisation des probabilités est que seule la loi d'une variable aléatoire X est importante. L'espace de probabilité Ω sous-jacent ainsi que la manière dont est définie la fonction $X : \Omega \to \mathbf{R}$ ne sont pas importants.

Exemple. Voici deux manières différentes de modéliser le lancer d'un dé

1. On peut prendre $\Omega = \{1, \ldots, 6\}$, $X : \Omega \to \mathbf{R}$ la fonction définie par $X(\omega) = \omega$ et $\mathbf{P}$ la probabilité uniforme sur Ω.
2. On peut prendre Ω l'ensemble des conditions initiales (vitesse, force, angle du lancer) et des paramètres (vent, température, ...) qui interviennent dans les équations physiques qui sous-tendent l'expérience du lancer du dé. La mesure $\mathbf{P}$ et la fonction X sont alors extrêmement compliquées, mais ont la propriété que $\mathbf{P}(X = k) = \frac{1}{6}$ pour tout entier k de 1 à 6.

Bien évidemment, les calculs que l'on peut faire sur les statistiques des lancers de dés donneront les mêmes résultats dans chacune de ces deux modélisations.

On peut aussi illustrer par un exemple informatique l'idée que seule les lois des v.a. comptent et non les détails de leur implémentation sur un espace de probabilité : quand un algorithme probabiliste appelle la fonction `random` pour générer des bits aléatoires indépendants, il n'est pas nécessaire de connaître les détails de l'implémentation de cette fonction (sujet par ailleurs passionnant) pour étudier la performance de l'algorithme.

Il y a deux classes importantes de variables aléatoires réelles :

1. Les *variables aléatoires discrètes*, qui prennent leurs valeurs dans un ensemble fini ou dénombrable. Soit X est une variable aléatoire à valeurs dans un sous-ensemble fini ou dénombrable $C \subset \mathbf{R}$. Si pour a dans C on pose $p_a = \mathbf{P}(X = a)$, alors on a $\sum_{a \in C} p_a = 1$.
2. Les *variables aléatoires continues.* Étant donné une fonction $f_X : \mathbf{R} \to \mathbf{R}^+$ continue par morceaux vérifiant $\int_{-\infty}^{\infty} f_X(s)\,\mathrm{d}s = 1$, il existe une variable aléatoire X dont la loi vérifie, pour tout $a < b$

$$\mathbf{P}_X([a,b]) = \mathbf{P}(X \in [a,b]) = \int_a^b f_X(s)\,\mathrm{d}s = 1.$$

On dit que X est une variable aléatoire de densité f_X.

Il existe des variables aléatoires qui ne sont ni discrètes ni continues : par exemple la loi d'un nombre aléatoire dans $[0,1]$ obtenu en choisissant à l'aide d'une suite de bits aléatoires les décimales de son développement en base 10 comme valant soit 3 soit 7.

Si un espace de probabilité admet une suite infinie de bits aléatoires, alors on peut définir dessus une variable aléatoire ayant n'importe quelle loi prescrite.

Exercice. Définir une variable aléatoire ayant une loi uniforme sur $\{1,2,3\}$ à partir d'une suite infinie de bits aléatoires. Est-ce possible à partir d'une suite finie ?

Indépendance de variables aléatoires

Définition. On dit que des variables aléatoires $(X_i)_{i\in I}$ sont *indépendantes* si, quels que soient les réels $(t_i)_{i\in I}$, les événements $\{X_i \leqslant t_i\}$ sont indépendants.

Remarque. Dans le cas discret (où I est fini et les variables aléatoires sont à valeurs dans un ensemble E fini ou dénombrable), les variables aléatoires $(X_i)_{i\in I}$ sont indépendantes si et seulement si la relation

$$\mathbf{P}(\forall i \in I, X_i = x_i) = \prod_{i\in I} \mathbf{P}(X_i = x_i)$$

est vérifiée pour tous les choix de (x_i) dans E.

Lemme (Lemme des coalitions)**.** *Soit $(X_i)_{i\in I}$ des variables aléatoires indépendantes, $I = \bigcup_\alpha I_\alpha$ une partition de I. Alors, si on pose*

$$Y_\alpha = f_\alpha((X_i)_{i\in I_\alpha})$$

(les fonctions $f_\alpha : \mathbf{R}^{I_\alpha} \to \mathbf{R}$ étant «boréliennes»), les variables aléatoires (Y_α) sont indépendantes.

En particulier, si X et Y sont indépendantes, alors des variables aléatoires de la forme $f(X)$ et $g(Y)$ sont indépendantes.

Voici un dernier théorème d'existence.

Théorème. *Étant donnée une suite (μ_n) de mesures de probabilités sur $\mathbf{R}$, il existe un espace de probabilité Ω, et pour tout n une variable aléatoire $X_n : \Omega \to \mathbf{R}$ de loi μ_n, tels que les variables aléatoires (X_n) sont indépendantes.*

On dira que les variables aléatoires (X_n) sont i.i.d. (indépendantes et identiquement distribuées) si elles sont indépendantes et de même loi.

1.5 Espérance d'une variable aléatoire

Si X est une variable aléatoire, on veut définir son espérance $\mathbf{E}[X]$ comme la valeur moyenne qu'elle prend.

Dans le cas discret, si X prend les valeurs réelles $x_1, \ldots, x_n$, on pose

$$\mathbf{E}[X] = \sum_{k=1}^{n} x_k \mathbf{P}(X = x_k).$$

Dans le cas général, on procède en plusieurs étapes.

1. Pour tout événement A, on pose $\mathbf{E}[\mathbf{1}_A] = \mathbf{P}(A)$.

2. On étend cette définition par linéarité : si $X = \sum_i \lambda_i \mathbf{1}_{A_i}$ (somme finie), on pose
$$\mathbf{E}[X] = \sum \lambda_i \mathbf{P}(A_i).$$
On vérifie que cette définition est cohérente : si on a $\sum \lambda_i \mathbf{1}_{A_i} = \sum \mu_j \mathbf{1}_{B_j}$, alors on a $\sum \lambda_i \mathbf{P}(A_i) = \sum \mu_j \mathbf{P}(B_j)$. Cette étape permet de définir l'espérance d'une variable aléatoire prenant un nombre fini de valeurs.
3. Si X est une variable aléatoire positive, on peut l'écrire comme $X = \lim X_n$ où (X_n) est une suite croissante de variables aléatoires prenant un nombre fini de valeurs, et on pose alors
$$\mathbf{E}[X] = \lim \mathbf{E}[X_n]$$
en vérifiant que cette définition ne dépend pas du choix de la suite X_n. Cette limite existe dans $[0, +\infty]$ comme limite d'une suite croissante.
4. Si X est une variable aléatoire telle que $\mathbf{E}[|X|] < +\infty$ (une telle variable est dite *intégrable*), on écrit $X = X^+ - X^-$ (où $X^+ = \max(0, X)$ et $X^- = \max(0, -X)$ sont des variables aléatoires positives) et on pose
$$\mathbf{E}[X] = \mathbf{E}[X^+] - \mathbf{E}[X^-].$$

La raison pour laquelle on se restreint aux variables aléatoires intégrables pour définir l'espérance est qu'on veut éviter d'écrire une forme indéterminée du type $(+\infty) - (+\infty)$.

Proposition (Linéarité de l'espérance)**.** *Si X et Y sont des variables aléatoires intégrables et $c \in \mathbf{R}$, alors*
$$\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y],$$
$$\mathbf{E}[cX] = c\,\mathbf{E}[X].$$

Pour les variables à valeurs dans $\mathbf{N}$, on a la formule suivante.

Proposition. *Soit Y une variable aléatoire à valeurs dans $\mathbf{N}$. Alors*
$$\mathbf{E}[Y] = \sum_{k=1}^{\infty} \mathbf{P}(Y \geqslant k).$$

En effet, $\mathbf{P}(Y \geqslant k) = \sum_{n=k}^{\infty} \mathbf{P}(Y = n)$ et on inverse les sommes.

Proposition. *Si X et Y sont des variables aléatoires indépendantes et intégrables, alors la variable aléatoire XY est intégrable et*
$$\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y].$$

Démonstration. Par approximation, il suffit de traiter le cas où X et Y prennent un nombre fini de valeurs. Écrivons
$$X = \sum \lambda_i \mathbf{1}_{A_i}, \quad Y = \sum \mu_j \mathbf{1}_{B_j},$$
les événements (A_i) (resp. (B_j)) étant disjoints. Quels que soient les indices i et j, les événements $A_i = X^{-1}(\lambda_i)$ et $B_j = Y^{-1}(\mu_j)$ sont indépendants et donc $\mathbf{P}(A_i \cap B_j) = \mathbf{P}(A_i)\mathbf{P}(B_j)$. Puisque
$$XY = \sum_{i,j} \lambda_i \mu_j \mathbf{1}_{A_i \cap B_j},$$
on a
$$\mathbf{E}[XY] = \sum_{i,j} \lambda_i \mu_j \mathbf{P}(A_i \cap B_j) = \left(\sum_i \lambda_i \mathbf{P}(A_i) \right) \left(\sum_j \mu_j \mathbf{P}(B_j) \right) = \mathbf{E}[X]\,\mathbf{E}[Y],$$
d'où le résultat. □

Corollaire. *Si les variables aléatoires $X_1, \dots, X_n$ sont indépendantes, et si $f_1, \dots, f_n$ sont des fonctions telles que les variables $f_i(X_i)$ sont intégrables, alors*

$$\mathbf{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] = \prod_{i=1}^{n} \mathbf{E}[f_i(X_i)].$$

Enfin, mentionnons comment on calcule l'espérance d'une fonction d'une variable aléatoire continue.

Proposition («Formule du transfert»)**.** *Soit X une variable aléatoire continue admettant une densité f_X. Pour toute fonction $h : \mathbf{R} \to \mathbf{R}$, on a*

$$\mathbf{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)\, \mathrm{d}x$$

dès lors que l'intégrale a un sens.

En particulier, l'espérance d'une variable aléatoire intégrable de densité f_X s'obtient comme

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, \mathrm{d}x$$

Fin cours # 2 du 17 septembre

1.6 Exemple : `QuickSort` randomisé

Nous allons décrire un exemple qui illustre l'efficacité du principe de linéarité de l'espérance.

Supposons que l'on doive trier une liste S de n nombres que l'on suppose distincts (c'est le cas le plus dur). L'algorithme récursif `QuickSort` consiste à choisir un élément x de S, que l'on compare à tous les autres éléments pour écrire la partition

$$S = S_- \cup \{x\} \cup S_+$$

où $S_- = \{y \in S \ : \ y < x\}$ et $S_+ = \{y \in S \ : \ y > x\}$, puis à trier S_- et S_+ par des appels récursifs à `QuickSort`.

La complexité C_n de l'algorithme (que l'on définit comme le nombre total de comparaisons effectuées) dépend du choix des pivots : c'est une variable aléatoire. Dans le pire cas, le pivot choisi est toujours le plus petit possible et alors

$$C_n = (n-1) + (n-2) + \cdots + 1 = \frac{n(n-1)}{2}$$

(toutes les comparaisons possibles ont été effectuées). Dans le meilleur cas, le pivot choisi est toujours la médiane de l'ensemble considéré et on a

$$C_n = (n-1) + C_{\lceil \frac{n}{2} \rceil} + C_{\lfloor \frac{n}{2} \rfloor},$$

d'où on tire l'estimation $C_n = O(n \log n)$ qui est la complexité optimale d'un algorithme de tri.

L'algorithme `Randomized Quicksort` est la variante de l'algorithme `Quicksort` où les pivots sont choisis au hasard à chaque étape, indépendamment et selon la loi uniforme. On s'intéresse alors au temps moyen d'exécution $\mathbf{E}[C_n]$. Nous allons voir que le principe de la linéarité de l'espérance permet un calcul élégant de la complexité moyenne.

Théorème. *Pour* `Randomized Quicksort`*, on a* $\mathbf{E}[C_n] \sim 2n \log n$ *quand* $n \to \infty$.

Démonstration. Soit $S = \{x_1, \ldots, x_n\}$ avec $x_1 < x_2 < \cdots < x_n$. Remarquons que chaque couple d'éléments distincts de S sera comparé 0 ou 1 fois au cours de l'algorithme. Pour $i < j$, soit A_{ij} l'événement «les éléments x_i et x_j ont été comparés au cours de l'exécution de l'algorithme». On a

$$\begin{aligned} \mathbf{E}[C_n] &= \mathbf{E}\left[\sum_{i<j} \mathbf{1}_{A_{ij}}\right] \\ &= \sum_{i<j} \mathbf{P}(A_{ij}). \end{aligned}$$

L'observation cruciale est la suivante ; deux éléments $x_i < x_j$ ont été comparés pendant l'exécution de l'algorithme si et seulement si, la première fois qu'un pivot est choisi parmi $\{x_i, x_{i+1}, \ldots, x_j\}$, ce pivot est x_i ou x_j. On a donc $\mathbf{P}(A_{ij}) = \frac{2}{j-i+1}$. On a donc

$$\mathbf{E}[C_n] = \sum_{i<j} \frac{2}{j-i+1} = 2\sum_{k=1}^{n-1} \frac{1}{k+1}(n-k) = 2(n+1)\sum_{k=1}^{n-1} \frac{1}{k+1} - 2(n-1)$$

d'où le résultat. □

1.7 La loi géométrique

Si a est un réel, la *mesure de* DIRAC en a, notée δ_a est la loi d'une variable aléatoire X telle que $\mathbf{P}(X = a) = 1$. On dit aussi que X est presque sûrement égale à a.

Soit $p \in [0, 1]$. La loi de BERNOULLI de paramètre p, notée $\mathsf{B}(p)$ est la loi $p\delta_1 + (1-p)\delta_0$. Une variable aléatoire X a pour loi $\mathsf{B}(p)$, ce qu'on note $X \sim \mathsf{B}(p)$, si et seulement si $\mathbf{P}(X = 1) = p$ et $\mathbf{P}(X = 0) = 1 - p$. La loi $\mathsf{B}(\frac{1}{2})$ est la loi d'un bit aléatoire.

Soient $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires i.i.d. de loi $\mathsf{B}(p)$. On considère la variable aléatoire

$$Y = \min\{k \geqslant 1 \ : \ X_k = 1\}$$

donnée comme l'indice du premier 1. On a $\mathbf{P}(Y = k) = (1-p)^{k-1}p$. Si on suppose $0 < p \leqslant 1$, alors

$$\sum_{k=1}^{\infty} (1-p)^{k-1}p = p\sum_{j=0}^{\infty} (1-p)^j = 1$$

et donc la variable aléatoire Y prend presque sûrement une valeur finie. La loi de Y est appelée *loi géométrique* de paramètre p et notée $\mathsf{G}(p)$.

Si $Y \sim \mathsf{G}(p)$, alors (par un calcul ou un raisonnement) $\mathbf{P}(Y > k) = (1-p)^k$. De plus, $\mathbf{E}[Y] = \frac{1}{p}$.

Proposition (Absence de mémoire de la loi géométrique)**.** *Soit Y une variable aléatoire de loi $\mathsf{G}(p)$. Alors pour tous k, $n > 0$*

$$\mathbf{P}(Y = n + k | Y > k) = \mathbf{P}(Y = n).$$

Autrement dit, la loi conditionnelle de $Y - k$ sachant que $Y > k$ est la même que la loi de Y.

Démonstration. Il est équivalent de montrer que $\mathbf{P}(Y > n+k|Y > k) = \mathbf{P}(Y > n)$ et c'est immédiat au vu de la formule $\mathbf{P}(Y > k) = (1-p)^k$. □

Exercice. Soient $Y_1 \sim \mathsf{G}(p_1)$ et $Y_2 \sim \mathsf{G}(p_2)$ deux variables aléatoires indépendantes. Quelle est la loi de $\min(Y_1, Y_2)$? (Il est possible répondre sans aucun calcul.)

Voici un exemple important où intervient la loi géométrique : le problème du *collectionneur de vignettes*.

Soit E un ensemble fini de cardinal N et $(X_n)_{n\geqslant 1}$ des variables aléatoires i.i.d. de loi uniforme sur E (penser à une collection d'images Panini). On considère

$$Y = \min\{k \ : \ \{X_1, \dots, X_k\} = E\},$$

le nombre de vignettes qu'il faut amasser avant d'avoir une collection complète. On veut calculer $\mathbf{E}[Y]$, la valeur moyenne de Y.

Introduisons pour $1 \leqslant j \leqslant N$ les variables aléatoires

$$T_j = \min\{k \ : \ |\{X_1, \dots, X_k\}| = j\},$$

de sorte que $Y = T_N$. On a $T_1 = 1$ et $T_2 - 1 \sim \mathsf{G}(\frac{N-1}{N})$. Plus généralement, on a

Proposition. *Les variables aléatoires $Z_1, \dots, Z_N$ définies par $Z_1 = 1$ et $Z_j = T_j - T_{j-1}$ pour $1 < j \leqslant N$ sont indépendantes. De plus Z_j suit la loi $\mathsf{G}(\frac{N+1-j}{N})$.*

Esquisse de démonstration. Nous devons montrer que pour tout choix d'entiers $k_2, \dots, k_N$, on a

$$\mathbf{P}(Z_2 = k_2, \dots, Z_N = k_N) = \prod_{j=2}^{N} \left[\left(\frac{j-1}{N} \right)^{k_j - 1} \left(1 - \frac{j-1}{N} \right) \right]$$

On peut réécrire le membre de gauche comme

$$\mathbf{P}(Z_2 = k_2)\mathbf{P}(Z_3 = k_3|Z_2 = k_2) \cdots \mathbf{P}(Z_N = k_N|Z_2 = k_2, \dots, Z_{N-1} = k_{N-1})$$

Fixons j et soit $\overline{x} = (x_1, \dots, x_\ell) \in E^\ell$ tel que $|\{x_1, \dots, x_\ell\}| = j-1$ et $x_i \neq x_\ell$ si $i < \ell$. On considère l'événement $H(\overline{x}) = \{X_1 = x_1, \dots, X_\ell = x_\ell\}$. On a

$$\mathbf{P}(Z_j = k_j|H(\overline{x})) = \left(\frac{j-1}{N} \right)^{k_j - 1} \frac{N-j+1}{N}.$$

On en déduit que pour tous $k_2, \dots, k_{j-1}$

$$\mathbf{P}(Z_j = k_j|Z_2 = k_2, \dots, Z_{j-1} = k_{j-1}) = \left(\frac{j-1}{N} \right)^{k_j - 1} \frac{N-j+1}{N},$$

par la propriété élémentaire suivante des probabilités conditionnelles : si (B_i) est une famille finie d'événements disjoints tels que $\mathbf{P}(A|B_i) = p$ pour tout i, alors $\mathbf{P}(A|\bigcup B_i) = p$. (On utilise le fait que l'événement «$Z_2 = k_2$, …, $Z_j = k_j$» est réunion disjointe de tels événements $H(\overline{x})$). □

On peut donc écrire la variable aléatoire

$$Y = T_N = Z_1 + \cdots + \dots Z_N$$

comme une somme de variables aléatoires indépendantes de loi géométrique. Par linéarité de l'espérance, on en déduit

$$\begin{aligned}\mathbf{E}[Y] &= \mathbf{E}[Z_1] + \cdots + \mathbf{E}[Z_N] \\ &= \sum_{j=1}^{N} \frac{N}{N+1-j} \\ &= N \sum_{k=1}^{N} \frac{1}{k} \\ &\sim N \log N\end{aligned}$$

Chapitre 2

Moments et déviations

On a vu quelques calculs d'espérance, par exemple pour le temps d'exécution de `QuickSort` ou pour le problème du collectionneur de vignettes. Mais l'espérance d'une variable aléatoire ne suffit bien sur pas à déterminer la loi. Par exemple, les deux variables aléatoires suivantes ont une espérance de 1

1. une variable aléatoire X telle que $\mathbf{P}(X = 1) = 1$,
2. une variable aléatoire Y telle que $\mathbf{P}(Y = n) = \frac{1}{n}$ et $\mathbf{P}(Y = 0) = \frac{n-1}{n}$, où $n \gg 1$.

On aimerait savoir a priori qu'une variable aléatoire est souvent proche de son espérance ; c'est le cas de la variable X mais pas de la variable Y.

2.1 Les inégalités de Markov et de Tchebychev

Théorème (Inégalité de Markov)**.** *Soit X une variable aléatoire à valeurs $\geqslant 0$. Alors, pour tout $a > 0$,*

$$\mathbf{P}(X \geqslant a) \leqslant \frac{\mathbf{E}[X]}{a}.$$

Démonstration. On a $X \geqslant a\mathbf{1}_{X \geqslant a}$, et donc $\mathbf{E}[X] \geqslant a\mathbf{P}(X \geqslant a)$. □

En général, la borne donnée par l'inégalité de Markov est trop faible. On peut l'améliorer en remplaçant l'espérance par des « moment plus grands ». Soit $k \in \mathbf{N}$. Lorsque la variable aléatoire X^k est intégrable, on dit que X *admet un moment d'ordre k* et la quantité $\mathbf{E}[X^k]$ s'appelle le *moment d'ordre k* de X.

Si une variable aléatoire positive X admet un moment d'ordre k, alors pour tout $a > 0$,

$$\mathbf{P}(X \geqslant a) \leqslant \frac{\mathbf{E}[X^k]}{a^k}$$

comme on le voit en appliquant l'inégalité de Markov à la variable aléatoire X^k.

Les moments de différents ordres sont comparés à l'aide de l'inégalité suivante.

Lemme. *Soit $1 \leqslant p \leqslant q$. Alors pour toute variable aléatoire X on a*

$$(\mathbf{E}[|X|^p])^{1/p} \leqslant (\mathbf{E}[|X|^q])^{1/q}.$$

Démonstration. Puisque l'inégalité à montrer se réécrit en $\mathbf{E}[|Y|] \leqslant (\mathbf{E}[|Y|^r])^{1/r}$ avec $Y = |X|^p$ et $r = q/p$, il suffit de traiter le cas où $p = 1$.

Par homogénéité, on peut également supposer que $\mathbf{E}[|X|^q] = 1$. Par convexité de $x \mapsto x^q$, on a pour tout $x \geqslant 0$,

$$qx \leqslant x^q + (q - 1),$$

d'où on déduit en prenant l'espérance

$$q\,\mathbf{E}[|X|] \leqslant \mathbf{E}[|X|^q] + (q-1) = q,$$

puis $\mathbf{E}[|X|] \leqslant 1$. □

Si X est une variable aléatoire qui admet un moment d'ordre 2, sa *variance* est définie comme

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

et son *écart-type* (en anglais *standard deviation*) comme

$$\sigma(X) = \sqrt{\mathbf{Var}[X]}.$$

La variance est homogène d'ordre 2, au sens où $\mathbf{Var}[s + tX] = t^2\,\mathbf{Var}[X]$.

Si X et Y sont deux variables aléatoire définies sur le même espace de probabilité qui admettent un moment d'ordre 2, leur *covariance* est donnée par

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y]).$$

L'inégalité de Cauchy–Schwarz implique que $|\,\mathbf{Cov}(X, Y)| \leqslant \sigma(X)\sigma(Y)$.

On a également

$$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\,\mathbf{Cov}(X, Y).$$

On en déduit

Proposition. *Si X et Y sont des variables aléatoires indépendantes, alors $\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$. Plus généralement, si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes, alors*

$$\mathbf{Var}[X_1 + \cdots + X_n] = \mathbf{Var}[X_1] + \cdots + \mathbf{Var}[X_n].$$

La version «moment d'ordre 2» de l'inégalité de Markov est connue sous le nom d'inégalité de Tchebychev. C'est une inégalité de déviations : il est peu probable qu'une variable aléatoire prenne ses valeurs en dehors d'un intervalle autour de sa moyenne et de largeur proportionnelle à l'écart-type.

Proposition (Inégalité de Tchebychev)**.** *Si une variable aléatoire X admet un moment d'ordre 2, alors pour tout $a > 0$,*

$$\mathbf{P}(|X - \mathbf{E}[X]| \geqslant a) \leqslant \frac{\mathbf{Var}[X]}{a^2}.$$

Démonstration. On écrit $\mathbf{P}(|X - \mathbf{E}[X]| \geqslant a) = \mathbf{P}(|X - \mathbf{E}[X]|^2 \geqslant a^2) \leqslant \mathbf{E}[(X - \mathbf{E}[X])^2]/a^2$ par l'inégalité de Markov. □

Revenons enfin sur le problème du collectionneur de vignettes. On avait écrit le temps T nécessaire pour avoir une collection complète comme

$$T_N = Z_1 + \cdots + Z_N$$

où les variables aléatoires Z_i sont indépendantes, et $Z_i \sim \mathsf{G}(\frac{i}{N})$. Le calcul de l'espérance $\mathbf{E}[T_N] \sim N \log N$ n'a utilisé que la linéarité de l'espérance. On peut exploiter l'indépendance en écrivant

$$\mathbf{Var}[T_N] = \mathbf{Var}[Z_1] + \cdots + \mathbf{Var}[Z_N].$$

Si $X \sim \mathsf{G}(p)$, alors $\mathbf{Var}[X] = \frac{1-p}{p^2}$ (exercice) et en particulier $\mathbf{Var}[X] \leqslant \frac{1}{p^2}$. On a donc

$$\mathbf{Var}[T_N] \leqslant \sum_{i=1}^{N} \left(\frac{N}{i}\right)^2 \leqslant CN^2.$$

On a donc $\mathbf{Var}[T_N] = o(\mathbf{E}[T_N]^2)$ et on peut alors conclure que T_N est de l'ordre de $\mathbf{E}[T_N]$ avec grande probabilité : pour tout $\varepsilon > 0$

$$\mathbf{P}(|T_N - \mathbf{E}[T_N]| > \varepsilon\, \mathbf{E}[T_N]) \leqslant \frac{\mathbf{Var}[T_N]}{\varepsilon^2(\mathbf{E}[T_N])^2} \to 0$$

Fin cours #3 du 24 septembre

2.2 La loi faible des grands nombres

On dit qu'une suite (X_n) de variables aléatoires *converge en probabilité* vers une variable aléatoire X si

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbf{P}(|X_n - X| > \varepsilon) = 0.$$

Par exemple, si T_N est l'exemple donné précédemment dans le contexte du problème du collectionneur de coupons, alors la suite (X_N) définie par $X_N = \frac{N}{N \log N}$ converge en probabilité vers la v.a. constante égale à 1.

Théorème (Loi faible des grands nombres)**.** *Soit (X_n) une suite de variables aléatoires i.i.d. admettant un moment d'ordre 2. Soit $Y_n = \frac{1}{n}(X_1 + \cdots + X_n)$ la suite de ses moyennes de Cesàro. Alors (Y_n) converge en probabilité vers une variable aléatoire constante égale à $\mathbf{E}[X_1]$.*

Démonstration. Par linéarité de l'espérance on a $\mathbf{E}[Y_n] = \mathbf{E}[X_1]$. Par additivité de la variance pour des sommes indépendantes, on a $\mathbf{Var}[Y_n] = \frac{1}{n}\mathbf{Var}[X_1]$. On a donc, pour tout $\varepsilon > 0$,

$$\mathbf{P}[|Y_n - \mathbf{E}[X_1]| > \varepsilon] = \mathbf{P}[|Y_n - \mathbf{E}[Y_n]| > \varepsilon] \leqslant \frac{\mathbf{Var}[Y_n]}{\varepsilon^2} = \frac{\mathbf{Var}[X_1]}{n\varepsilon^2}$$

qui tend bien vers 0. □

Voici une conséquence de la loi des grands nombres. Soit $p \in (0,1)$ et (X_n) une suite de variables aléatoires i.i.d. de loi de Bernoulli $\mathsf{B}(p)$. La loi de la somme

$$Y_n = X_1 + \cdots + X_n$$

s'appelle la *loi binomiale de paramètres n et p* et se note $\mathsf{B}(n,p)$. On calcule $\mathbf{E}[Y_n] = n\,\mathbf{E}[X_1] = np$ et $\mathbf{Var}[Y_n] = n\,\mathbf{Var}[X_1] = np(1-p)$. La loi binomiale est décrite plus explicitement par la formule

$$\mathbf{P}(Y_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Dans le cas particulier important où $p = \frac{1}{2}$, on a alors $\mathbf{P}(Y_n = k) = 2^{-n}\binom{n}{k}$. La loi faible des grands nombres implique alors le résultat suivant : lorsque $n \gg 1$, quasiment toute la masse dans la $n$ème ligne du triangle de Pascal se concentre dans les 1% de coefficients centraux :

$$\sum_{(\frac{1}{2}-\varepsilon)n \leqslant k \leqslant (\frac{1}{2}+\varepsilon)n} \binom{n}{k} = 2^n(1 - o(1)).$$

2.3 Les inégalités de Chernoff

Si X est une variable aléatoire, on appelle *fonction génératrice des moments* de X la fonction

$$M_X(t) = \mathbf{E}[e^{tX}].$$

Cette fonction contient toutes les informations sur les moments de X.

Théorème. *Soit X une variable aléatoire vérifiant $M_X(t) < \infty$ pour $|t| < t_0$. Alors X admet des moments de tous les ordres et on a la relation*

$$M_X(t) = \sum_{k=0}^{\infty} \mathbf{E}[X^k] \frac{t^k}{k!}$$

pour tout $|t| < t_0$.

Le théorème s'obtient en écrivant la série entière définissant e^{tX} et en justifiant les calculs à l'aide du théorème de convergence dominée : *si une suite (Z_n) de variables aléatoires converge vers Z, et s'il existe une variable aléatoire intégrable Y telle que $|Z_n| \leqslant Y$, alors* $\mathbf{E}[Z] = \lim \mathbf{E}[Z_n]$.

La fonction génératrice des moments permet de calculer les moments. Par exemple, si $X \sim \mathsf{G}(p)$, alors

$$\mathbf{E}[e^{tX}] = \sum_{k=1}^{\infty} p(1-p)^{k-1} e^{tk} = pe^t \sum_{k=0}^{\infty} (e^t(1-p))^k = \frac{pe^t}{1-(1-p)e^t}$$

dès lors que $|t| < |\ln(1-p)|$. La loi géométrique admet donc des moments de tous les ordres, que l'on peut calculer à l'aide du développement limité en 0 de la fonction $t \mapsto \frac{pe^t}{1-(1-p)e^t}$.

Proposition. *Si X et Y sont des variables aléatoires indépendantes, alors $M_{X+Y}(t) = M_X(t) M_Y(t)$.*

Démonstration. On écrit $\mathbf{E}[e^{t(X+Y}] = \mathbf{E}[e^{tX}]\,\mathbf{E}[e^{tY}]$ par indépendance. □

Par exemple, si X suit la loi $\mathsf{B}(n,p)$, on a

$$M_X(t) = ((1-p) + pe^t)^n$$

puisque X a la même loi qui la somme de variables aléatoires i.i.d. de loi $\mathsf{B}(p)$.

Voici l'inégalité de déviation la plus importante.

Théorème (Inégalité de Chernoff I)**.** *Soit X une variable aléatoire de loi $\mathsf{B}(n, \frac{1}{2})$. On note $\mu = \mathbf{E}[X] = n/2$. Pour tout $a > 0$, on a*

$$\mathbf{P}(X \geqslant \mu + a) \leqslant \exp(-2a^2/n)$$

$$\mathbf{P}(X \leqslant \mu - a) \leqslant \exp(-2a^2/n)$$

Voici une version équivalente ou les valeurs 0 et 1 des lois de Bernoulli sont remplacées par les valeurs -1 et 1.

Théorème (Inégalité de Chernoff I, variante)**.** *Soient $Y_1, \ldots, Y_n$ des variables aléatoires i.i.d. de loi uniforme sur $\{-1, 1\}$ et $Y = Y_1 + \cdots + Y_n$. Pour tout $x > 0$, on a*

$$\mathbf{P}(Y \geqslant x\sqrt{n}) \leqslant \exp(-x^2/2)$$

$$\mathbf{P}(Y \leqslant -x\sqrt{n}) \leqslant \exp(-x^2/2)$$

Les deux versions sont équivalentes : si les variables X_i et Y_i sont reliées par la relation $Y_i = 2X_i - 1$, alors

$$X_1, \dots, X_n \text{ i.i.d. de loi } \mathsf{B}(1/2) \iff Y_1, \dots, Y_n \text{ i.i.d. de loi unforme sur } \{-1, 1\}$$

Si on pose $X = X_1 + \cdots + X_n$ et $Y = Y_1 + \cdots + Y_n$, alors $Y = 2X - n$ et donc

$$X \geqslant \mu + a \iff Y \geqslant 2a$$

$$X \leqslant \mu - a \iff Y \leqslant 2a$$

et l'on passe d'un énoncé à l'autre par la formule $2a = x\sqrt{n}$.

Démonstration. Montrons la seconde version. L'idée est d'appliquer l'inégalité de Markov a une fonction bien choisie de Y. Pour tout réel $t > 0$, on a

$$\mathbf{P}(Y \geqslant x\sqrt{n}) = \mathbf{P}(\exp(tY) \geqslant \exp(tx\sqrt{n})) \leqslant e^{-tx\sqrt{n}}\, \mathbf{E}[e^{tY}] = e^{-tx\sqrt{n}} M_Y(t)$$

Par ailleurs, on a $M_Y(t) = M_{Y_1}(t)^n = \cosh(t)^n$. On utilise maintenant le

Lemme. *Pour tout réel t, on a $\cosh(t) \leqslant \exp(t^2/2)$.*

qui se montre en comparant terme à terme les deux séries entières. On a donc $M_Y(t) \leqslant \exp(nt^2/2)$ puis

$$\mathbf{P}(Y \geqslant x\sqrt{n}) \leqslant e^{nt^2/2 - tx\sqrt{n}}.$$

Enfin, on optimise sur t en choisissant la valeur $t = x/\sqrt{n}$, d'où le résultat. La seconde partie du théorème s'obtient en remarquant que $Y \sim -Y$. □

Cette majoration est BEAUCOUP plus précise que l'inégalité de Tchebychev. Par exemple, si $X \sim \mathsf{B}(n, \frac{1}{2})$, on a

$$\mathbf{P}(X \geqslant \frac{3}{4}n) \leqslant \frac{2}{3} \quad \text{par l'inégalité de Markov}$$

$$\mathbf{P}(X \geqslant \frac{3}{4}n) \leqslant \frac{4}{n} \quad \text{par l'inégalité de Tchebychev}$$

$$\mathbf{P}(X \geqslant \frac{3}{4}n) \leqslant \exp(-n/8) \quad \text{par l'inégalité de Chernoff I}$$

L'inégalité de Chernoff est extrêmement précise. On verra plus tard (par le théorème central limite) que si $Y^{(n)}$ est une somme de n v.a. i.i.d. de loi uniforme sur $\{-1, 1\}$;

$$\lim_{n\to\infty} \mathbf{P}(Y^{(n)} \geqslant x\sqrt{n}) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-u^2/2)\, \mathrm{d}u$$

et cette quantité est équivalente à $\exp(-x^2/2)/x\sqrt{2\pi}$ lorsque x tend vers l'infini : l'exposant dans l'exponentielle donné par l'inégalité de Chernoff est optimal.

Il existe aussi une inégalité de Chernoff qui couvre le cas général d'une sommes de variables de Bernoulli indépendantes.

Théorème (Inégalité de Chernoff II)**.** *Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes, avec $X_k \sim \mathsf{B}(p_k)$. On pose $X = X_1 + \cdots + X_n$ et $\mu = \mathbf{E}[X] = p_1 + \cdots + p_n$. Alors*

1. *Pour tout $\delta > 0$, on a*

$$\mathbf{P}(X \geqslant (1+\delta)\mu) \leqslant \left(\frac{\exp(\delta)}{(1+\delta)^{1+\delta}}\right)^{\mu} \leqslant \exp\left(-\frac{\delta^2}{2+\delta}\mu\right)$$

2. *Pour tout $R \geqslant 6\mu$, on a $\mathbf{P}(X \geqslant R) \leqslant 2^{-R}$.*
3. *Pour tout $0 < \delta < 1$, on a*

$$\mathbf{P}(X \leqslant (1-\delta)\mu) \leqslant \left(\frac{\exp(-\delta)}{(1-\delta)^{1-\delta}}\right)^{\mu} \leqslant \exp\left(-\frac{\delta^2}{2}\mu\right)$$

Démonstration. On applique la même stratégie. Pour $t > 0$ à déterminer, on a

$$\mathbf{P}(X \geqslant (1+\delta)\mu) = \mathbf{P}(e^{tX} \geqslant e^{t(1+\delta)\mu}) \leqslant e^{-t(1+\delta)\mu} M_X(t).$$

On a $M_{X_i}(t) = (1-p_i) + p_i e^t = 1 + p_i(e^t - 1) \leqslant \exp(p_i(e^t-1))$, et donc par indépendance

$$M_X(t) = \prod M_{X_i}(t) \leqslant \exp(\mu(e^t - 1)).$$

On choisit maintenant la valeur $t_1 = \ln(1+\delta)$ pour obtenir

$$\mathbf{P}(X \geqslant (1+\delta)\mu) \leqslant \exp(\mu(e^{t_1}-1) - (1+\delta)t_1\mu) = \left(\frac{\exp(\delta)}{(1+\delta)^{1+\delta}}\right)^{\mu}.$$

Pour le dernier point, on écrit pour $t < 0$ à déterminer

$$\mathbf{P}(X \leqslant (1-\delta)\mu) = \mathbf{P}(e^{tX} \geqslant e^{t(1-\delta)\mu}) \leqslant e^{-t(1-\delta)\mu} M_X(t).$$

En choisissant la valeur $t_2 = \ln(1-\delta)$, il vient

$$\mathbf{P}(X \leqslant (1-\delta)\mu) \leqslant \left(\frac{\exp(-\delta)}{(1-\delta)^{1-\delta}}\right)^{\mu}$$

Les inégalités dans le premier et dernier points découlent des lemmes suivants, qui peuvent se démontrer par de banales études de fonctions.

Lemme. *Pour tout $\delta > 0$, on a $\delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{2+\delta} \leqslant 0$*

Lemme. *Pour tout $0 < \delta < 1$, on a $-\delta - (1-\delta)\ln(1-\delta) + \frac{\delta^2}{2} \leqslant 0$.*

Il reste à montrer le deuxième point. En écrivant $R = (1+\delta)\mu$, il vient (puisque $1 + \delta \geqslant 6$)

$$\mathbf{P}(X \geqslant R) \leqslant \left(\frac{\exp(\delta)}{(1+\delta)^{1+\delta}}\right)^{\mu} \leqslant \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu} \leqslant (e/6)^R \leqslant 2^{-R}. \qquad \square$$

Les bornes données par l'inégalité de Chernoff dans la cas général ne vont pas intervenir une décroissante sous-gaussienne (en $\exp(-ct^2)$) mais plutôt sous-exponentielle (en $\exp(-ct)$). L'observation remarque suivant montre que c'est inévitable.

On appelle loi de Poisson de paramètre $\lambda > 0$ la loi d'une variable aléatoire X à valeurs entières telle que, pour tout $k \in \mathbf{N}$,

$$\mathbf{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

On vérifie que $\sum_{k=0}^{\infty} \mathbf{P}(X = k) = 1$. Dans ce cas, on note $X \sim \mathsf{P}(\lambda)$.

La loi de Poisson apparaît dans la limite des événements rares, comme le montre la proposition suivante.

Proposition. *Soit (X_j) une suite de variables aléatoires, avec $X \sim \mathsf{B}(n_i, p_j)$, où les paramètre n_j et p_j sont tels que*

$$\lim_{j\to\infty} n_j = \infty, \quad \lim_{j\to\infty} p_j = 0, \quad \lim_{j\to\infty} n_j p_j = \lambda \in]0, \infty[.$$

Soit X une variable aléatoire de loi $\mathsf{P}(\lambda)$. Alors, pour tout $k \in \mathbf{N}$,

$$\lim_{j\to\infty} \mathbf{P}(X_j = k) = \mathbf{P}(X = k).$$

Démonstration. On a, pour tout $k \in \mathbf{N}$,

$$\mathbf{P}(X_j = k) = \binom{n_j}{k} p_j^k (1-p_j)^{n_j - k}.$$

Sous les hypothèses de la proposition, on a les équivalents

$$\binom{n_j}{k} \sim \frac{n_j^k}{k!}, \quad (1-p_j)^{-k} \sim 1$$

et on conclut en utilisant le fait que

$$\lim_{j\to\infty} \log[(1-p_j)^{n_j}] = \lim_{j\to\infty} n_j \log(1-p_j) = \lambda$$

puisque $\log(1-x) \sim -x$ lorsque x tend vers 0. □

Dans l'inégalité de Chernoff II, considérons le cas où $X_j \sim \mathsf{B}(j, 1/j)$. On a alors $\mathbf{E}[X_j] = 1$ pour tout entier j. Dans ce cas, l'inégalité donnée par le théorème s'écrit

$$\mathbf{P}(X_j \geqslant 1+\delta) \leqslant \exp\left(-\frac{\delta^2}{2+\delta}\right)$$

En choisissant $\delta = t-1$ pour un entier t, on a par la proposition précédente avec $X \sim \mathsf{P}(1)$

$$\mathbf{P}(X_j \geqslant 1+\delta) \geqslant \mathbf{P}(X_j = t) \overset{j\to\infty}{\to} \mathbf{P}(X = t) = \frac{1}{t!e} = \exp\Big(-t\log(t)(1+o(1))\Big)$$

Fin cours # 4 du 1er octobre

Concluons avec un dernier résultat de concentration (dont on ne donne pas la preuve) pour une sommes de variables aléatoires indépendants bornées.

Théorème (Inégalité de Hoeffding)**.** *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes, où pour tout i la variable aléatoire X_i est à valeurs dans un intervalle $[a_i, b_i]$. On pose $X = X_1 + \cdots + X_n$ et $\mu = \mathbf{E}[X]$. Alors, pour tout $t > 0$,*

$$\mathbf{P}\left(|X-\mu| \geqslant t\right) \leqslant 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Notons $\ell_i = b_i - a_i$ la longueur de l'intervalle $[a_i, b_i]$. L'inégalité de Hoeffding peut s'interpréter comme suit : alors que l'inégalité triangulaire permet de conclure que *toutes* les valeurs prises par X sont contenues dans un intervalle de longueur $\ell_1 + \cdots + \ell_n$, l'inégalité de Hoeffding implique qu'un intervalle de longueur $O(\sqrt{\ell_1^2 + \cdots + \ell_n^2})$ contient la très grande majorité des valeurs effectivement prises par X. Dans la plupart des cas d'intérêt, comme celui où $\ell_i = 1$, on a

$$\sqrt{\sum \ell_i^2} \ll \sum \ell_i$$

et l'inégalité de Hoeffding est donc plus précise.

2.4 Applications des inégalités de Chernoff

2.5 Partage équilibré

Soit A une matrice $n \times m$ à coefficients dans $\{0,1\}$. On cherche un vecteur $b \in \{-1,1\}^m$ qui minimise la quantité

$$\|Ab\|_\infty = \max_{1 \leqslant i \leqslant n} |(Ab)_i|.$$

Voici une interprétation. Chacune des m colonnes de la matrice correspond à un individu d'une population et chacune des n lignes de la matrice correspond à une caractéristique. La matrice A est déterminée par la condition

$$a_{ij} = 1 \iff \text{l'individu } j \text{ possède la caractéristique } i.$$

On souhaite diviser la population en deux groupes $+1$ et -1, de façon aussi équilibrée que possible pour chacune des caractéristiques. Si on identifie le partage à un vecteur $b \in \{-1,1\}^m$, minimiser $\|Ab\|_\infty$ revient à minimiser le déséquilibre de la caractéristique la plus déséquilibrée.

Une idée naturelle est de faire un partage aléatoire. On a alors

Proposition. *Si b est choisi selon la loi uniforme dans $\{-1,1\}^m$, alors*

$$\mathbf{P}(\|Ab\|_\infty \geqslant \sqrt{4m \log n}) \leqslant 2/n.$$

Démonstration. Par la borne de l'union,

$$\mathbf{P}(\|Ab\|_\infty \geqslant \sqrt{4m \log n}) \leqslant \sum_{i=1}^{n} \mathbf{P}(|(Ab)_i| \geqslant \sqrt{4m \log n})$$

Soit k_i le nombre de 1 dans la ligne i de la matrice A, ou encore le nombre d'individus partageant la caractéristique i. Puisque $|(Ab)_i| \leqslant k_i$, lorsque $k_i < \sqrt{4m \log n}$, on a

$$\mathbf{P}(|(Ab)|_i \geqslant \sqrt{4m \log n}) = 0.$$

Si $k_i \geqslant \sqrt{4m \log n}$, on a en utilisant l'inégalité de Chernoff puis le fait que $k_i \leqslant m$

$$\mathbf{P}(|(Ab)_i| \geqslant \sqrt{4m \log n}) \leqslant 2 \exp\left(-\frac{4m \log n}{2k_i}\right) \leqslant 2 \exp\left(-\frac{4m \log n}{2m}\right) = \frac{2}{n^2}. \qquad \square$$

2.6 Répartition entre serveurs

Cet exemple est similaire au précédent, avec un partage en plus de 2 groupes. Supposons que n tâches doivent être attribuées à k serveurs. Lorsque les tâches sont attribuées au hasard (uniformément, indépendamment), quel est la charge maximale d'un serveur ? Cette dernière est toujours au moins n/k, mais quelle valeur prend-elle dans une situation typique ?

Pour $1 \leqslant i \leqslant k$, soit X_i le nombre de tâches assignées au serveur i. Chacune des variables aléatoire X_i suit la loi binomiale $\mathsf{B}(n, 1/k)$. On notera que ces variables ne sont pas indépendantes. On s'intéresse à la charge maximale $M = \max(X_1, \ldots, X_k)$ dans deux régimes différents : d'abord quand $n = k$ puis quand $n \gg k$.

Quand $n = k$, la charge de chaque serveur est bien approximée par une loi de Poisson de paramètre 1. On peut montrer que le maximum des n variables aléatoires i.i.d. de loi $\mathsf{P}(1)$ est de l'ordre de $\frac{\log n}{\log \log n}$ avec probabilité tendant vers 1 quand n tend vers l'infini. Même si les charges entre serveurs ne sont pas indépendantes, cette heuristique est correcte. On a

Proposition. *Si on note $M^{(n)}$ la charge maximale d'un serveur quand n tâches sont affectées aléatoirement à n serveurs, alors*

$$\lim_{n\to\infty} \mathbf{P}\left(M^{(n)} \geqslant \frac{e\log n}{\log\log n}\right) = 0.$$

La preuve montre en réalité plus (exercice) : pour tout $\varepsilon > 0$, on a

$$\lim_{n\to\infty} \mathbf{P}\left(M^{(n)} \geqslant \frac{(1+\varepsilon)\log n}{\log\log n}\right) = 0.$$

On peut aussi démontrer (nous ne le ferons pas) que

$$\lim_{n\to\infty} \mathbf{P}\left(M^{(n)} \leqslant \frac{(1-\varepsilon)\log n}{\log\log n}\right) = 0.$$

Démonstration. Écrivons $M^{(n)} = \max(X_1^{(n)}, \dots, X_n^{(n)})$ où les variables $X_i^{(n)}$ suivent la loi binomiale $\mathsf{B}(n, 1/n)$ (et ne sont pas indépendantes). Pour tout entier d, on a par la borne de l'union

$$\mathbf{P}(M^{(n)} \geqslant d) \leqslant n\mathbf{P}(X_1^{(n)} \geqslant d) \leqslant n\binom{n}{d}\left(\frac{1}{n}\right)^d.$$

La seconde inégalité s'explique par le fait que l'événement «$X_1^{(n)} \geqslant d$» est la réunion, pour $I \subset \{1,\dots,n\}$ de cardinal d, des événements «pour tout $i \in I$, la $i$ème tâche a été affectée au premier serveur». On utilise ensuite les inégalités $\binom{n}{d} \leqslant \frac{n^d}{d!}$ et $\frac{d^d}{d!} \leqslant e^d$ pour obtenir

$$\mathbf{P}(M^{(n)} \geqslant d) \leqslant n\left(\frac{e}{d}\right)^d$$

Pour $d = \left\lceil \frac{e\log n}{\log\log n} \right\rceil$, on a donc

$$\begin{aligned}\mathbf{P}\left(M^{(n)} \geqslant \frac{e\log n}{\log\log n}\right) &\leqslant n\left(\frac{\log\log n}{\log n}\right)^{\frac{e\log n}{\log\log n}} \\ &\leqslant \exp\left(\log n - e\log n + e\frac{\log n\cdot\log\log\log n}{\log\log n}\right)\end{aligned}$$

et cette quantité tend vers 0 car le terme dominant dans l'exponentielle est $(1-e)\log n$. □

Dans le régime où $n \gg k$, on a par exemple le résultat suivant.

Proposition. *Si $n \geqslant 9k\log k$, alors*

$$\mathbf{P}\left(M \geqslant \frac{n}{k} + 3\sqrt{\log k}\sqrt{n/k}\right) \leqslant \frac{1}{k^2}$$

Démonstration. On utilise la borne de l'union et le fait que les variables aléatoires (X_i) sont identiquement distribuées pour écrire

$$\mathbf{P}\left(M \geqslant \frac{n}{k} + 3\sqrt{\log k}\sqrt{n/k}\right) \leqslant k\mathbf{P}\left(X_1 \geqslant \frac{n}{k}(1+\varepsilon)\right)$$

avec $\varepsilon = 3\sqrt{\log k}/\sqrt{n/k}$. Sous l'hypothèse de la proposition, on a $\varepsilon \leqslant 1$. Par l'inégalité de Chernoff II, on peut donc écrire

$$\mathbf{P}\left(X_1 \geqslant \frac{n}{k}(1+\varepsilon)\right) \leqslant \exp\left(-\frac{n}{k}\frac{\varepsilon^2}{3}\right) = 1/k^3,$$

d'où le résultat. □

2.7 Graphes aléatoires

Les graphes de la vie réelle (internet, réseaux sociaux...) sont souvent très compliqués et peuvent être appréhendés par l'étude de graphes aléatoires. On se contera ici du modèle le plus simple dans lequel tous les sommets jouent un rôle symétrique.

Étant donné deux paramètres $n \in \mathbf{N}$ et $p \in [0,1]$, le graphe d'Erdős–Rényi est défini comme suit. On part d'une famille $(X_{ij})_{1 \leqslant i < j \leqslant n}$ de variables aléatoires i.i.d. de loi de Bernoulli $\mathsf{B}(p)$ et on considère le graphe $G = (V, E)$ où $V = \{1, \dots, n\}$ et E est défini par

$$\{i, j\} \in E \iff X_{ij} = 1.$$

Le graphe ainsi obtenu est aléatoires (c'est une variable aléatoire à valeurs dans l'ensemble des graphes possibles) et on note $\mathsf{G}_{n,p}$ sa loi.

Remarquons que $\mathsf{G}_{n,1/2}$ est la loi uniforme sur l'ensemble de tous les graphes de sommets $\{1, \dots, n\}$. Le nombre d'arêtes $|E|$ est distribué selon la loi $\mathsf{B}(\binom{n}{2}, p)$. Le degré de chaque sommet est distribué selon la loi $\mathsf{B}(n-1, p)$.

On étudie en général le graphe d'Erdős–Rényi dans la limite $n \to \infty$ en distinguant plusieurs régimes, comme par exemple

- le cas où p est constant ; on a alors un graphe dense qui contient $\Omega(n^2)$ arêtes avec grande probabilité (conséquence des inégalités de Chernoff)
- le cas où $p = \Theta(1/n)$; on a alors un graphe creux où le degré d'un sommet est approximé par une loi de Poisson.

Théorème. *Soit $c > 0$ fixé et posons $p = c\log(n)/n$ et soit G_n un graphe aléatoire de loi $\mathsf{G}_{n,p}$. Alors*

- *Si $c < 1$, alors*
$$\lim_{n\to\infty} \mathbf{P}(G_n \text{ a un sommet isolé}) = 1$$
- *Si $c > 1$, alors*
$$\lim_{n\to\infty} \mathbf{P}(G_n \text{ a un sommet isolé}) = 0$$

Démonstration. Soit N le nombre de sommets isolés. Par linéarité de l'espérance

$$\mathbf{E}[N] = n(1-p)^{n-1} = n\exp\left(n\ln(1-p)\right)/(1-p) \sim \frac{n}{1-p}\exp(-c\ln n) \sim \frac{n^{1-c}}{1-p}.$$

Si $c > 1$, alors $\mathbf{E}[N] \to 0$ et donc $\mathbf{P}(N \geqslant 1) \leqslant \mathbf{E}[N] \to 0$.

Si $c < 1$, alors $\mathbf{E}[N] \to \infty$ mais cela ne suffit pas à conclure. On peut écrire par l'inégalité de Tchebycheff

$$\mathbf{P}(N = 0) \leqslant \mathbf{P}\left(|N - \mathbf{E}[N]| \geqslant \mathbf{E}[N]\right) \leqslant \frac{\mathbf{Var}(N)}{\mathbf{E}[N]^2} = \frac{\mathbf{E}[N^2]}{\mathbf{E}[N]^2} - 1$$

et on est ramené à montrer que $\mathbf{E}[N^2] \sim \mathbf{E}[N]^2$. On calcule donc

$$\begin{aligned}\mathbf{E}[N^2] &= \mathbf{E}\left[\sum_{i,j} \mathbf{1}_{\{i \text{ isolé et} j \text{ isolé.}\}}\right] \\ &= n\mathbf{P}(1 \text{ isolé}) + n(n-1)\mathbf{P}(1 \text{ et } 2 \text{ isolés}) \quad = n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3}\end{aligned}$$

et donc

$$\frac{\mathbf{E}[N^2]}{\mathbf{E}[N]^2} = \frac{1}{n(1-p)^{n-1}} + \frac{n-1}{n(1-p)}$$

tend bien vers 1. □

Fin cours #5 du 8 octobre

On peut en réalité montrer mieux.

Théorème. *Sous les hypothèses du théorème précédent, si* $c > 1$

$$\lim_{n\to\infty} \mathbf{P}(G_n \text{ est connexe}) = 1$$

Démonstration. Remarquons que G_n est non connexe si et seulement si il existe un sous ensemble $S \subset V$ avec $|S| \leqslant n/2$ sans arête entre S et $V \setminus S$. On a donc

$$\mathbf{P}(G_n \text{ non connexe}) \leqslant \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k} (1-p)^{k(n-k)}$$

Pour simplifier l'analyse on suppose $c > 2$ (exercice : montrer le résultat sous l'hypothèse $c > 1$). On a en écrivant $1 - x \leqslant e^{-x}$ et $\binom{n}{k} \leqslant n^k$

$$\mathbf{P}(G_n \text{ non connexe}) \leqslant \sum_{k=1}^{\lfloor n/2 \rfloor} \underbrace{n^k \exp\left(-pk(n-k)\right)}_{\alpha}$$

Comme $\log(\alpha) \leqslant k \log n - \frac{c \log n}{n}(n-k) \leqslant k \log n(1 - c/2)$, on a

$$\mathbf{P}(G_n \text{ non connexe}) \leqslant \sum_{k=1}^{\infty} n^{k(1-c/2)} = \frac{n^{1-c/2}}{1 - n^{1-c/2}} \to 0$$

d'où le résultat. □

Chapitre 3

Convergence des variables aléatoires et théorème central limite

3.1 Convergence presque sûre et loi forte des grands nombres

Lorsque (X_n) est une suite de variables aléatoires, il y a plusieurs notions possibles de convergence de la suite (X_n) vers une variable aléatoire X.

Il y a une notion de convergence déjà rencontrée : la *convergence en probabilité*. On dit que (X_n) converge en probabilité vers X si

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbf{P}(|X_n - X| > \varepsilon) = 0.$$

Ainsi, lorsque (X_n) est une suite de variables aléatoires i.i.d. ayant un moment d'ordre 2, la loi faible des grands nombres s'énonce en disant que la suite (S_n) des moyennes de Cesáro converge en probabilité vers la variable aléatoire constante égale à $\mathbf{E}[X_1]$.

Un autre notion de convergence est la notion de convergence presque sûre. On dit que (X_n) converge presque sûrement vers X si

$$\mathbf{P}(\{\lim X_n = X\}) = 1.$$

Proposition. *Si (X_n) converge vers X presque sûrement, alors (X_n) converge vers X en probabilité.*

Démonstration. Fixons $\varepsilon > 0$. Pour $m \in \mathbf{N}$, on considère l'événement

$$A_m = \{\exists n \geqslant m \ : \ |X_n - X| > \varepsilon\}$$

C'est une suite décroissante d'événements ; il découle de la σ-additivité (considérer les événements complémentaires) que

$$\mathbf{P}\left(\bigcap_{m\geqslant 1} A_m\right) = \lim_{m\to\infty} \mathbf{P}(A_m).$$

Mais

$$\mathbf{P}\left(\bigcap_{m\geqslant 1} A_m\right) \leqslant \mathbf{P}\left(\text{la suite } (X_n) \text{ ne converge pas vers } X\right) = 0.$$

On a donc $\mathbf{P}(|X_m - X| > \varepsilon) \leqslant \mathbf{P}(A_m) \to 0$. □

Pour bien comprendre la différence, voici un exemple de suite qui converge en probabilité mais pas presque sûrement. Soit (X_i) une suite de bits aléatoires, c'est-à-dire de variables aléatoires i.i.d. de loi $\mathsf{B}(1/2)$. On considère l'ensemble $\{0,1\}^*$ des mots finis sur l'alphabet $\{0,1\}$. C'est un ensemble dénombrable, que l'on écrit comme une suite (w_n) en l'ordonnant de façon arbitraire (pour fixer les idées, on peut l'ordonner par longueur de mot, puis par ordre lexicographique pour les mots de même longueur). On note Y_n la variable aléatoire à valeurs $\{0,1\}$ qui vaut 1 si et seulement si w_n est un segment initial de la suite (X_i).

Alors (Y_n) converge en probabilité vers la variable aléatoire constante égale à 0, puisque pour tout $0 < \varepsilon < 1$,

$$\mathbf{P}(|Y_n| > \varepsilon) = \mathbf{P}(Y_n = 1) = \frac{1}{2^{|w_n|}}$$

tend vers 0 quand n tend vers l'infini. Mais il n'est pas vrai que (Y_n) converge presque sûrement vers 0 puisque la suite (Y_n) admet une sous-suite (aléatoire) dont tous les termes sont égaux à 1, celle obtenue en prenant comme mots les segments initiaux de la suite (X_i).

Néanmoins, dans le cas de la loi des grands nombres, on a le résultat suivant.

Théorème (Loi forte des grands nombres)**.** *Soit (X_n) une suite de variables aléatoires i.i.d. admettant un moment d'ordre 1. Posons $\mu = \mathbf{E}[X_1]$ et $S_n = X_1 + \cdots + X_n$. Alors la suite (S_n/n) converge presque sûrement vers la variable aléatoire constante égale à μ.*

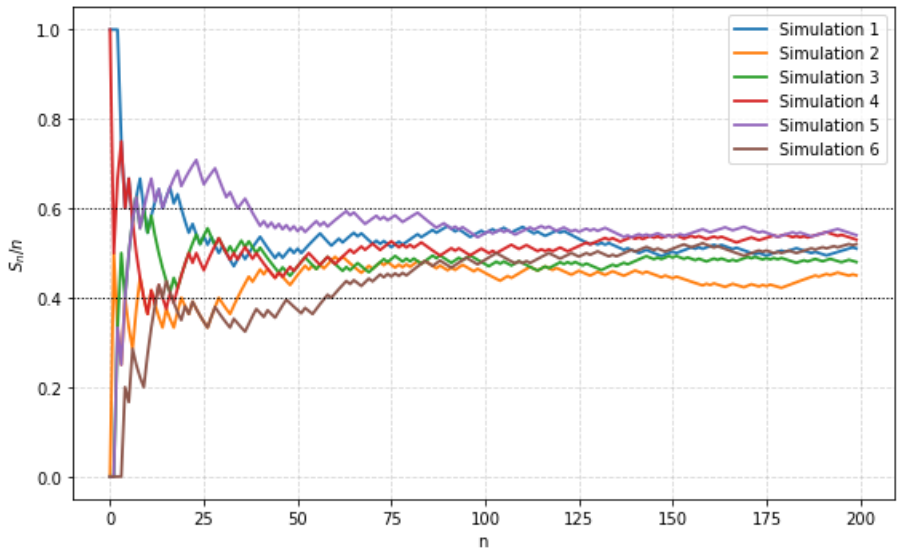


FIGURE 3.1 – Loi des grands nombres pour une somme de variables aléatoires de loi de Bernoulli $\mathsf{B}(1/2)$

La loi des grands nombres est illustrée dans la figure 3.1. La loi faible des grands nombres affirme que la proportion de simulations qui sont dans la bande délimitée par les deux lignes pointillées d'ordonnée $\mu - \varepsilon$ et $\mu + \varepsilon$ tend vers 1 quand n tend vers l'infini. La loi forte des grands nombres affirme que (presque) toute simulation est confinée dans cette bande pour n assez grand.

On a déjà démontré la loi faible sous l'hypothèse que X_1 admet un moment d'ordre 2. On va maintenant expliquer comment montrer la loi forte.

Lemme. *Soient (X_n) et X des variables aléatoires. Alors*

$$X_n \overset{p.s.}{\to} X \iff \forall \varepsilon > 0,\ \mathbf{P}\left(|X_n - X| > \varepsilon \text{ pour une infinité d'indices } n\right) = 0$$

Démonstration. L'implication directe est immédiate. Pour la réciproque, on l'applique à $\varepsilon = 1/k$ pour tout $k \in \mathbf{N}^*$ et on utilise le fait qu'une union dénombrable d'événements de mesure nulle est de mesure nulle. □

Lemme (Lemme de Borel–Cantelli)**.** *Soit (A_n) une suite d'événements tels que la série $\sum \mathbf{P}(A_n)$ converge. Alors*

$$\mathbf{P}(\text{ une infinité des événements } (A_n) \text{ est vraie }) = 0.$$

Démonstration. Soit E l'événement en question. Alors

$$\mathbf{P}(E) \leqslant \mathbf{P}\left(\bigcap_{m\geqslant 1} \bigcup_{n\geqslant m} A_n\right) = \lim_{m\to\infty} \mathbf{P}\left(\bigcup_{n\geqslant m} A_n\right) \leqslant \sum_{n=m}^{\infty} \mathbf{P}(A_n)$$

qui tend vers 0 comme reste d'une série convergente. □

Pour résumer,

1. Si $\forall \varepsilon > 0$ on a $\mathbf{P}(|X_n - X| > \varepsilon) \to 0$ alors (X_n) converge vers X en probabilité (c'est la définition)
2. Si $\forall \varepsilon > 0$ on a $\sum \mathbf{P}(|X_n - X| > \varepsilon) < \infty$ alors (X_n) converge vers X presque sûrement (on peut appliquer le lemme de Borel–Cantelli).

En un sens, la différence entre ces deux notions de convergence est similaire à la différence entre le fait qu'une série converge et le fait que son terme général tend vers 0.

Preuve de la loi forte des grands nombres sous l'hypothèse de 4ème moment fini. On peut (quitte à remplacer X_n par $X_n - \mu$) supposer que $\mu = 0$. On calcule alors

$$\mathbf{E}[S_n^4] = \sum_{i,j,k,l} \mathbf{E}[X_i X_j X_k X_l]$$

En utilisant l'indépendance et le fait que $\mu = 0$, on observe qu'un terme $\mathbf{E}[X_i X_j X_k X_l]$ est nul en dehors des cas suivants

— $i = j = k = l$
— $i = j$ et $k = l$
— $i = k$ et $j = l$
— $i = l$ et $j = k$

On a donc

$$\mathbf{E}[S_n^4] = n\,\mathbf{E}[X_1^4] + 3n(n-1)\,\mathbf{E}[X_1^2]^2 \leqslant Cn^2$$

pour une constante $C > 0$. Ainsi, $\mathbf{E}[(S_n/n)^2] \leqslant C/n^2$ et l'inégalité de Markov permet de conclure que, pour tout $\varepsilon > 0$

$$\forall \varepsilon > 0,\ \mathbf{P}(|S_n/n|^4 \geqslant \varepsilon) \leqslant \frac{C}{n^2 \varepsilon}.$$

Puisque la série $\sum \frac{C}{n^2\varepsilon}$ est convergente, on conclut à l'aide du lemme de Borel–Cantelli. □

Dans la loi des grands nombres, la limite est une variable aléatoire constante. Voici un exemple simple de convergence presque sûre vers une variable aléatoire non-constante.

Soit U une variable aléatoire de loi uniforme sur l'intervalle $[0,1]$. On considère une suite (p_n) de réels dans $[0,1]$ qui converge vers p. Si on considère les variables aléatoires

$$X_n = \begin{cases} 1 \text{ si } U \leqslant p_n \\ 0 \text{ si } U > p_n \end{cases} \qquad X = \begin{cases} 1 \text{ si } U \leqslant p \\ 0 \text{ si } U > p \end{cases}$$

alors (X_n) converge presque sûrement vers X (le seul cas où on peut avoir $\lim X_n \neq X$ est le cas où $U = p$, qui est un événement de probabilité nulle).

3.2 Convergence en distribution et théorème central limite

La convergence en distribution (ou convergence en loi) s'intéresse aux variables aléatoires uniquement à travers leur loi.

Définition. Soient (X_n) et X des variables aléatoires. On dit que (X_n) converge vers X en distribution si, pour tout t point de continuité de $t \mapsto \mathbf{P}(X \leqslant t)$,

$$\lim_{n\to\infty} \mathbf{P}(X_n \leqslant t) = \mathbf{P}(X \leqslant t).$$

Remarquons que pour définir la convergence en distribution, les variables aléatoires (X_n) et X n'ont pas besoin d'être définies sur le même espace de probabilité. Cette notion dépend seulement des lois de (X_n) et X. En particulier, si (X_n) converge en distribution vers X et si $X \sim Y$, alors (X_n) converge en distribution vers Y.

Lemme. *Si (X_n) converge vers X en probabilité, alors (X_n) converge vers X en distribution.*

Démonstration. On note $F_X(t) = \mathbf{P}(X \leqslant t)$. Soit t un point de continuité de F_X. Pour tout $\varepsilon > 0$, il existe $\alpha > 0$ tel que $F_X(t-\alpha) \geqslant F_X(t) - \varepsilon$ et $F_X(t+\alpha) \leqslant F_X(t) + \varepsilon$. Pour n assez grand, on a $\mathbf{P}(|X_n - X| > \alpha) \leqslant \varepsilon$, d'où

$$\mathbf{P}(X_n \leqslant t) \leqslant \mathbf{P}(X \leqslant t+\alpha) + \mathbf{P}(|X_n - X| > \alpha) \leqslant F_X(t) + 2\varepsilon$$

$$\mathbf{P}(X_n \leqslant t) \geqslant \mathbf{P}(X \leqslant t-\alpha) - \mathbf{P}(|X_n - X| > \alpha) \geqslant F_X(t) - 2\varepsilon$$

d'où le résultat. □

Théorème (Théorème de Lévy, admis)**.** *Soient (X_n) et X des variables aléatoires. On a l'équivalence entre*

1. *(X_n) converge vers X en distribution,*
2. *Pour tout $t \in \mathbf{R}$, on a*
$$\lim_{n\to\infty} \mathbf{E}[e^{itX_n}] = \mathbf{E}[e^{itX}]$$

La fonction $\Phi_X : t \mapsto \mathbf{E}[e^{itX}]$ s'appelle la fonction caractéristique de X ; c'est l'analogue de la transformée de Fourier en analyse. Elle partage cette propriétés de la fonction génératrice des moments, comme l'identité $\Phi_{X+Y} = \Phi_X \Phi_Y$ lorsque X et Y sont indépendantes, mais elle est toujours définie même sans aucune hypothèse d'existence de moments.

Soit (X_n) une suite de variables aléatoires admettant un moment d'ordre 2 et vérifiant $\mathbf{E}[X_1] = 0$. Posons $S_n = X_1 + \cdots + X_n$. Par la loi forte des grands nombres on a presque sûrement $S_n = o(n)$. Peut-on préciser le développement asymptotique de S_n ? Puisque $\mathbf{Var}(S_n) = n\,\mathbf{Var}(X_1)$, on a $\mathbf{Var}(S_n/\sqrt{n}) = \mathbf{Var}(X_1)$ et on s'attend à ce que S_n soit de

l'ordre de $\sqrt{n}$. C'est bien le cas, mais ce terme d'ordre $\sqrt{n}$ est aléatoire et fait intervenir la loi gaussienne.

On appelle loi gaussienne (ou normale) standard ou $\mathsf{N}(0,1)$ la loi de densité

$$x \mapsto \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Si $X \sim \mathsf{N}(0,1)$, alors $\mathbf{E}[X] = 0$ et $\mathbf{Var}[X] = \sigma^2$. Plus généralement, étant donnés des réels m et σ, on note $\mathsf{N}(m, \sigma^2)$ ma loi de densité

$$x \mapsto \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}.$$

Si X suit la loi $\mathsf{N}(0,1)$, alors la variable aléatoire $Y := m + \sigma X$ suit loi loi $\mathsf{N}(m, \sigma^2)$.

Fin cours #6 du 15 octobre

Théorème (Théorème central limite)**.** *Soit (X_n) une suite de variables aléatoires i.i.d. admettant un moment d'ordre 2. On pose $\mu = \mathbf{E}[X_1]$ et $\sigma = \sqrt{\mathbf{Var}(X_1)}$, supposé > 0. Soit $S_n = X_1 + \cdots + X_n$. Alors, la suite*

$$\left(\frac{S_n - \mu n}{\sigma\sqrt{n}} \right)$$

converge en distribution vers une variable de loi $\mathsf{N}(0,1)$.

C'est un résultat d'universalité : la limite ne dépend pas de X_1 mais uniquement de sa variance. Remarquons que la condition $\sigma > 0$ équivaut à dire que X n'est pas constante.

Si Z suit la loi $\mathsf{N}(0,1)$, alors la fonction

$$t \mapsto \mathbf{P}(Z \leqslant t) = \int_{-\infty}^{t} \exp(-x^2/2) \frac{\mathrm{d}x}{\sqrt{2\pi}}$$

est continue. La conclusion du théorème central limite peut donc s'écrire ainsi : pour tout $t \in \mathbf{R}$,

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{S_n - \mu n}{\sqrt{n}} \leqslant t \right) = \mathbf{P}(Z \leqslant t)$$

Un calcul élémentaire montre que la fonction caractéristique d'une variable aléatoire $Z \sim \mathsf{N}(0,1)$ est donnée par

$$\Phi_Z(t) = e^{-t^2/2}$$

(le plus simple pour le montrer est d'observer que Φ_Z est solution de l'équation différentielle $y'(t) = -ty(y)$ à l'aide d'une intégration par parties).

Démonstration. On peut supposer que $\mu = 0$ et $\sigma = 1$, quitte à remplacer X_n par $\frac{X_n - \mu}{\sigma}$. On effectue ensuite un développement limité de la fonction caractéristique au voisinage de 0. L'approximation $e^{itX_1} = 1 + itX_1 - \frac{t^2}{2}X_1^2 + o(t^2)$ implique (cela ce justifie par le théorème de convergence dominée) que $\Phi_{X_1}(t) = \mathbf{E}[e^{itX_1}] = 1 - t^2/2 + o(t^2)$.

On a en utilisant l'indépendance des (X_n) que

$$\Phi_{S_n/\sqrt{n}}(t) = \Phi_{S_n}(t/\sqrt{n}) = \Phi_{X_1}(t/\sqrt{n})^n = (1 - t^2/2n + o(1/n))^n = \exp(-t^2/2) + o(1),$$

puis on conclut à l'aide du théorème de Lévy. □

Chapitre 4

La méthode probabiliste : exemples

La méthode probabiliste montre l'existence d'objets (souvent de nature combinatoire, mais pas uniquement) en montrant qu'un choix aléatoire convient avec probabilité > 0. Nous allons illustrer ce principe sur 4 exemples, de complexité croissante.

4.1 Exemple 1 : satisfiabilité

On appelle formule k-SAT une formule booléenne qui est une conjonction de clauses, chaque clause étant la disjonction de k variables ou leur négation, ces k variables étant 2 à 2 distinctes. Une telle formule est du type

$$(x_1 \vee \overline{x_3} \vee x_4) \wedge (x_5 \vee x_6 \vee \overline{x_8}) \wedge \dots$$

Le problème de satisfiabilité demande s'il existe une affectation des variables booléennes rendant vraie la formule ci-dessous. C'est un problème NP-difficile pour $k \geqslant 3$.

Une variante est de demander combien de clauses peuvent être satisfaites. On a alors le résultat suivant.

Proposition. *Soit une formule k-SAT écrite comme la disjonction de m clauses. Il existe une affectation des variables qui satisfait au moins $m(1 - 2^{-k})$ des clauses.*

Pour $k = 3$, cela montre qu'il est toujours possible de satisfaire une proportion $7/8$ des clauses d'une formule 3-SAT. La preuve est très simple.

Démonstration. On affecte au hasard les valeurs des variables, indépendamment et uniformément sur l'ensemble $\{\texttt{vrai}, \texttt{faux}\}$. Pour toute clause C_i, par indépendance, l'événement «la clause C_i est satisfaite» a probabilité $1 - 2^{-k}$. On a donc, par linéarité de l'espérance

$$\mathbf{E}[\text{nombre de clauses satisfaites}] = m(1 - 2^{-k}),$$

d'où le résultat. □

La preuve utilise le principe suivant : si une variable aléatoire X intégrable a pour espérance μ, alors $\mathbf{P}(X \geqslant \mu) > 0$ (dans notre cas, X est le nombre de clauses satisfaites).

4.2 Exemple 2 : nombres de Ramsey

On note $R(k, l)$ l'entier n minimal tel que tout coloriage des arêtes du graphe complet K_n en deux couleurs (rouge et bleu) contient un sous-graphe K_k dont toutes les arêtes sont rouges ou un sous-graphe K_l dont toutes les arêtes sont bleues.

On calcule par exemple que $R(2,2) = 2$ et $R(3,3) = 6$.

Exercice. Montrer l'inégalité $R(k,l) \leqslant R(k-1,l) + R(k,l-1)$ et en déduire que $R(k,l) \leqslant 2^{k+l}$ et en particulier $R(k,k) \leqslant 4^k$.

Voici une borne inférieure

Proposition. *Si $k \geqslant 3$, alors $R(k,k) > \lfloor 2^{k/2} \rfloor$*

Démonstration. On considère un coloriage aléatoire du graphe complet $K_n = (V_n, E_n)$ où chaque arête est coloriée en rouge ou bleu aléatoirement, uniformément et indépendemment.

Si $S \subset V_n$ est un sous-ensemble de taille k, alors

$$\mathbf{P}(S \text{ est monochromatique}) = 2 \cdot 2^{-\binom{k}{2}}.$$

Par la borne de l'union, on en déduit

$$\begin{aligned}\mathbf{P}(\exists S \subset V_n \text{ monochromatique de taille } k) &\leqslant \binom{n}{k} 2^{1-\binom{k}{2}} \\ &\leqslant \frac{n^k}{k!} 2 \cdot 2^{-\frac{k(k-1)}{2}}\end{aligned}$$

En choisissant $n = \lfloor 2^{k/2} \rfloor$, cette quantité est $\leqslant \frac{2 \cdot 2^{k/2}}{k!} < 1$ pour $k \geqslant 3$, d'où le résultat : il existe un coloriage de K_n sans clique monochromatique de taille k. □

L'argument précédent peut être réécrit comme un argument de comptage, mais le point de vue probabiliste est en général plus fructueux.

Un problème ouvert important est de déterminer la limite

$$\ell = \lim_{k \to \infty} R(k,k)^{1/k}$$

(il n'est pas clair que la limite existe). La proposition implique $\ell \geqslant 1/2$ (Erdős 1947) et l'exercice $\ell \leqslant 4$ (Ramsey 1929). Un progrès remarquable récent (2023) améliore cette borne en $\ell \leqslant 4 - \varepsilon$ avec ε de l'ordre de 2^{-10}.

4.3 Exemple 3 : borne inférieure pour le problème de partage équilibré

On rappelle qu'on a montré le résultat suivant : étant donnée A une matrice $n \times n$ à coefficients dans $\{0,1\}$, alors si b est choisi uniformément dans $\{-1,1\}^n$,

$$\mathbf{P}(\|Ab\|_\infty \leqslant \sqrt{4n \log n}) \to 1.$$

Nous allons voir que cette estimation en $\sqrt{4n \log n}$ pour le meilleur partage équilibré est essentiellement optimale.

Proposition. *Il existe une constante réelle $c > 0$, un entier n_0 et pour tout $n \geqslant n_0$ une matrice $A_n \in \{0,1\}^{n \times n}$ telle que*

$$\min_{b \in \{-1,1\}^n} \|A_n b\|_\infty \geqslant c\sqrt{n}$$

On va bien sûr choisir A_n au hasard en prenant pour coefficients des bits aléatoires ! On utilisera le lemme suivant

Lemme. *Il existe une constante réelle $c > 0$ et un entier n_0 tels que, pour tout $n \geqslant n_0$, si $b_1, \dots, b_n$ sont dans $\{-1, 1\}$ (fixés) et $X_1, \dots, X_n$ sont i.i.d. de loi $\mathsf{B}(1/2)$, alors*

$$\mathbf{P}\left(\left|\sum_{i=1}^{n} b_i X_i\right| \leqslant c\sqrt{n}\right) < 1/2$$

Preuve de la proposition. Soit $A = (a_{ij})$ une matrice de coefficients i.i.d. de loi $\mathsf{B}(1/2)$. Pour tout $b \in \{-1, 1\}^n$, on a par indépendance des lignes de A

$$\mathbf{P}(\|Ab\|_\infty < c\sqrt{n}) = \mathbf{P}(\forall i,\ |(Ab)_i| < c\sqrt{n}) < (1/2)^n.$$

Soit N le nombre de $b \in \{-1, 1\}^n$ tels que $\|Ab\|_\infty < c\sqrt{n}$. Par linéarité de l'espérance,

$$\mathbf{E}[N] < 2^n (1/2)^n = 1$$

et donc il existe A tel que $N = 0$, ce qui veut dire que $\|Ab\|_\infty \geqslant c\sqrt{n}$ pour tout $b \in \{-1, 1\}^n$. □

Preuve du lemme. Posons

$$Y_i = b_i X_i + \frac{1 - b_i}{2} = \begin{cases} X_i & \text{si } b_i = 1 \\ 1 - X_i & \text{si } b_i = -1 \end{cases}$$

et remarquons que les v.a. (Y_i) sont i.i.d. de loi $\mathsf{B}(1/2)$. Soit $S_n = Y_1 + \cdots + Y_n$ (qui suit une loi binomiale $\mathsf{B}(n, 1/2)$ et x l'entier $\frac{n}{2} - \frac{b_1 + \cdots + b_n}{2}$. On a, pour tout entier ℓ

$$\left|\sum b_i X_i\right| \leqslant \ell \iff \left|-x + \sum Y_i\right| \leqslant \ell \iff S \in [x - \ell, x + \ell]$$

Puisque la fonction $k \mapsto \binom{n}{k}$ est croissante pour $k \leqslant \frac{n}{2}$ et décroissante pour $k \geqslant \frac{n}{2}$, la quantité $\mathbf{P}(S \in [x - \ell, x + \ell])$ est maximale pour $x = \lfloor n/2 \rfloor$. Il s'ensuit que

$$\mathbf{P}\left(\left|\sum b_i X_i\right| \leqslant c\sqrt{n}\right) \leqslant \mathbf{P}\left(|S - \lfloor n/2 \rfloor| \leqslant c\sqrt{n}\right) =: \alpha_n$$

Par le théorème central limite, on a pour Z de loi $\mathsf{N}(0, 1)$

$$\lim_{n \to \infty} \alpha_n = \mathbf{P}(|Z| \leqslant c) = \int_{-c/2}^{c/2} \exp(-x^2) \frac{\mathrm{d}x}{\sqrt{2\pi}}$$

et cette quantité peut être rendue $< 1/2$ en choisissant la constante c suffisamment petite. On a donc $\alpha_n < 1/2$ pour n assez grand. □

Fin cours #7 du 22 octobre

4.4 Le lemme local de Lovász

Lorsqu'on utilise la méthode probabiliste, on veut prouver que certains événements «mauvais» $A_1, \dots, A_n$ sont simultanément évités avec probabilité non nulle. Il y a deux idées simples pour cela

- La borne de l'union : si $\sum \mathbf{P}(A_i) < 1$ alors $\mathbf{P}(\overline{A_1} \cap \cdots \cap \overline{A_n}) > 0$,
- L'indépendance : si les événements (A_i) sont indépendants et vérifient $\mathbf{P}(A_i) < 1$, alors $\mathbf{P}(\overline{A_1} \cap \cdots \cap \overline{A_n}) > 0$.

Le lemme local de Lovász combine de manière astucieuse ces deux situations. Soient $A, A_1, \dots, A_n$ des événements. On dit que A est indépendant de $\{A_1, \dots, A_n\}$ si pour tout $I \subset \{1, \dots, n\}$ tel que $\mathbf{P}(\bigcap_{j\in I} A_j) > 0$, on a

$$\mathbf{P}\left(A \,\middle|\, \bigcap_{j\in I} A_j\right) = \mathbf{P}(A).$$

Cette condition est satisfaite lorsque les événements $A, A_1, \dots, A_n$ sont indépendants, mais elle est plus faible : par exemple, elle n'implique pas que A_1 et A_2 sont indépendants.

Soit $(A_i)_{i\in V}$ une famille d'événements. Un *graphe de dépendance* est un graphe non orienté $G = (V, E)$ tel que, pour tout $i \in V$, l'événement A_i est indépendant de $\{A_j \ : (i, j) \notin E\}$.

Théorème (Lemme Local de Lovász)**.** *Soient $A_1, \dots, A_n$ des événements tels que*

1. *Pour tout i, on a $\mathbf{P}(A_i) \leqslant p$,*
2. *Les événements (A_i) admettent un graphe de dépendance dans lequel tout sommet a degré $\leqslant d$,*
3. *$4dp \leqslant 1$.*

Alors $\mathbf{P}(\overline{A_1} \cap \dots \cap \overline{A_n}) > 0$.

Commençons par donner une application du lemme local de Lovász.

Proposition. *Soit $k \geqslant 4$. Une forme k-SAT où chaque variable apparaît au plus $\frac{2^k}{4k}$ fois est satisfiable.*

L'énoncé est trivial pour $k = 4$ (une formule où chaque variable n'apparaît qu'une fois est évidemment satisfiable). Pour $k = 8$, on obtient qu'une formule 8-SAT où chaque variable apparaît au plus 8 fois est satisfiable.

Démonstration. Soient $C_1, \dots, C_N$ les clauses apparaissant dans la formule. On assigne les valeurs booléennes des variables indépendamment et uniformément. Soit A_i l'événement «La clause C_i n'est pas satisfaite». On a $\mathbf{P}(A_i) = 2^{-k} = p$.

Considérons le graphe $G = (V, E)$ où $V = \{1, \dots, N\}$ et $(i, j) \in E$ si les clauses C_i et C_j ont une variable en commun. C'est un graphe de dépendance pour les événements $(A_i)_{1\leqslant i\leqslant N}$ dont le degré est $\leqslant k\frac{2^k}{4k} = \frac{2^k}{4} = d$.

Puisque $4pd \leqslant 1$, le lemme local de Lovász s'applique et nous pouvons conclure que

$$\mathbf{P}(\overline{A_1} \cap \dots \cap \overline{A_N}) > 0,$$

d'où le résultat. □

La preuve du lemme local de Lovász repose sur une récurrence astucieuse.

Preuve du lemme local de Lovász. Il est commode de noter $B_i = \overline{A_i}$ et $B_S = \bigcap_{i\in S} \overline{A_i}$ pour $S \subset \{1, \dots, n\}$. On montre par récurrence sur $s \in \{0, \dots, n\}$ que si $|S| \leqslant s$ alors

$$\mathbf{P}(B_S) > 0 \text{ et } \forall k \notin S,\ \mathbf{P}(A_k|B_S) \leqslant 2p.$$

Le cas $s = 0$ est trivialement vrai puisque $B_\emptyset = \Omega$. Supposons la propriété vraie au rang $s - 1$ et montrons-la au rang s. Il suffit de le faire pour $S = \{1, \dots, s\}$. On a

$$\mathbf{P}(B_S) = \mathbf{P}(B_1)\mathbf{P}(B_2|B_1)\mathbf{P}(B_3|B_{\{1,2\}}) \dots \mathbf{P}(B_s|B_{\{1,\dots,s-1\}}) \geqslant (1 - 2p)^s > 0$$

par hypothèse de récurrence. Soit maintenant $k \notin S$ et considérons la partition $S = S_1 \cup S_2$ où S_1 est le sous-ensemble formé des sommets reliés à k dans le graphe de dépendance. Si $S_1 = \emptyset$, alors $\mathbf{P}(A_k|B_S) = \mathbf{P}(A_k) \leqslant p \leqslant 2p$. Sinon, on a $|S_2| \leqslant s-1$. On écrit (l'hypothèse de récurrence garantissant que $\mathbf{P}(B_{S_2}) > 0$)

$$\mathbf{P}(A_k|B_S) = \frac{\mathbf{P}(A_k \cap B_S)}{\mathbf{P}(B_S)} = \frac{\mathbf{P}(A_k \cap B_{S_1} \cap B_{S_2})}{\mathbf{P}(B_{S_1} \cap B_{S_2})} = \frac{\mathbf{P}(A_k \cap B_{S_1}|B_{S_2})}{\mathbf{P}(B_{S_1}|B_{S_2})}.$$

On estime séparément le numérateur et le dénominateur.

$$\mathbf{P}(A_k \cap B_{S_1}|B_{S_2}) \leqslant \mathbf{P}(A_k|B_{S_2}) = \mathbf{P}(A_k) \leqslant p$$

$$\mathbf{P}(\overline{B_{S_1}}|B_{S_2}) \leqslant \sum_{i \in S_1} \mathbf{P}(A_i|B_{S_2}) \leqslant 2p|S_1| \leqslant 2pd \leqslant \frac{1}{2}$$

On a donc $\mathbf{P}(B_{S_1}|B_{S_2}) \geqslant \frac{1}{2}$ et donc $\mathbf{P}(A_k|B_s) \leqslant 2p$, concluant la récurrence. □

4.5 Application du lemme local de Lovász : routage de paquets

On considère un graphe non orienté (V, E) et un ensemble de paquets $p_1, \ldots, p_n$. A chaque paquet p_i est associé un itinéraire, formé d'un sommet de départ s_i, d'un sommet d'arrivée t_i et d'un chemin dans le graphe allant de s_i à t_i. A chaque étape de temps discret, un paquet peut attendre ou être déplacé vers l'étape suivante de son itinéraire avec la contrainte qu'une arête ne peut être empruntée à chaque étape que par un seul paquet. Un planning de durée T est la donnée pour chaque paquet et chaque instant $t \in \{1, \ldots, T\}$ d'une instruction «avance !» ou «attends !». Le planning est valide si chaque paquet complète son itinéraire et si chaque arête est utilisée par au plus un paquet à chaque étape. On cherche à minimiser la durée d'un planning valide.

Il y a deux paramètres pertinents : la dilatation

$$d = \max_i \{ \text{ longueur de l'itinéraire du paquet } p_i \}$$

et la congestion

$$c = \max_e \{ \text{ nombre d'itinéraires utilisant l'arête } e \}$$

Il est évident que tout planning valide nécessite une durée $\geqslant \max(c, d)$. Il est clair aussi qu'il existe un planning valide de durée $\leqslant cd$.

Théorème. *Il existe un planning valide de temps $O(\max(c, d))$.*

Posons $m = \max(c, d)$. Nous allons montrer à l'aide du lemme local de Lovász une version plus faible de ce théorème : il existe un planning valide de temps $O(m\beta^{\log^*(m)})$ où β est une constante à déterminer et $\log^*$ est le logarithme itéré (i.e. le nombre d'itérations de la fonction logarithme nécessaires pour obtenir une valeur < 1).

On donne un algorithme récursif qui consiste à utiliser un planning arbitraire si $m < m_0$, et si $m \geqslant m_0$ à diviser l'intervalle $\{1, \ldots, \beta m\}$ en phases de longueur $\log m$. Dans chaque phase, chaque paquet se voit attribuer un sous-itinéraire de son itinéraire initial, de sorte qu'on retrouve l'itinéraire initial d'un paquet en mettant bout à bout les sous-itinéraires. Nous allons voir que pour $m \geqslant m_0$ on peut faire en sorte que la congestion soit $\leqslant \log m$

pour chacune des phases. Il est évident que la dilatation dans chacune des phases est $\leqslant \log m$. Si on note $T(m)$ la durée du planning ainsi construit, on a pour $m \geqslant m_0$

$$T(m) \leqslant T(\log m)\frac{\beta m}{\log m}$$

et donc $T(m) = O(m\beta^{\log^*(m)})$. (La profondeur de la récursion est $\log^*(m)$ et chaque appel récursif multiplie la durée par un facteur β).

On considère un planning du type suivant : chaque paquet p_i reçoit l'ordre d'attendre pendant un temps X_i, puis de faire toutes les étapes de son itinéraire. Posons $\alpha = \beta - 1$. On choisit X_i aléatoirement, indépendamment selon la loi uniforme sur $\{1, \dots, \alpha m\}$. Pour $e \in E$, on note A_e l'événement «il existe une phase où l'arête e apparaît dans $> \log m$ sous-itinéraires». Nous allons montrer que $\mathbf{P}(\bigcap_e \overline{A_e}) > 0$, ce qui permettra de conclure.

Lemme. *On peut choisir la valeur de α de sorte que $\mathbf{P}(A_e) \leqslant \frac{1}{4m^2}$ pour m assez grand.*

Soit $F_e \subset E$ la réunion des itinéraires contenant l'arête e. On a $|F_e| \leqslant cd \leqslant m^2$. L'événement A_e est indépendant de $(A_f)_{f \notin F_e}$ et donc les événements $(A_e)_{e \in E}$ admettent un graphe de dépendance de degré $\leqslant m^2$. Le lemme local de Lovász implique que l'événement $\bigcap_e \overline{A_e}$ est non vide.

Preuve du lemme. Fixons une phase $i \in \{1, \dots, \frac{\beta m}{\log m}\}$ et $N_{e,i}$ le nombre de sous-itinéraires utilisant l'arête e au cours de la phase i. La variable aléatoire $N_{e,i}$ est une somme de v.a. indépendantes de loi de Bernoulli (chacun des $\leqslant d$ paquets ayant l'arête e dans leur itinéraire ont probabilité $\leqslant \frac{\log m}{\alpha m}$ de l'inclure dans le sous-itinéraire de l'étape i)

$$\mathbf{E}[N_{e,i}] \leqslant d \cdot \frac{\log(m)}{\alpha m} = \frac{\log m}{\alpha}$$

On a vu dans la preuve de Chernoff II que si X est une somme de v.a. de Bernoulli indépendantes avec $\mathbf{E}[X] \leqslant \mu$, alors pour tout $\delta > 0$ on a $\mathbf{P}(X \geqslant (1+\delta)\mu) \leqslant \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$. On a donc, avec $\mu = \frac{\log m}{\alpha}$

$$\mathbf{P}(N_{e,i} > \log m) = \mathbf{P}(N_{e,i} > \alpha\mu) \leqslant \left(\frac{e^\alpha}{\alpha^\alpha}\right)^\mu = \left(\frac{e}{\alpha}\right)^{\log m} = \frac{m}{m^{\log \alpha}}.$$

Par la borne de l'union,

$$\mathbf{P}(A_e) = \mathbf{P}(\exists i \ : \ N_{e,i} > \log m) \leqslant \frac{(\alpha+1)m}{\log m}\frac{m}{m^{\log \alpha}} \leqslant \frac{1}{4m^2},$$

la dernière inégalité étant vraie pour m et α suffisamment grands. □

Fin cours #8 du 12 novembre

Chapitre 5

Chaînes de Markov

Les chaînes de Markov sont un exemple de suite de variables aléatoires (X_n) non indépendantes, où la loi de X_{n+1} dépend uniquement de X_n. Pour une chaîne de Markov, «le futur ne dépend du passé qu'à travers le présent».

5.1 Définition

On se donne un ensemble fini ou dénombrable S, appelé l'ensemble des états. On supposera souvent que $S = \{1, \dots, n\}$ ou $S = \mathbf{N}$.

Définition. On dit qu'une suite $(X_n)_{n\geqslant 0}$ de variables aléatoires à valeurs dans S est une chaîne de Markov s'il existe une fonction $Q : S \times S \to [0, 1]$ telle que l'on ait

$$\mathbf{P}(X_n = a_n | X_0 = a_0,\ X_1 = a_1, \dots,\ X_{n-1} = a_{n-1}) = Q(a_{n-1}, a_n)$$

pour tous $a_0, \dots, a_{n-1}$ dans S tels que $\mathbf{P}(X_0 = a_0,\ \dots,\ X_{n-1} = a_{n-1}) > 0$.

On dit que Q est la *matrice de transition* de la chaîne de Markov. Elle est à valeurs positives et vérifie la condition

$$\sum_{b\in S} Q(a, b) = 1$$

pour tout $a \in S$ (une matrice vérifiant ces conditions est dite *stochastique*). La loi de (X_n) est entièrement déterminée par Q et par la donnée de la loi de X_0. On a en effet, pour tous $a_0, a_1, \dots, a_n$ dans S

$$\mathbf{P}(X_0 = a_0, X_1 = a_1, \dots, X_n = a_n) = \mathbf{P}(X_0 = a_0)Q(a_0, a_1)Q(a_1, a_2) \dots Q(a_{n-1}, a_n)$$

Dans le cas particulier important où X_0 est constante égale à $a \in S$ (on parle de «chaîne de Markov issue de a»), on a,

$$\mathbf{P}(X_1 = a_1, \dots, X_n = a_n) = Q(a, a_1)Q(a_1, a_2) \dots Q(a_{n-1}, a_n)$$

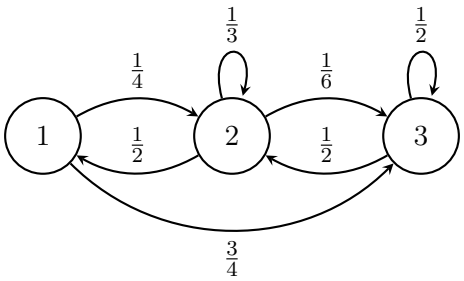
Soit μ_0 la loi de X_0, vue comme un vecteur ligne, de coefficients $\mu_0(a) = \mathbf{P}(X_0 = a)$ pour $a \in S$. Si μ_n est la loi de X_n, alors

$$\mu_1(b) = \mathbf{P}(X_1 = b) = \sum_{a\in S} \mathbf{P}(X_1 = b,\ X_0 = a) = \sum_{a\in S} Q(a, b)\mu_0(a)$$

et on a donc la relation $\mu_1 = \mu_0 Q$ au sens de la multiplication matricielle. Plus généralement, si μ_n est la loi de X_n, on a $\mu_n = \mu_0 Q^n$. La matrice Q^n correspond à la matrice de transition après n pas de la chaîne de Markov.

Il est utile de représenter la matrice de transition sous forme de graphe orienté étiqueté de sommets S, dans lequel (x, y) est une arête si et seulement si $Q(x, y) > 0$. Par exemple, voici une matrice de transition suivante pour l'espace d'états $S = \{1, 2, 3\}$ et sa représentation graphique.

$$\begin{bmatrix} 0 & 1/4 & 3/4 \\ 1/2 & 1/3 & 1/6 \\ 0 & 1/2 & 1/2 \end{bmatrix}$$



5.2 Un algorithme probabiliste pour 2-SAT

Voici un exemple concret qui illustre l'intérêt des chaînes de Markov pour l'étude des algorithmes probabilistes. Une formule 2-SAT est du type

$$(x_1 \vee \overline{x_2}) \wedge (\overline{x_1} \vee x_3) \wedge (x_1 \vee x_5) \wedge (\overline{x_4} \vee \overline{x_1}) \wedge \dots$$

Étant donné une formule 2-SAT en n variables, on voudrait déterminer si elle est satisfiable, c'est-à-dire s'il existe une affectation des variables booléennes $x_1, \dots, x_n$ qui la rendre vraie. Ce problème peut résolu par un algorithme déterministe en temps polynomial, mais on propose l'algorithme probabiliste suivant, en temps polynomial.

1. On initialise avec une affectation arbitraire des variables.
2. Répéter $200n^2$ fois, en s'arrêtant si la formule est satisfaite
   (a) Choisir arbitrairement une clause non satisfaite
   (b) Choisir uniformément au hasard une des deux variables apparaissant dans cette clause, et la remplacer par sa négation.
3. Répondre «la formule est satisfiable» si l'algorithme s'est arrêté au cours de l'étape 2, et «la formule n'est pas satisfiable» sinon.

Théorème. *Cet algorithme a une probabilité d'erreur* $\leqslant 2^{-100}$.

Le seul cas où l'algorithme peut se tromper est si la formule est satisfiable. Traitons donc ce cas. Soit A une affectation des variables satisfaisant la formule. Pour l'analyse de l'algorithme, nous allons étudier une modification où la condition d'arrêt dans la boucle est remplacée par «en s'arrêtant si l'affectation coïncide avec A». Nous allons montrer que la probabilité que l'algorithme modifié ne s'arrête pas est $\leqslant 2^{-100}$. Cela implique que la probabilité que l'algorithme initial ne s'arrête pas est $\leqslant 2^{-100}$.

On note X_i le nombres des variables ayant la même valeur que dans l'affectation A après i tours de boucle dans l'algorithme modifié. L'algothime s'arrête au $i$ème tour de boucle si et seulement si $X_i = n$. Dans ce cas, on pose $X_j = n$ pour tout $j > i$.

Si $X_i < n$, la variable X_{i+1} est égale soit à $X_i + 1$ soit à $X_i - 1$, selon que la variable remplacée par sa négation lors du $i$ème tour de boucle était en désaccord ou non avec S. On a

$$\mathbf{P}(X_{i+1} = 1 | X_i = 0) = 1$$

$$\mathbf{P}(X_{i+1} = k+1 | X_i = k) \geqslant 1/2$$

(puisque toute clause non satisfaite a au moins une variable en désaccord avec S) et donc, par passage à l'événement complémentaire,

$$\mathbf{P}(X_{i+1} = k-1 | X_i = k) \leqslant 1/2$$

Pour que (X_i) soit défini même après l'arrêt de l'algorithme, on rajoute la condition

$$\mathbf{P}(X_{i+1} = n | X_i = n) = 1.$$

La suite (X_i) de variables aléatoires n'est pas une chaîne de Markov ! Néanmoins, on peut la comparer à une chaîne de Markov qui serait une version pessimiste de (X_i). Définissons une chaîne de Markov (Y_i) sur l'espace d'états $\{0, \ldots, n\}$ de matrice de transition

$$Q(0,1) = Q(n,n) = 1$$

$$Q(k, k+1) = Q(k, k-1) = \frac{1}{2} \text{ si } 0 < k < n$$

On peut faire un couplage de (X_i) et (Y_i) (c'est-à-dire les définir sur le même espace de probabilité) de telle sorte que l'on ait $Y_0 = X_0 = k_0$ et pour tout j l'inégalité $Y_j \leqslant X_j$. Si on définit les variables aléatoires

$$T_{k_0}^{(n)} = \min\{j \geqslant 0 \ : \ X_j = n\}, \quad S_{k_0}^{(n)} = \min\{j \geqslant 0 \ : \ Y_j = n\}$$

on a alors presque sûrement l'inégalité $T_{k_0}^{(n)} \leqslant S_{k_0}^{(n)}$.

Lemme. *Pour tout $k_0 \in \{0, \ldots, n\}$, on a $\mathbf{E}[T_{k_0}^{(n)}] \leqslant n^2$.*

Démonstration. Posons $u_k = \mathbf{E}[S_k^{(n)}]$. On va montrer que $u_k \leqslant n^2$ et le lemme en découlera. La suite (u_k) vérifie la relation de récurrence $u_n = 0$, $u_0 = 1 + u_1$ et

$$u_k = 1 + \frac{u_{k+1} + u_{k-1}}{2} \text{ si } 0 < k < n$$

Cette relation de récurrence admet la solution explicite $u_k = n^2 - k^2$, d'où le résultat. □

Par l'inégalité de Markov, on a donc $\mathbf{P}(T_{k_0}^{(n)} > 2n^2) \leqslant \frac{1}{2}$. Divise les $200n^2$ itérations de l'algorithme en 100 blocs de longueur $2n^2$ et pour $1 \leqslant i \leqslant 100$, soit A_i l'événement «l'algorithme s'arrête au cours du $i$ème bloc». Il découle de l'analyse précédente que

$$\mathbf{P}(A_1) \geqslant \frac{1}{2}, \quad \mathbf{P}(A_2 | \overline{A_1}) \geqslant \frac{1}{2}, \quad \ldots \quad \mathbf{P}(A_i | \overline{A_1} \cap \cdots \cap \overline{A_{i-1}}) \geqslant \frac{1}{2}$$

et donc

$$\mathbf{P}(\overline{A_1} \cap \cdots \cap \overline{A_{100}}) \leqslant 2^{-100}$$

5.3 Classification des états

On fixe une chaîne de MARKOV (X_n) de matrice de transition Q et d'ensemble d'états S.

Pour i, j dans S, on dit que j est *accessible depuis* i et on note $i \rightsquigarrow j$ s'il existe un entier $n \geqslant 0$ tel que $Q^n(i,j) > 0$. Cela revient à dire qu'il existe un chemin de i vers j dans le graphe orienté associé. On dit que i et j *communiquent* si $i \rightsquigarrow j$ et $j \rightsquigarrow i$. Remarquons que si $i \rightsquigarrow j$ et $j \rightsquigarrow k$ alors $i \rightsquigarrow k$ (preuve : si $Q^m(i,j) > 0$ et $Q^n(j,k) > 0$ alors $Q^{m+n}(i,k) \geqslant Q^m(i,j)Q^n(j,k) > 0$).

On dit qu'une chaîne de MARKOV est *irréductible* si tous ses états communiquent.

Afin d'alléger les notations, pour $x \in S$, on utilisera les notations $\mathbf{P}_x$ ou $\mathbf{E}_x$ pour signifier que l'on considère la chaîne de MARKOV $(X_n)_{n\geqslant 0}$ issue de x, c'est-à-dire telle que $\mathbf{P}(X_0 = x) = 1$.

Dans l'étude du comportement asymptotique d'une chaîne de MARKOV (X_n), on associe à chaque état $x \in S$ deux variables aléatoires à valeurs entières

— On note N_x le nombre de visites en x, défini comme

$$N_x = \sum_{n\geqslant 0} \mathbf{1}_{\{X_n = x\}}.$$

— On note T_x le *temps d'atteinte* de x, ou encore l'instant de première visite en x, définie comme

$$T_x = \inf\{n > 0 \ : \ X_n = x\}$$

avec la convention habituelle $T_x = \infty$ si l'ensemble est vide.

Les états d'une chaîne de MARKOV se classifient selon la dichotomie suivante

— Un état x est dit *récurrent* si

$$\mathbf{P}_x(T_x < \infty) = 1.$$

— Un état x est dit *transitoire* (ou transient) si

$$\mathbf{P}_x(T_x < \infty) < 1$$

.

Si x est récurrent, la chaîne issue de x revisite presque sûrement x au bout d'un temps fini, puis presque sûrement revisite une seconde fois x au bout d'une temps fini, ... Il s'ensuit que le nombre de visite en x est presque sûrement infini : $\mathbf{P}_x(N_x = \infty) = 1$.

Si x est transitoire, posons $p = \mathbf{P}_x(T_x = \infty) > 0$. Pour la chaîne issue de x, le nombre de visites en x suit alors une loi géométrique de paramètre p et est donc d'espérance finie : on a $\mathbf{P}_x(N_x = \infty) = 0$ et $\mathbf{E}_x[N_x] = \frac{1}{p} < \infty$.

Les raisonnements ci-dessous utilisent de manière intuitive ce que les mathématiciens appellent la *propriété de* MARKOV : conditionnellement à l'événement $\{T_x < \infty\}$, la loi $(X_{T_x+n})_{n\geqslant 0}$ est identique à loi de (X_n) sachant $X_0 = x$. Ces considérations peuvent être rendues rigoureuses mais nous n'introduirons pas le formalisme nécessaire et nous contenterons de raisonnements intuitifs.

On peut montrer que deux états qui communiquent ont même nature (ils sont soit tous deux récurrents, soit tous deux transitoires). On peut aussi montrer qu'une chaîne de MARKOV sur un espace d'états fini admet au moins un état récurrent.

5.4 Probabilités invariantes et convergence des chaînes de Markov

On considère une chaîne de Markov (X_n) d'espace d'états S et de matrice de transition Q.

Définition. Une mesure de probabilité π sur S est dite *invariante* si elle vérifie la relation $\pi = \pi Q$, c'est-à-dire que pour tout $y \in S$

$$\pi(y) = \sum_{x \in S} \pi(x) Q(x, y).$$

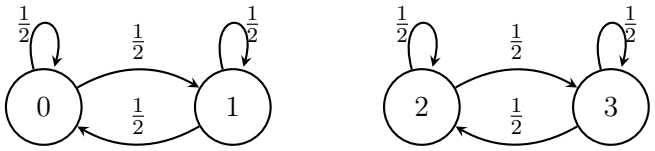
L'interprétation est la suivante : si $X_n \sim \pi$ alors $X_{n+1} \sim \pi$. On s'intéresse au comportement en temps long des chaînes de Markov à travers la quantité

$$\tilde{\pi}(x) = \lim_{n \to \infty} \mathbf{P}(X_n = x).$$

On remarque que si la limite existe et si $\tilde{\pi}$ est une mesure de probabilité, alors $\tilde{\pi}$ est invariante. Dans la situation idéale, il existe une unique probabilité invariante π qui coïncident avec le $\tilde{\pi}$ ci-dessus. Mais il y a plusieurs obstructions.

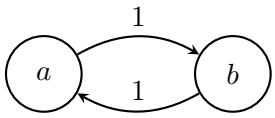
— Une obstruction liée à l'infini : il peut ne pas y avoir de probabilité invariante. C'est le cas par exemple pour la marche aléatoire sur $\mathbf{Z}$. En effet, une probabilité invariante vérifie la relation $\pi(k) = \pi(k+1)/2 + \pi(k-1)/2$. Les fonction $\pi : \mathbf{Z} \to \mathbf{R}$ solutions de cette équation sont de la forme $\pi(n) = \alpha n + \beta$, qui ne sont pas compatibles avec les conditions $\pi \geqslant 0$ et $\pi(\mathbf{Z}) = 1$. On verra le résultat suivant.
Théorème. *Une chaîne de Markov à espace d'états fini admet une probabilité invariante.*

— Une obstruction liée à la non-irréductibilité : il peut y avoir plusieurs probabilités invariantes. C'est le cas par exemple de l'exemple suivant



La probabilité π définie par $\pi(0) = \pi(1) = \frac{1}{2}$ et $\pi(2) = \pi(3) = 0$ est invariante, de même que la probabilité $\overline{\pi}$ définie par $\overline{\pi}(0) = \overline{\pi}(1) = 0$ et $\overline{\pi}(2) = \overline{\pi}(3) = \frac{1}{2}$. Plus généralement, pour tout t dans $[0, 1]$, la probabilité $t\pi + (1-t)\overline{\pi}$ est invariante. On verra le résultat suivant.

— Une obstruction de nature arithmétique : il peut y avoir une unique probabilité invariante sans que la quantité $\mathbf{P}(X_n = x)$ ait une limite quand n tend vers l'infini. C'est le cas car exemple de la chaîne de Markov suivante



pour laquelle l'unique probabilité invariante est la loi uniforme sur $\{a, b\}$. Si par exemple X_0 vaut a, alors la variable aléatoire X_{2n} est constante égale à a ou b selon la parité de n. Cette obstruction liée à la parité est évidente sur cet exemple mais se retrouve dans de nombreuses situations (par exemple considérer la marche aléatoire d'un cavalier sur un échiquier).

Fin cours #9 du 26 novembre

Théorème (Existence et unicité de la probabilité invariante)**.** *Soit (X_n) une chaîne de* MARKOV *(X_n) irréductible à espace d'états finis S. Alors tous les états sont récurrents ; elle admet une unique mesure de probabilité invariante π, donnée pour $x \in S$ par*

$$\pi(x) = \frac{1}{\mathbf{E}_x[T_x]} > 0.$$

Démonstration. Une mesure de probabilité π est invariante si (identifiée à un vecteur ligne) elle vérifie l'équation $\pi Q = \pi$, autrement si c'est un vecteur propre à gauche de valeur propre 1 pour la matrice Q.

Puisque la somme des lignes de Q vaut 1, le vecteur $\mathbf{1} = (1, \dots, 1)$ est vecteur propre à droite. Réciproquement, si $f = (f(x))_{x \in S}$ vérifie $Qf = f$, montrons que f est un multiple de $\mathbf{1}$. Soit $x \in S$ tel que $f(x)$ est maximal et soit $y \neq x$. Par irréductibilité, il existe un entier n tel que $Q^n(x, y) > 0$. On a alors

$$f(x) = Q^n f(x) = \sum_{y \in S} Q^n(x, y) f(y) \leqslant \sum_{y \in S} Q^n(x, y) f(x) = f(x)$$

et donc $f(y) = f(x)$. Ainsi, tout vecteur propre à droite de valeur propre 1 est un multiple de $\mathbf{1}$. Comme les espaces propres à droite et à gauche ont même dimension, l'espace propre à gauche de valeur propre 1 est aussi de dimension 1. On en déduit l'unicité (si existence) d'une mesure de probabilité invariante. Cet argument d'algèbre linéaire donne l'existence d'un vecteur non nul π tel que $\pi Q = \pi$, mais il n'est pas clair que ce vecteur soit à coefficients positifs.

Pour tous x, y dans S, on note $n_{x,y}$ le plus petit entier $n > 0$ tel que $Q^n(x, y) > 0$ (un tel n existe par irréductibilité ; c'est la longueur minimale d'un chemin de x à y). On pose aussi $N = \max\{n_{x,y} \ : \ x, y \in S\}$, puis on choisit $\varepsilon > 0$ tel que, pour tous $x, y \in S$ on ait $Q^n(x, y) \geqslant \varepsilon$ pour un $n \in \{1, \dots, N\}$.

Fixons x, y dans S et soit l'événement

$$A_k = \text{« il existe un entier } n \text{ vérifiant } kN \leqslant n < (k+1)N \text{ et } X_n = y \text{ »}.$$

On a $\mathbf{P}_x(A_0) \geqslant \varepsilon$, $\mathbf{P}_x(A_1 | \overline{A_0}) \geqslant \varepsilon$ et plus généralement $\mathbf{P}_x(A_{k+1} | \overline{A_0} \cap \overline{A_1} \cap \cdots \cap \overline{A_k}) \geqslant \varepsilon$ pour tout k. On en déduit que

$$\mathbf{P}_x(T_y \geqslant kN) \leqslant (1 - \varepsilon)^k$$

ce qui implique que $\mathbf{E}_x[T_y] = \sum_{l \geqslant 0} \mathbf{P}_x(T_y > l) \leqslant N \sum_{k=0}^{\infty} (1 - \varepsilon)^k < \infty$. Fixons un état z et définissons $\mu_z : S \to \mathbf{R}^+$ par la formule

$$\mu_z(y) = \mathbf{E}_z \left[\sum_{k=0}^{T_z - 1} \mathbf{1}_{\{X_k = y\}} \right] = \sum_{k=0}^{\infty} \mathbf{P}_z(X_k = y, \ T_z \geqslant k + 1).$$

Autrement dit, $\mu_z(y)$ est le nombre moyen de visites en y entre deux visites en z. Remarquons que $\mu_z(y) \leqslant \mathbf{E}_z[T_z]$, donc μ_z est à valeurs finies et tous les états sont récurrents. Pour tout $y \in S$, on a

$$(\mu_z Q)(y) = \sum_{x \in S} \mu_z(x) Q(x, y) = \sum_{x \in S} \sum_{k=0}^{\infty} \mathbf{P}_z(X_k = x, \ T_z \geqslant k + 1) Q(x, y)$$

Mais on a, pour tout $k \geqslant 0$, puisque l'événement $\{T_z \geqslant k+1\}$ peut s'exprimer en fonction de $X_0, \dots, X_k$,

$$\sum_{x \in S} \mathbf{P}_z(X_k = x, \ T_z \geqslant k + 1) Q(x, y) = \mathbf{P}_z(X_{k+1} = y, \ T_z \geqslant k + 1)$$

On a donc, en faisant le changement d'indice $k+1 \to k$ dans la seconde somme,

$$\begin{aligned}\mu_z(y) - (\mu_z Q)(y) &= \sum_{k=0}^{\infty} \mathbf{P}_z(X_k = y,\ T_z \geqslant k+1) - \sum_{k=1}^{\infty} \mathbf{P}_z(X_k = y,\ T_z \geqslant k) \\ &= \mathbf{P}_z(X_0 = y) - \sum_{k=1}^{\infty} \mathbf{P}_z(X_k = y,\ T_z = k) \\ &= 0\end{aligned}$$

car les deux termes de cette différence valent tous deux 1 si $y = z$ et tous deux 0 sinon. On a donc $\mu_z Q = \mu_z$. Puisque $\sum_{y \in S} \mu_z(y) = \mathbf{E}_z[T_z]$, on en déduit que l'unique probabilité invariante est donnée par $\pi(x) = \frac{\mu_z(x)}{\mathbf{E}_z[T_z]}$. Puisque par ailleurs $\mu_z(z) = 1$, on en déduit que $\pi(x) = \frac{1}{\mathbf{E}_x[T_x]}$ pour tout $x \in S$. □

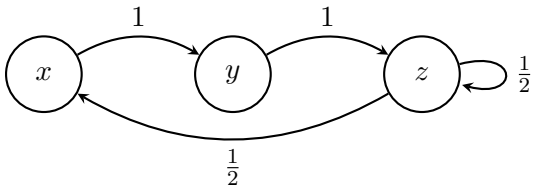
Nous allons maintenant détailler l'obstruction d'ordre arithmétique pour la convergence en grand temps vers la mesure de probabilité invariante. Pour alléger les notations, on écrira désormais Q_{xy} ou Q^k_{xy} plutôt que $Q(x,y)$ ou $Q^k(x,y)$.

Définition. La *période* d'un état x est

$$d_x = \mathrm{PGCD}\{n \geqslant 1\ :\ Q^n_{xx} > 0\}$$

(rappelons que le PGCD d'un ensemble de nombres entiers est leur Plus Grand Commun Diviseur). Une chaîne de Markov est *apériodique* si tout état a période 1.

Dans l'exemple suivant, on a $d_z = 1$ puisque $Q_{zz} > 0$ mais aussi $d_x = 1$ puisque $Q^3_{xx} > 0$ et $Q^4_{xx} > 0$ (de même, $d_y = 1$). Remarquons que comme la notion d'irréductibilité, la notion d'apériodicité de dépend pas de la valeur des étiquettes du graphe.



Lemme. *Dans une chaîne de Markov, deux états qui communiquent ont même période.*

Corollaire. *Dans une chaîne de Markov irréductible, tous les états ont même période.*

Démonstration. Supposons $x \rightsquigarrow y \rightsquigarrow x$. Il existe donc deux entiers m et n tels que $Q^m_{xy} > 0$ et $Q^n_{yx} > 0$. Soient d_x et d_y les périodes de x et y.

— Puisque $Q^{m+n}_{xx} \geqslant Q^m_{xy} Q^n_{yx} > 0$, l'entier d_x divise $m+n$.

— Soit p un entier tel que $Q^p_{yy} > 0$. On a $Q^{m+p+n}_{xx} \geqslant Q^m_{xy} Q^p_{yy} Q^n_{yx} > 0$, ce qui fait que l'entier d_x divise $m+p+n$, et donc également p d'après le point précédent.

Puisque $d_x | p$ pour tout p tel que $Q^p_{yy} > 0$, on déduit de la définition du PGCD que $d_x | d_y$. Par symétrie, on a donc $d_x = d_y$. □

Théorème (Théorème de convergence)**.** *Soit (X_n) une chaîne de Markov irréductible apériodique à espaces d'états finis S, et soit π sa mesure de probabilité invariante. Alors, pour tous x, y dans S*

$$\lim_{n \to \infty} \mathbf{P}_x(X_n = y) = \pi(y).$$

Lemme. *Il existe un entier N tel que l'on ait $Q^n_{xy} > 0$ pour tout $n \geqslant N$ et $x, y \in S$.*

Démonstration. L'espace d'états étant fini, il suffit de voir que pour tout $x, y \in S$, il existe un entier $N_{x,y}$ tel que $Q^n_{xy} > 0$ pour tout $n \geqslant N_{x,y}$. Mais ceci est une conséquence directe du résultat suivant, qui est un exercice d'arithmétique : étant donnés des entiers naturels $n_1, \ldots, n_k$ tels que $\mathrm{PGCD}(n_1, \ldots, n_k) = 1$, il existe un entier N tel que tout entier $n \geqslant N$ s'écrit comme $n = a_1 n_1 + \cdots + a_k n_k$ avec $a_1, \ldots, a_n$ dans $\mathbf{N}$. □

Preuve du théorème de convergence. Il faut montrer que pour tous x, y dans S, on a

$$\lim_{n\to\infty} Q^n_{xy} = \pi(y).$$

On utilise un argument de couplage en définissant une nouvelle chaîne de Markov d'espace d'états $S \times S$ et de matrice de transition $\overline{Q}$ donnée par

$$\overline{Q}_{(y,z),(y',z')} = Q_{y,y'} Q_{z,z'}.$$

Si (Y_n, Z_n) est une chaîne de Markov de matrice de transition $\overline{Q}$ issue de (y, z), alors (Y_n) et (Z_n) sont deux chaînes de Markov de matrice de transition Q issues respectivement de y et z ; de plus (Y_n) et (Z_n) sont indépendantes.

La chaîne de Markov de matrice de transition $\overline{Q}$ est irréductible. C'est ici qu'on utilise l'apériodicité de Q : si N est donné par le lemme, alors pour $n \geqslant N$

$$\overline{Q}^n_{(y,z),(y',z')} = Q^n_{y,y'} Q^n_{z,z'} > 0.$$

En tant que chaîne de Markov irréductible à espaces d'états fini, cette chaîne est récurrente. La mesure de probabilité $\overline{\pi}$ définie sur $S \times S$ par $\overline{\pi}(y, z) = \pi(y)\pi(z)$ (c'est-à-dire, c'est la loi d'un couple de deux variables aléatoires indépendantes de loi π) est invariante pour $\overline{Q}$.

Prouvons maintenant le théorème. Fixons $(w, x) \in S \times S$ et soit (Y_n, Z_n) la chaîne de Markov de matrice de transition $\overline{Q}$ issue de (w, x). Pour alléger les notations, on notera $\mathbf{P}_* = \mathbf{P}_{(w,x)}$). Fixons un état arbitraire $z \in S$ et considérons

$$\tau = \inf\{n \geqslant 0 \ : \ (Y_n, Z_n) = (z, z)\}.$$

Puisque la chaîne est irréductible et récurrente, on a $\mathbf{P}_*(\tau < \infty) = 1$. Pour tout y dans S, on a

$$\begin{aligned}
\left|Q^n_{wy} - Q^n_{xy}\right| &= |\mathbf{P}_*(Y_n = y) - \mathbf{P}_*(Z_n = y)| \\
&= \Big| \sum_{k=0}^{n} \mathbf{P}_*(Y_n = y,\ \tau = k) + \mathbf{P}_*(Y_n = y,\ \tau > n) \\
&\qquad - \sum_{k=0}^{n} \mathbf{P}_*(Z_n = y,\ \tau = k) - \mathbf{P}_*(Z_n = y,\ \tau > n) \Big|
\end{aligned}$$

Par symétrie (i.e., $\overline{\pi}(s, t) = \overline{\pi}(t, s)$ pour tous s et t dans S), on a pour tout $k \leqslant n$ l'égalité

$$\mathbf{P}_*(Y_n = y,\ \tau = k) = \mathbf{P}_*(Z_n = y,\ \tau = k).$$

On en conclut que

$$\left|Q^n_{wy} - Q^n_{xy}\right| \leqslant |\mathbf{P}_*(Y_n = y,\ \tau > n) - \mathbf{P}_*(Z_n = y,\ \tau > n)| \leqslant \mathbf{P}(\tau > n)$$

et donc $\lim_{n\to\infty} |Q^n_{xy} - Q^n_{x'y}| = 0$. Finalement, on utilise l'égalité $\pi Q^n = \pi$ pour écrire

$$\pi(y) - Q^n_{xy} = \left(\sum_{w\in S} \pi(w) Q^n_{wy}\right) - Q^n_{xy} = \sum_{w\in S} \pi(w)\left(Q^n_{wy} - Q^n_{xy}\right)$$

et cette quantité tend vers 0. □

Fin cours #10 du 28 novembre

5.5 Calcul de la probabilité invariante

On se donne une chaîne de MARKOV irréductible à espace d'états finis, qui admet donc une unique probabilité invariante. Le calcul de la mesure invariante peut toujours se ramener à une résolution de système linéaire. Par exemple, pour la chaîne de matrice de transition

$$Q = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/3 & 1/6 & 1/2 \end{pmatrix}$$

la mesure invariante π vérifie l'équation $\pi Q = \pi$, donc est solution du système

$$\pi(1) = \frac{1}{2}\pi(1) + \frac{1}{4}\pi(2) + \frac{1}{3}\pi(3)$$

$$\pi(2) = 0\pi(1) + \frac{1}{2}\pi(2) + \frac{1}{6}\pi(3)$$

$$\pi(3) = \frac{1}{2}\pi(1) + \frac{1}{4}\pi(2) + \frac{1}{2}\pi(3)$$

Ces trois équations ne sont pas linéairement indépendantes (leur somme donne $1 = 1$, puisque la somme des lignes de Q vaut 1) mais il faut rajouter la condition

$$\pi(1) + \pi(2) + \pi(3).$$

La résolution de tels systèmes est vite fastidieuse, et donc on préfère l'éviter si possible. Une alternative possible est celle de la méthode des coupes, qui se base sur l'équation suivante : étant donné une partition $S = S_1 \cup S_2$ de l'espace des états, on a l'équation

$$\sum_{x\in S_1}\sum_{y\in S_2} \pi(x)Q(x,y) = \sum_{x\in S_1}\sum_{y\in S_2} \pi(y)Q(y,x).$$

Cette équation s'interprète ainsi : puisque la probabilité invariante correspond à un état d'équilibre, le flux sortant de S_1 vers S_2 (terme de gauche) est égal au flux entrant de S_2 vers S_1 (terme de droite).

Démontrons l'égalité : on calcule

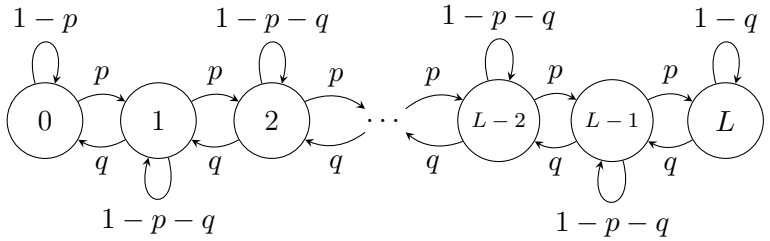
$$\sum_{x\in S_1, y\in S_2} \pi(x)Q_{xy} = \sum_{x\in S_1, y\in S} \pi(x)Q_{xy} - \sum_{x\in S_1, y\in S_1} \pi(x)Q_{xy} = \sum_{x\in S_1} \pi(x) - \sum_{x,y\in S_1} \pi(x)Q_{xy}$$

$$\sum_{x\in S_1, y\in S_2} \pi(y)Q_{yx} = \sum_{x\in S_1, y\in S} \pi(y)Q_{yx} - \sum_{x\in S_1, y\in S_1} \pi(y)Q_{yx} = \sum_{x\in S_1} \pi(x) - \sum_{x,y\in S_1} \pi(x)Q_{xy}$$

où on a utilisé la relation $\pi Q = Q$ dans la dernière égalité.

Exemple. Considérons un modèle de file d'attente de longueur maximale L. Étant donné deux paramètres $p, q > 0$ tels que $p+q \leqslant 1$, le système évolue selon la dynamique suivante. A chaque instant $n \in \mathbf{N}$,

1. si la file d'attente est de longueur $< L$, un nouveau client d'y installe avec probabilité p,
2. si la file d'attente est de longueur > 0, un client est servi (et quitte la file) avec probabilité q.



C'est une chaîne de Markov d'espace d'états $S = \{0, \dots, L\}$, donc la matrice de transition est donnée par $Q(i, i+1) = p$ (si $0 \leqslant i < L$), $Q(i-i-1) = q$ (si $0 < i \leqslant L$), $Q(0,0) = 1-p$, $Q(L, L) = 1 - q$ et $Q(i, i) = 1 - p - q$ (si $0 < i < L$), les autres termes étant nuls. Cette chaîne est irréductible et apériodique et admet donc une unique probabilité invariante π. Pour calculer cette dernière, on considère pour chaque sommet $0 \leqslant i < L$ la partition $S = \{0, \dots, i\} \cup \{i+1, \dots, L\}$. L'équation de coupe s'écrit alors simplement

$$\pi(i)Q_{i,i+1} = \pi(i+1)Q_{i+1,i}$$

et donc $p\pi(i) = q\pi(i+1)$. En posant $\alpha = p/q$, on a donc $\pi(i+1) = \alpha\pi(i)$ puis $\pi(i) = \alpha^i\pi(0)$. Si $\alpha = 1$, la mesure invariante est la mesure uniforme sur S ; sinon on a

$$\pi(i) = \frac{\alpha^i}{\sum_{k=1}^{L} \alpha^k} = \frac{\alpha^i(1-\alpha)}{1-\alpha^{L+1}}.$$

Remarquons que si $\alpha < 1$, on obtient la loi géométrique de paramètre α dans la limite $L \to \infty$.

5.6 La marche aléatoire sur un graphe

On se donne un graphe fini $G = (V, E)$ non orienté, sans boucle ni arête multiple. On suppose de plus qu'aucun sommet n'est isolé. On appelle marche aléatoire sur G la chaîne de Markov d'espace d'états V et de matrice de transition donnée par

$$Q(x, y) = \begin{cases} \frac{1}{\deg x} & \text{si } x \sim y \\ 0 & \text{sinon.} \end{cases}$$

Remarquons que

— la marche aléatoire sur G est irréductible si et seulement si G est connexe,

— la marche aléatoire sur G est apériodique si et seulement si G est non bipartite.

Pour justifier ce dernier point, remarquons que la période de tout sommet vaut 1 ou 2, et qu'elle vaut 1 si et seulement si ce sommet est contenue dans un cycle de longueur impaire ; or les graphes bipartites sont les graphes sans cycle de longueur impaire.

Dans la suite on suppose que le graphe G est connexe et fini. La marche aléatoire sur G admet donc une unique probabilité invariante.

Proposition. *La probabilité invariante pour la marche aléatoire sur G est donnée par*

$$\pi(v) = \frac{\deg(v)}{2|E|}.$$

Démonstration. C'est bien une probabilité puisque la sommet de tous les degrés vaut $2|E|$. On calcule, pour $y \in V$

$$\sum_{x \in V} \pi(x) Q(x,y) = \sum_{x \sim y} \frac{\deg x}{2|E|} \frac{1}{\deg x} = \frac{\deg y}{2|E|} = \pi(y),$$

ce qui montre que π est invariante. □

Corollaire. *Pour tout sommet v, on a $\mathbf{E}_v[T_v] = \frac{2|E|}{\deg x}$.*

Exercice. Une pièce d'échecs (fou, tour, cavalier, reine ou roi) se déplace aléatoirement sur un échiquier. Quels choix de pièce et de case de départ maximisent/minimisent le nombre moyen de déplacements nécessaires pour revenir sur la case de départ ?

On s'intéresse maintenant au temps de recouvrement d'un graphe, qui est défini par

$$T_{rec}(G) = \max_{x \in V} \mathbf{E}_x \left[\max_{y \in V} T_y \right]$$

C'est le temps moyen nécessaire, partant du pire point, pour que la marche aléatoire soit passée par tous les sommets du graphe.

Proposition. *Pour un graphe connexe G, on a $T_{rec}(G) \leqslant 4|V| \cdot |E|$.*

On commence par montrer un lemme

Lemme. *Si $x \sim y$ alors $\mathbf{E}_x[T_y] \leqslant 2|E|$.*

Démonstration. Soit A l'ensemble des voisins de y. On a

$$\frac{2|E|}{\deg y} = \pi(y)^{-1} = \mathbf{E}_y[T_y] = 1 + \sum_x Q_{xy} \, \mathbf{E}_x[T_y] = 1 + \sum_{x \sim y} \frac{1}{\deg y} \, \mathbf{E}_x[T_y]$$

On en tire l'inégalité $\sum_{x \sim y} \mathbf{E}_x[T_y] \leqslant 2|E|$ et le lemme en découle puisque les quantités sommées sont positives. □

Preuve de la proposition. Soit $n = |V|$ et fixons $x \in V$. On se donne un arbre couvrant de G (qui a donc $n-1$ arêtes, énuméré selon l'ordre de parcours comme)

$$x = x_0 \sim x_1 \sim x_2 \sim \cdots \sim x_{2n-3} \sim x_{2n-2} = x.$$

Soit τ le premier instant où les sommets de l'arbre couvrant ont été visités dans cet ordre, c'est à dire

$$\tau = \inf\{N \ : \ \exists 0 \leqslant t_0 < t_1 < \ldots t_{2n-2} \leqslant N \ : \ X_{t_0} = x_0, X_{t_1} = x_1, \ldots, X_{t_{2n-2}} = x_{2n-2}\}$$

On a $\max_{y \in V} T_y \leqslant \tau$. En utilisant la propriété de Markov, on a

$$\mathbf{E}_x[\tau] \leqslant \mathbf{E}_x[T_{x_1}] + \mathbf{E}_{x_1}[T_{x_2}] + \cdots + \mathbf{E}_{x_{2n-3}}[T_{x_{2n-2}}].$$

Par le lemme, chacun de termes de la somme est majoré par $2|E|$. On a donc $T_{rec}(G) \leqslant (2n-2)|E| \leqslant 4|V||E|$ □

Il est découle de la proposition que si G est un graphe connexe à n sommets, alors $T_{rec}(G) \leqslant 2n^3$. Donnons quelques exemples de temps de recouvrement.

1. Le calcul du temps de recouvrement du graphe complet se ramène au problème du collectionneur de vignettes : on a $T_{rec}(K_n) \sim n \log n$.
2. Si $G = L_n$ est le graphe linéaire à n sommets (où les seules arêtes sont $\{i, i+1\}$, on a déjà considéré ce problème dans l'étude du problème 2-SAT ; on peut montrer que $T_{rec}(G) = \Theta(n^2)$.
3. On peut combiner les exemples précédent pour former le graphe «sucette» obtenu en recollant K_n et L_n. Le temps de recouvrement est alors $\Theta(n^3)$, ce qui montre que la proposition précédente ne peut pas être améliorée.

Comme application de la notion de temps de recouvrement, on donne un algorithme probabiliste de mauvaise complexité mais extrêmement économe en mémoire pour le problème suivant. On se donne un graphe G à n sommets, de degré borné. Étant donnés deux sommets x et y, il faut décider s'ils sont reliés dans le graphe. Ce problème a une solution déterministe simple de complexité $O(n)$ et de mémoire $O(n)$ qui consiste à effectuer en parcours du graphe en profondeur (par exemple).

Pour réduire la mémoire utilisée, on suppose qu'on a accès à un oracle qui, interrogé sur un sommet du graphe, renvoie la liste de ses voisins. On peut alors considérer l'algorithme probabiliste suivant : on effectue la marche aléatoire sur G issue de x pendant $4n^3$ étapes. Si la marche passe par y, on répond que x et y sont reliés dans G (et on ne se trompe pas). Sinon, on répond que x et y ne sont pas reliés. La probabilité d'erreur est alors majorée par l'inégalité de Markov

$$\mathbf{P}_x(T_y > 4n^3) \leqslant \frac{T_{rec}(G)}{4n^3} \leqslant \frac{1}{2}$$

et peut être rendue arbitrairement petite en répétant l'algorithme. La complexité est $O(n^3)$ et l'algorithme nécessite une mémoire $O(\log n)$. En effet il suffit de stocker uniquement le sommet actuellement visité par la marche aléatoire, et cette information peut être encodée sur $\log n$ bits.

5.7 Vitesse de convergence vers la probabilité invariante

Soit S un ensemble fini. On définit la *distance en variation totale* entre deux probabilités μ_1 et μ_2 sur S par

$$d_{TV}(\mu_1, \mu_2) = \frac{1}{2} \sum_{x \in S} |\mu_1(x) - \mu_2(x)|.$$

On a aussi, via la formule $\min(a, b) = \frac{a+b-|a-b|}{2}$,

$$d_{TV}(\mu_1, \mu_2) = 1 - \sum_{x \in S} \min(\mu_1(x), \mu_2(x)).$$

Remarquons que $d_{TV}(\mu_1, \mu_2) = 1$ si et seulement si μ_1 et μ_2 sont à supports disjoints.

Cette quantité s'interprète en termes de couplages de μ_1 et μ_2.

Proposition. *Étant données deux probabilités μ_1 et μ_2 sur un ensemble fini S, on a*

$$d_{TV}(\mu_1, \mu_2) = \inf \mathbf{P}(X \neq Y)$$

où la borne inférieure porte sur l'ensemble des couples (X, Y) de variables aléatoires, tels que $X \sim \mu_1$ et $Y \sim \mu_2$.

Démonstration. Pour tout x dans S, on a

$$\mathbf{P}(X = Y = x) \leqslant \min(\mu_1(x), \mu_2(x))$$

et donc en sommant sur x

$$\mathbf{P}(X = y) \leqslant \sum_{x \in S} \min(\mu_1(x), \mu_2(x)) = 1 - d_{TV}(\mu_1, \mu_2)$$

On a donc l'inégalité $d_{TV}(\mu_1, \mu_2) \leqslant \mathbf{P}(X \neq Y)$. Pour montrer l'inégalité, il suffit de donner une loi pour (X, Y) telle que $\mathbf{P}(X = Y = x) = \min(\mu_1(x), \mu_2(x)) =: m(x)$ pour tout $x \in S$. Le choix

$$\mathbf{P}(X = x, Y = y) = \begin{cases} m(x) & \text{si } x = y \\ \frac{(\mu_1(x) - m(x))(\mu_2(y) - m(y))}{1 - \sum_z m(z)} & \text{si } x \neq y \end{cases}$$

répond aux conditions voulues. □

Considérons une chaîne de Markov irréductible apériodique (X_n) à espace d'états finis, de matrice de transition Q et de probabilité invariante π. Pour étudier la convergence vers l'équilibre, on introduit

$$\Delta(n) = \max_{x \in S} d_{TV}(P_x^n, \pi)$$

où $P_x^n = (Q^n(x, y))_y$ est la loi de X_n sachant $X_0 = x$. On définit aussi le *temps de mélange* d'ordre $\varepsilon > 0$ par

$$t_{mix}(\varepsilon) = \inf\{n \ : \ \Delta(n) \leqslant \varepsilon\}$$

Il découle du théorème de convergence que $\lim_{n \to \infty} \Delta(n) = 0$. Vérifions d'abord que la convergence vers l'équilibre est monotone.

Proposition. *La suite $(\Delta(n))_n$ est décroissante.*

Démonstration. On a

$$\begin{aligned} d_{TV}(P_x^{n+1}, \pi) &= \frac{1}{2} \sum_{z \in S} |Q^{n+1}(x, z) - \pi(z)| \\ &= \frac{1}{2} \sum_{z \in S} \left| \sum_{y \in S} Q(x, y) Q^n(y, z) - \sum_{y \in S} Q(x, y) \pi(z) \right| \\ &\leqslant \sum_{y \in S} Q(x, y) \cdot \frac{1}{2} \sum_{z \in S} |Q^n(y, z) - \pi(z)| \\ &\leqslant \sum_{y \in S} Q(x, y) \Delta(n) = \Delta(n) \end{aligned}$$

et il suffit de prendre la borne supérieure sur x. □

Fin cours #11 du 3 décembre

Un couplage pour une chaîne de Markov (X_n) de matrice de transition Q (sur un espace d'états S) es la donnée d'une chaîne de Markov (Y_n, Z_n) d'espace d'états $S \times S$ dont la matrice de transition R vérifie

$$\forall y, y', z \qquad \sum_{z' \in S} R((y, z), (y', z')) = Q(y, y')$$

$$\forall y, z, z' \qquad \sum_{y' \in S} R((y,z),(y',z')) = Q(z,z')$$

$$R((x,x),(y',z')) = \begin{cases} Q(x,y') & \text{si } y' = z' \\ 0 & \text{sinon} \end{cases}$$

Les deux premières conditions peuvent se réécrire comme

$$\mathbf{P}(Y_{n+1} = y' \mid (Y_n, Z_n) = (y,z)) = Q(y,y')$$

$$\mathbf{P}(Z_{n+1} = z' \mid (Y_n, Z_n) = (y,z)) = Q(z,z').$$

La dernière condition revient à demander qu'une fois que les deux coordonnées de la chaîne de Markov sont égales, elles le demeurent pour tout le futur.

Un exemple de couplage utilisé dans la preuve du théorème de convergence est de faire évoluer les deux coordonnées indépendamment jusqu'à ce qu'elles se rencontrent. Cela correspond à demander que lorsque $y \neq z$

$$R((y,z),(y',z')) = Q(y,y')Q(z,z').$$

Lemme (Lemme de couplage)**.** *Si (Y_n, Z_n) est un couplage pour la chaîne de Markov (X_n) tel que*

$$\forall y, z, \in S \qquad \mathbf{P}(Y_N \neq Z_N | Y_0 = y, Z_0 = z) \leqslant \varepsilon,$$

alors $\Delta(N) \leqslant \varepsilon$.

Démonstration. Considérons la chaîne avec $Y_0 = y$ et $Z_0 \sim \pi$. Alors

$$\mathbf{P}(Y_N \neq Z_N) = \sum_z \pi(z) \mathbf{P}(Y_N \neq Z_N | Y_0 = y, Z_0 = z) \leqslant \varepsilon.$$

Comme $Y_N \sim P_y^N$ et $Z_N \sim \pi$, on a $d_{TV}(P_y^N, \pi) \leqslant \varepsilon$, d'où le résultat. □

Nous allons montrer que la convergence vers l'équilibre est toujours exponentiellement rapide.

Proposition. *Il existe des constantes $C > 0$ et $\alpha < 1$ telles que $\Delta(n) \leqslant C\alpha^n$.*

Démonstration. Supposons d'abord que tous les coefficients de la matrice Q soient > 0. Soit μ_x la loi de X_1 sachant $X_0 = x$. C'est la ligne x de la matrice Q. On a $d_{TV}(\mu_y, \mu_z) < 1$; posons

$$\alpha = \max_{y,z} d_{TV}(\mu_y, \mu_z) < 1$$

Par un lemme précédent, pour tous y, z dans S, il existe une loi $\mu_{y,z}$ sur $S \times S$ telle que si $(Y, Z) \sim \mu_{y,z}$,

$$\mathbf{P}(Y \neq Z) = d_{TV}(\mu_y, \mu_z) \leqslant \alpha$$

On définit un couplage (Y_n, Z_n) par

$$\mathbf{P}((Y_{n+1}, Z_{n+1}) = (y',z') \,|\, (Y_n, Z_n) = (y,z)) = \mu_{y,z}(y',z')$$

de telle sorte que, quels que soient (y, z)

$$\mathbf{P}(Y_{n+1} \neq Z_{n=1} \,|\, (Y_n, Z_n) = (x,y)) \leqslant \alpha$$

et donc

$$\mathbf{P}(Y_{n+1} \neq Z_{n+1} \,|\, Y_n \neq Z_n) \leqslant \alpha$$

Puisque $Y_k = Z_k$ implique $Y_{k+1} = Z_{k+1}$, on a alors

$$\mathbf{P}(Y_n \neq Z_n | Y_0 = y, Z_0 = z) = \mathbf{P}(Y_n \neq Z_n, Y_{n-1} \neq Z_{n-1}, \dots, Y_1 \neq Z_1 | Y_0 = y, Z_0 = z)$$

$$= \mathbf{P}(X_n \neq Y_n | X_{n-1} \neq Y_{n-1}) \cdots \mathbf{P}(X_1 \neq Y_1 | X_0 = x, Y_0 = x) \leqslant \alpha^n$$

et donc, par le lemme de couplage, on a $\Delta(n) \leqslant \alpha^n$.

Pour le cas général, soit p un entier telle que tous les coefficients de Q^p soient > 0 (un tel entier existe par apériodicité, cf. preuve du théorème de convergence). Le raisonnement précédent appliqué à Q^p montre que

$$\Delta(pn) \leqslant \alpha^n$$

et la décroissance de Δ permet de conclure que

$$\Delta(n) \leqslant \Delta(p\lfloor n/p \rfloor) \leqslant \alpha^{\lfloor n/p \rfloor} \leqslant C\beta^n$$

pour $\beta = \alpha^{1/p}$ et $C = 1/\alpha$. □

Montrons enfin que la valeur de ε n'a pas grande importance lorsqu'on définit le temps de mélange

Proposition. *Pour tout $0 < \varepsilon < 1/4$, on a*

$$t_{mix}(\varepsilon) \leqslant \lceil \log_2 1/\varepsilon \rceil t_{mix}(1/4).$$

Démonstration. Soit $N = t_{mix}(1/4)$. On a alors, pour tout x, y dans S

$$d_{TV}(P_x^N, P_y^N) \leqslant d_{TV}(P_x^N, \pi) + d_{TV}(\pi, P_y^N) \leqslant 1/2.$$

On peut donc appliquer l'argument de la proposition précédente à la matrice Q^N pour obtenir

$$\Delta(kN) \leqslant (1/2)^k$$

d'où le résultat. □

5.7.1 Exemple : mélange de cartes

On considère un paquet de N cartes que l'on mélange en itérant l'opération suivante : on choisit uniformément au hasard une des cartes du paquet, et on la place au-dessus du paquet. Au bout de combien d'étapes est-ce que le jeu est mélangé ?

On peut voir ce processus comme une chaîne de Markov d'espaces d'états $S = \mathfrak{S}_N$ et de matrice de transition

$$Q(\sigma, \tau) = \frac{1}{N}$$

si $\sigma = (x_1, x_2, \dots, x_N)$ et $\tau = (x_i, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$ pour $i \in \{1, \dots, N\}$. La chaîne de Markov est irréductible et apériodique et la mesure invariante est donnée par la mesure uniforme sur $\mathfrak{S}_N$.

Définir un couplage revient à faire la chose suivante : on a deux paquets de N cartes et on fait une opération qui revient pour chaque paquet à faire un pas de la chaîne de Markov ci-dessous.

Une idée naïve est de tirer au sort un entier i, et mettre sur le dessus la $i$ème carte du premier paquet ainsi que la $i$ème carte du second paquet. Ce couplage n'est pas intéressant

pour étudier la convergence : si on note $(Y_n^{(N)}, Z_n^{(N)})$ la chaîne de MARKOV correspondant, alors on a $\mathbf{P}(Y_{n+1}^{(N)} = Z_{n+1}^{(N)}) = \mathbf{P}(Y_n^{(N)} = Z_n^{(N)})$.

Une meilleure idée est de tirer au sort un entier i, de mettre sur le dessus la $i$ème carte du premier paquet, et de mettre sur le dessus du second paquet la carte de même valeur que celle-ci. Notons $(Y_n^{(N)}, Z_n^{(N)})$ cette chaîne de MARKOV. C'est bien un couplage (chaque carte du second paquet a probabilité $1/N$ d'être choisie). On peur remarquer que les cartes qui ont été manipulées resteront toujours dans la même position dans les deux paquets. Si on pose

$$T^{(N)} = \inf\{n \ : \ \text{toutes les cartes du paquet ont été vues entre les temps 1 et } n\}$$

Alors $n \geqslant T^{(N)}$ implique $Y_n^{(N)} = Z_n^{(N)}$ et donc

$$\mathbf{P}(Y_n^{(N)} \neq Z_n^{(N)}) \leqslant \mathbf{P}(T^{(N)} \geqslant n)$$

L'étude de $T^{(N)}$ se ramène au problème du collectionneur de vignettes, on a vu que

$$\lim_{N\to\infty} \mathbf{P}(T^{(N)} \geqslant (1+\alpha)N\log N) = 0$$

pour tout $\alpha > 0$ et donc $t_{mix}^{(N)}(\varepsilon) \leqslant (1+o(1))N\log N$ pour tout $\varepsilon > 0$.

5.7.2 Exemple : marche aléatoire sur l'hypercube

Soit (X_n) la marche aléatoire sur le graphe de l'hypercube $G_N = (V_N, E_N)$ où $V_N = \{0,1\}^N$ et où deux sommets sont reliés si et seulement si ils ne diffèrent que d'une coordonnée. Comme le graphe de l'hypercube est bipartite, cette chaîne de MARKOV n'est pas apériodique. On peut définir une variante, la marche aléatoire paresseuse, qui se déplace avec probabilité $1/2$ selon la marche aléatoire et ne se déplace pas avec probabilité $1/2$. C'est la chaîne de MARKOV donnée par la matrice de transition

$$Q(x,y) = \begin{cases} \frac{1}{2} & \text{si } x = y \\ \frac{1}{2N} & \text{si } x \sim y \\ 0 & \text{sinon} \end{cases}$$

Elle est irréductible et apériodique. La probabilité invariante est la probabilité uniforme sur V_N.

Pour estimer le temps de mélange de cette chaîne de MARKOV, on définit un couplage (Y_n, Z_n) de la façon suivante. A chaque étape de temps, on choisit $i_n \in \{1,\dots,N\}$ et $\varepsilon_n = \{0,1\}$ indépendemment et uniformément, et on définit Y_{n+1} (resp. Z_{n+1}) en effaçant la coordonnée i_n de Y_n (resp. de Z_n) et en la remplaçant par ε_n. C'est bien un couplage. Comme précédemment, si on note

$$T = \inf\{n \ : \ \text{card}\{i_1,\dots,i_n\} = N\}$$

alors l'événement $\{T > n\}$ est inclus dans l'événement $\{Y_n = Z_n\}$. On se ramène à nouveau au problème de collectionneur de vignettes et donc $t_{\text{mix}}(\varepsilon) \leqslant (1+o(1))N\log N$.

Chapitre 6

Statistiques et compléments sur les gaussiennes

6.1 Estimation de paramètres

En statistiques, on suppose qu'observe souvent des variables aléatoires i.i.d. (X_n) dont on ne connaît pas la loi, mais qu'on aimerait essayer d'identifier.

On se donne en général une famille paramétrée de mesures de probabilités (soit discrètes, soit continues) $(\mu_\theta)_{\theta\in\Theta}$. Par exemple :

1. Les variables X_n sont à valeurs dans $\{0,1\}$. Dans ce cas, la famille paramétrée est la famille des lois de Bernoulli $(\mathsf{B}(\theta))_{\theta\in[0,1]}$.
2. On fait l'hypothèse que les variables X_n suivent une loi géométrique. Dans ce cas, la famille paramétrée est la famille $(\mathsf{G}(\theta))_{\theta\in]0,1]}$.
3. On fait l'hypothèse que les variables X_n suivent une loi gaussienne. Dans ce cas, la famille paramétrée est la famille $\mathsf{N}(m,\sigma^2)_{m\in\mathbf{R},\sigma\geqslant 0}$

On appelle *échantillons* une suite de variables aléatoires (X_n) i.i.d. de loi inconnue parmi une famille $(\mu_\theta)_{\theta\in\Theta}$, discrètes ou continues ; on notera $\mathbf{P}_\theta$ la mesure de probabilité correspondant au cas où la loi est μ_θ. Dans le cas continu, on note f_θ la densité de la loi μ_θ. Un problème fondamental est le problème d'estimation de paramètres : on souhaite définir une pour tout n une fonction $F_n : \mathbf{R}^n \to \Theta$ de sorte que la variable aléatoire

$$\hat{\theta} = F_n(X_1,\dots,X_n)$$

soit aussi proche de θ que possible. On dit que la variable aléatoire $\hat{\theta}$ est un *estimateur*.

Un principe général pour définir des estimateurs est le *maximum de vraisemblance*. La vraisemblance (en anglais : likelihood) d'un paramètre θ connaissant les échantillons est dans le cas discret

$$\begin{aligned} L(\theta|x_1,\dots,x_n) &= \mathbf{P}_\theta(X_1 = x_1,\dots,X_n = x_n) \\ &= \prod_{i=1}^n \mu_\theta(x_i) \end{aligned}$$

et dans le cas continu

$$L(\theta|x_1,\dots,x_n) = \prod_{i=1}^n f_\theta(x_i).$$

On définit l'estimateur par maximum de vraisemblance comme

$$\hat{\theta}_n(X_1,\dots,X_n) = \operatorname{argmax} L(\theta|X_1,\dots,X_n)$$

Il est souvent plus simple de maximiser $\log L$, ce qui est bien sûr équivalent.

Voici une justification informelle du principe du maximum de vraisemblance. Le principe d'agnosticisme consiste à dire tous les choix de paramètres jouent le même rôle ; par exemple si $\Theta = [0,1]$ on peut supposer que le paramètre θ est choisi a priori selon la loi uniforme. On a («formule de Bayes»)

$$\mathbf{P}(\theta|X) = \frac{\mathbf{P}(X|\theta)\mathbf{P}(\theta)}{\mathbf{P}(X)}$$

et si on suppose que $\mathbf{P}(\theta)$ est constant par le principe d'agnosticisme, le maximum de vraisemblance revient à maximiser $\theta \mapsto \mathbf{P}(X|\theta)$, c'est-à-dire à choisir le paramètre qui rend les données observées les plus vraisemblables.

Fin cours #12 du 10 décembre

6.2 Exemples

Étudions en détail le cas de l'estimation du paramètre d'une loi de Bernoulli. La vraisemblance est

$$L(\theta|X_1,\dots,X_n) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} = \theta^{S_n}(1-\theta)^{n-S_n}$$

où l'on a posé $S_n = X_1 + \cdots + X_n$. On a

$$\log L = S_n \log\theta + (n - S_n)\log(1-\theta)$$

et cette fonction est maximale si

$$\frac{S_n}{\theta} - \frac{n-S_n}{1-\theta} = 0$$

ou encore $(1-\theta)S_n = \theta(n-S_n)$ soit $\theta = S_n/n$. L'estimateur par maximum de vraisemblance est donc la moyenne empirique

$$\hat{\theta} = \frac{S_n}{n}$$

Étudions l'erreur commise par cette estimation. Étant donné $\delta > 0$, on a en utilisant l'inégalité de Chernoff II

$$\begin{aligned}
\mathbf{P}(\theta \notin [\tilde{\theta}-\delta, \tilde{\theta}+\delta]) &= \mathbf{P}(\theta \notin [\tilde{\theta}-\delta, \tilde{\theta}+\delta]) \\
&= \mathbf{P}(S_n < n\theta(1-\delta/\theta) + \mathbf{P}(S_n < n\theta(1-\delta/\theta) \\
&\leqslant 2\exp\left(-\frac{\delta^2/\theta^2}{2+\delta/\theta}\theta n\right) \\
&= 2\exp\left(-\frac{\delta^2}{2\theta+\delta}n\right) \\
&\leqslant 2\exp\left(-\frac{\delta^2}{3}n\right)
\end{aligned}$$

On en déduit que si $n \geqslant 3\log(2/\gamma)/\delta^2$, alors

$$\mathbf{P}(\theta \in [\tilde{\theta}-\delta, \tilde{\theta}+\delta]) > 1-\gamma$$

est vérifiée. On dit que l'on a déterminé un *intervalle de confiance* pour le paramètre θ.

Exercice. Quel est l'estimateur par maximum de vraisemblance pour le paramètre d'une loi géométrique ?

Voici un exemple dans le cas continu. On considère la famille des lois gaussiennes $\mathsf{N}(\mu, \sigma^2)$, de densité

$$f_{\mu,\sigma^2} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Le paramètre est un couple $\theta = (\mu, v) \in \mathbf{R} \times \mathbf{R}_+$ (on pose $v = \sigma^2$). La vraisemblance vaut

$$\begin{aligned} L(\mu, v | X_1, \ldots, X_n) &= \prod_{i=1}^{n} f_{\mu,v}(X_i) \\ &= \frac{1}{(2\pi v)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (X_i - \mu)^2}{2v}\right) \end{aligned}$$

On a donc

$$\log L = -\frac{\sum_{i=1}^n (X_i - \mu)^2}{2v} - \frac{n}{2}\log(2\pi v)$$

Les conditions $\frac{\partial \log L}{\partial \mu} = \frac{\partial \log L}{\partial v}$ donnent

$$\sum_{i=1}^{n} (X_i - \mu) = 0 \quad \text{et} \quad \frac{\sum (X_i - \mu)^2}{2v^2} - \frac{n}{2v} = 0$$

On en déduit que l'estimateur par maximum de vraisemblance $\hat{\theta} = (\hat{\mu}, \hat{v})$ est donné par la moyenne et la variance empirique

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2$$

6.3 Vecteurs aléatoires gaussiens

On appelle vecteur aléatoire une variable aléatoire à valeurs dans $\mathbf{R}^n$. Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire. On dit que X a une densité s'il existe une fonction $f : \mathbf{R}^n \to \mathbf{R}^+$ telle que pour tout partie (borélienne) $A \subset \mathbf{R}^n$

$$\mathbf{P}((X_1, \ldots, X_n) \in A) = \int_A f(x_1, \ldots, x_n), \mathrm{d}x_1, \ldots \mathrm{d}x_n$$

et on dit que f est la densité du vecteur $(X_1, \ldots, X_n)$

Notons que si X est une variable aléatoire a densité, le vecteur aléatoire (X, X) n'est pas à densité.

Si $X_1, \ldots, X_n$ sont des variables aléatoires continues indépendantes, de densités respectives $f_1, \ldots, f_n$, alors le vecteur aléatoire $X = (X_1, \ldots, X_n)$ a pour densité la fonction

$$(x_1, \ldots, x_n) \mapsto f_1(x_1) \ldots f_n(x_n)$$

Si $X = (X_1, \ldots, X_n)$ est un vecteur aléatoire de densité $f_X : \mathbf{R}^n \to \mathbf{R}_+$, alors pour tout $1 \leqslant i \leqslant n$, la variable aléatoire X_i est continue, et sa densité f_{X_i} est donnée par

$$f_{X_i}(t) = \int_0^\infty \cdots \int_0^\infty f_X(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_{i-1} \mathrm{d}x_{i+1} \ldots \mathrm{d}x_n$$

On dit que $f_{X_1}, \ldots, f_{X_n}$ sont les densités marginales de f_X.

Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire telle que chaque variable aléatoire X_i admette un moment d'ordre 2. L'espérance de X est

$$\mathbf{E}[X] = (\mathbf{E}[X_1], \ldots, \mathbf{E}[X_n]) \in \mathbf{R}^n$$

et la matrice de variance-covariance est de X est la matrice $\mathrm{Cov}(X) = (\Sigma_{ij})$ donnée par

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbf{E}[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])$$

Autrement dit, $\Sigma = (\Sigma_{ij})$ est la matrice $\mathbf{E}[(X-m)(X-m)^t]$ avec $m = \mathbf{E}[X]$.

La matrice de covariance est une matrice symétrique et positive. On peut se ramener au cas où $\mathbf{E}[X] = 0$. Ensuite, pour tout $t = (t_1, \ldots, t_n) \in \mathbf{R}^2$, on a

$$\begin{aligned} \langle t, \mathrm{Cov}(X)t\rangle &= \sum_{i,j=1}^{n} t_i t_j \, \mathbf{E}[X_i X_j] \\ &= \mathbf{E}\left(\sum_{i=1}^{N} t_i X_i\right)^2 \\ &\geqslant 0 \end{aligned}$$

Soit $X = (X_1, \ldots, X_n)$ un vecteur aléatoire de moyenne $\mu \in \mathbf{R}^d$ et de matrice de covariance Σ. Pour toute matrice $A \in \mathsf{M}_n(\mathbf{R})$ et pour tout $b \in \mathbf{R}^n$, le vecteur aléatoire $Y = A(X) + b$ a pour moyenne $A(\mu) + b$ et pour matrice de covariance $A\Sigma A^t$.

C'est une simple conséquence de la linéarité de l'espérance. Pour la moyenne on a :

$$\mathbf{E}[AX] = A\,\mathbf{E}[X] = Am$$

et pour la matrice de covariance :

$$\begin{aligned} \mathbf{E}[(AX - Am)(AX - Am)^t] &= \mathbf{E}[A(X-m)(X-m)^t A^t] \\ &= A\,\mathbf{E}[(X-m)(X-m)^t]A^t \\ &= A\Sigma A^t. \end{aligned}$$

Définition. On dit qu'un vecteur aléatoire $(X_1, \ldots, X_n)$ est un *vecteur gaussien*, si pour tout $t \in \mathbf{R}^n$ la variable aléatoire

$$\langle t, X\rangle = \sum_{i=1}^{n} t_i X_i$$

suit une loi gaussienne (éventuellement constante).

Il ne suffit pas que chacune des coordonnées suive une loi gaussienne pour qu'un vecteur soit gaussien. Voici un exemple qui illustre ce point : si $X \sim N(0,1)$ et ε est une variable aléatoire indépendante de X et vérifiant $\mathbf{P}(\varepsilon = 1) = \mathbf{P}(\varepsilon = -1) = \frac{1}{2}$, alors $(X, \varepsilon X)$ n'est pas un vecteur gaussien (puisque $\mathbf{P}(X + \varepsilon X = 0) = 1/2$) bien que les variables aléatoires X et εX soient toutes deux gaussiennes.

Si $X = (X_1, \ldots, X_n)$ est un vecteur gaussien, pour toute matrice $A \in \mathsf{M}_n(\mathbf{R})$ et tout vecteur $b \in \mathbf{R}^n$, le vecteur aléatoire $AX + b$ est un vecteur gaussien.

Proposition. *Soit $m \in \mathbf{R}^n$ et Σ une matrice symétrique positive $n \times n$. Il existe un vecteur gaussien d'espérance m et de matrice de covariance Σ.*

Démonstration. Commençons par remarquer que si $X_1, \dots, X_n$ sont des variables aléatoires i.i.d. de loi $\mathsf{N}(0,1)$, alors le vecteur aléatoire $(X_1, \dots, X_n)$ est un vecteur aléatoire gaussien d'espérance 0 et de matrice de covariance la matrice identité I_n.

Pour le cas général, on utilise le fait que tout matrice symétrique positive Σ peut s'écrire comme $\Sigma = AA^t$ pour $A \in \mathsf{M}_n(\mathbf{R})$. On peut alors vérifier que le vecteur aléatoire $AX + b$ est un vecteur gaussien d'espérance m et de matrice de covariance Σ. □

Théorème (admis)**.** *Deux vecteurs gaussiens ont même loi si et seulement si ils ont la même espérance et la même matrice de covariance.*

On note $\mathsf{N}(m, \Sigma)$ la loi d'un vecteur gaussien de moyenne $m \in \mathbf{R}^n$ et de matrice de covariance Σ. On dit que la loi $\mathsf{N}(0, \mathrm{Id})$ est la loi d'un vecteur gaussien standard. Si X est un vecteur gaussien standard dans $\mathbf{R}^n$, ses coordonnées sont des variables aléatoires i.i.d. de loi $\mathsf{N}(0,1)$. De plus, si O est un matrice orthogonale (donc vérifiant $OO^t = \mathrm{Id}$), le vecteur gaussien $O(X)$ est un vecteur gaussien standard puisque sa covariance vaut $O \cdot I_n \cdot O^t = I_n$. On dit que la loi gaussienne est *invariante par rotation.* Une autre manière de dire les choses est que les coordonnées d'un vecteur gaussien standard calculées dans une base orthonormale quelconque de $\mathbf{R}^n$ sont indépendantes de loi $\mathsf{N}(0,1)$

Si Σ est inversible, on peut calculer que la loi $\mathsf{N}(m, \Sigma)$ a une densité donnée par

$$x \mapsto \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp\left(-\langle x - m, \Sigma^{-1}(x-m)\rangle\right) dx_1 \dots dx_n$$

Les lois gaussiennes sont omniprésentes dans l'étude des phénomènes de grande dimension, en particulier à cause du théorème central limite. Nous allons étudier deux problèmes qui illustrent ce phénomène.

6.4 Comment tirer une direction uniformément au hasard en grande dimension ?

On cherche à tirer dans l'espace euclidien $\mathbf{R}^n$ avec $n \gg 1$ une direction «uniformément au hasard». Cela revient à choisir un point sur la sphère $S^{n-1} = \{x \in \mathbf{R}^n \ : \ \|x\| = 1\}$ (où l'on note $\|x\| = \left(x_1^2 + \dots + x_n^2\right)^{1/2}$ la norme euclidienne) selon la «mesure de probabilité uniforme». Nous ne définirons pas exactement cette dernière ; c'est l'unique mesure de probabilité invariante par rotation.

Un algorithme naïf est de choisir un $Y_1, \dots, Y_n$ i.i.d. de loi uniforme dans l'intervalle $[-1, 1]$. Conditionnellement à l'événement $E = \{\|Y\| \leqslant 1\}$, le vecteur $\frac{Y}{\|Y\|}$ est de loi uniforme sur S^{n-1}.

L'inconvénient de cet algorithme est que son temps d'exécution est exponentiel ! En effet, son temps d'exécution suit une loi géométrique de paramètre $\mathbf{P}(E)$ et a donc pour espérance $\mathbf{P}(E)^{-1}$. On a

$$\mathbf{P}(E) = \mathbf{P}\left(\sum_{i=1}^n Y_i^2 \leqslant 1\right)$$

On est dans le cadre d'application des inégalités de Hoeffding puisque les variables aléatoires Y_i^2 sont indépendantes et à valeurs dans $[0,1]$. On calcule que $\mu = \mathbf{E}[Y_1^2 + \dots + Y_n^2] = n\,\mathbf{E}\,Y_1^2 = n/3$. On a donc pour $n \geqslant 6$

$$\mathbf{P}(E) \leqslant \mathbf{P}\left(\sum_{i=1}^n Y_i^2 \leqslant n/6\right) = \mathbf{P}\left(\sum_{i=1}^n Y_i^2 \leqslant \mu - n/6\right) \leqslant \exp(-n/3)$$

et le temps moyen d'exécution est donc $\geqslant \exp(n/3)$.

La bonne méthode est de choisir $(z_1, \dots, z_n)$ i.i.d. de loi $\mathsf{N}(0,1)$; alors le vecteur renormalisé $\frac{z}{\|z\|}$ est de loi uniforme sur S^{n-1} et cette méthode prend un temps $O(n)$.

6.5 Lemme de Johnson–Lindenstrauss

Théorème (Lemme de Johnson–Lindenstrauss)**.** *Soit $\varepsilon \in (0, 1/2)$, $Q \subset \mathbf{R}^d$ un ensemble de N points et $k = \lceil 20 \log(N)/\varepsilon^2 \rceil$. Il existe une application linéaire $f : \mathbf{R}^d \to \mathbf{R}^k$ telle que, pour tous u et v dans Q*

$$(1-\varepsilon)\|u-v\|^2 \leqslant \|f(u)-f(v)\|^2 \leqslant (1+\varepsilon)\|u-v\|^2.$$

Ce lemme permet de compresser la géométrie de l'ensemble Q dans un ensemble similaire de beaucoup plus petite dimension. Il est extrêmement utilisé, par exemple dans des problèmes d'apprentissage.

L'idée du lemme est de choisir f au hasard et de montrer qu'elle convient avec grande probabilité. C'est à nouveau une illustration de la méthode probabiliste.

Soit $X = (X_1, \dots, X_k)$. Un vecteur gaussien standard dans $\mathbf{R}^k$. La loi de $\|X\|^2 = X_1^2 + \dots X_k^2$ s'appelle *loi du chi-deux à k degrés de liberté* et se note $\chi^2(k)$. On utilise le lemme suivant

Lemme. *Soit Z une variable aléatoire de loi $\chi^2(p)$. Alors pour tout $0 < \varepsilon < 1/2$,*

$$\mathbf{P}(Z \geqslant (1+\varepsilon)k) \leqslant \exp\left(-k(\varepsilon^2-\varepsilon^3)/4\right)$$
$$\mathbf{P}(Z \leqslant (1-\varepsilon)k) \leqslant \exp\left(-k(\varepsilon^2-\varepsilon^3)/4\right)$$

Preuve du lemme de Johnson–Lindenstrauss. Soit A une matrice de taille $k \times d$ dont les coefficients sont i.i.d. de loi $\mathsf{N}(0,1)$. Il découle de la propriété d'invariance par rotation des vecteurs gaussiens que pour tout vecteur $X \in \mathbf{R}^d$ de norme 1, le vecteur AX est un vecteur gaussien standard dans $\mathbf{R}^k$.

On pose $f = A/\sqrt{k}$ et on calcule

$$\begin{aligned}
\mathbf{P}(f \text{ ne convient pas}) &\leqslant \sum_{u \neq v \in Q} \mathbf{P}\left(\|f(u)-f(v)\|^2 > (1+\varepsilon)\|u-v\|^2\right) \\
&\quad + \sum_{u \neq v \in Q} \mathbf{P}\left(\|f(u)-f(v)\|^2 < (1-\varepsilon)\|u-v\|^2\right) \\
&\leqslant \sum_{u \neq v \in Q} \mathbf{P}\left(\left\|\frac{f(u)-f(v)}{u-v}\right\|^2 > (1+\varepsilon)\right) \\
&\quad + \sum_{u \neq v \in Q} \mathbf{P}\left(\left\|\frac{f(u)-f(v)}{u-v}\right\|^2 < (1-\varepsilon)\right) \\
&\leqslant 2N^2 \exp(-k(\varepsilon^2-\varepsilon^3)/4)
\end{aligned}$$

et on vérifie que cette dernière quantité est < 1 pour le choix $k = 20 \log N/\varepsilon^2$. □

Enfin, le lemme se prouve de la même manière que les inégalités de Chernoff. On montre seulement la première inégalité, la seconde étant similaire. On peut écrire $Z = X_1^2 + \dots + X_k^2$ avec $(X_1, \dots, X_k)$ un vecteur gaussien standard. On a, pour tout $0 < \lambda < 1/2$

$$\begin{aligned}
\mathbf{P}(Z \geqslant (1+\varepsilon)k) &\leqslant \exp(-\lambda(1+\varepsilon)k)\, \mathbf{E}\left[\exp(\lambda X_1^2 + \dots + \lambda X_n^2)\right] \\
&= \exp(-\lambda(1+\varepsilon)k)\, \mathbf{E}\left[\exp(\lambda X_1^2)\right]^k
\end{aligned}$$

On calcule ensuite

$$\begin{aligned}\mathbf{E}[\exp(\lambda X_1^2)] &= \int_{\mathbf{R}} \exp(\lambda x^2)\exp(-x^2/2)\frac{\mathrm{d}x}{\sqrt{2\pi}} \\ &= \int_{\mathbf{R}} \exp(-x^2(1-2\lambda)/2)\frac{\mathrm{d}x}{\sqrt{2\pi}} \\ &= \frac{1}{\sqrt{1-2\lambda}}\end{aligned}$$

par le changement de variables $y = x\sqrt{1-2\lambda}$. On a donc

$$\mathbf{P}(Z \geqslant (1+\varepsilon)k) \leqslant \left(\frac{\exp(-\lambda(1+\varepsilon))}{\sqrt{1-2\lambda}}\right)^k.$$

On choisit finalement la valeur $\lambda = \frac{\varepsilon}{2(1+\varepsilon)}$, ce qui donne

$$\mathbf{P}(Z \geqslant (1+\varepsilon)k) \leqslant [(1+\varepsilon)\exp(-\varepsilon)]^{k/2}$$

et on conclut à l'aide de l'inégalité $(1+\varepsilon)\exp(-\varepsilon) \leqslant \exp(-(\varepsilon^2-\varepsilon^3)/2)$.

Fin du cours