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 $\mathbf{F}_2 = \{0, 1\}$ (pour simplifier). 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

$$C_{ij} = \sum_k A_{ik} B_{kj}$$

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 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 $\mathrm{mincut}(G)$. Il existe des algorithmes déterministes efficaces, mais il y a plus simple : l'algorithme probabiliste de Karger (1993)

Algorithme (Algorithme de Karger)**.** On choisit une arête de G uniformément au hasard et on la contracte en identifiant sa source et son but, puis on élimine les boucles éventuelles. On itère jusqu'à ce qu'il ne reste que 2 sommets et k arêtes, ce qui donne une coupe de G de taille k.

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*

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

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

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

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

Preuve du lemme. Soit $k = \mathrm{mincut}(G)$ 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

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

$$\vdots$$

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

On a donc

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

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 $A \subset \Omega$

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

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 tels 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

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

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.

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 $(\Omega, \mathcal{F}, \mathbf{P})$ 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 $\Omega = \{0,1\}^n$, $\mathcal{F} = \mathcal{P}(\Omega)$ et $\mathbf{P}$ la probabilité uniforme. On considère pour $k \in [n]$

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

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

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

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 $\bigcup_{i\in I} A_i$, qui a donc probabilité 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

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

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 $[0,1[$.

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 2 février

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.

Définition. Si X est une variable aléatoire, sa *fonction de répartition* (en anglais : *cumulative distribution function*) est la fonction $F_X : \mathbf{R} \to [0, 1]$ définie par

$$F_X(t) = \mathbf{P}(X \leqslant t).$$

Proposition. *Pour toute variable aléatoire X, la fonction F_X est*

1. *croissante,*
2. *continue à droite,*
3. *de limite 0 en $-\infty$,*
4. *de limite 1 en $+\infty$.*

La preuve de la proposition est élémentaire. Réciproquement, toute fonction vérifiant les propriétés 1. à 4. ci-dessus est la fonction de répartition d'une variable aléatoire.

Théorème. *Soient X et Y deux variables aléatoires. Alors $X \sim Y$ si et seulement si $F_X = F_Y$. Autrement dit, deux variables aléatoires ont même loi si et seulement si elles ont même fonction de répartition.*

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 $a < b$

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

La fonction de répartition F_X est la primitive de f_X qui vaut 0 en $-\infty$.

Il existe des variables aléatoires qui ne sont ni discrètes ni continues.

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 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 $\sum \lambda_i \mathbf{P}(A_i) = \sum \mu_j \mathbf{P}(B_j)$.
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, 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 $X_1, \ldots, X_n$ son indépendantes, et si $f_1, \ldots, f_n$ sont des fonctions telles que les variables $f_i(X_i)$ sont intégrables, alors*

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

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 $x_1, \ldots, x_n$ 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 autre 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 (le nombre total de comparaisons effectuées) dépend du choix des pivots. 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

$$C(n) \leqslant 2C\left(\left\lceil \frac{n}{2} \right\rceil\right) + \Theta(n)$$

et donc $C(n) = O(n \log n)$.

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 la méthode 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. Sans perte de généralité, supposons $x_1 < x_2 < \cdots < x_n$. Remarquons que chaque couple d'éléments distincts de S sera comparé 0 ou 1 fois au cours de l'algorithme. 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

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

d'où le résultat. □

Fin cours #2 du 9 février

1.7 Premiers exemples de loi

Si a est un réel, la *mesure de* DIRAC en a, notée δ_a est la loi d'une variable aléatoire qui 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 $Z_1, \ldots, Z_N$ 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. Étant donné $\overline{x} = (x_1, \ldots, x_\ell) \in E^\ell$, soit $H(\overline{x})$ l'événement «$X_1 = x_1$, …, $X_\ell = x_\ell$». Si on pose $j = |\{x_1, \ldots, x_\ell\}|$, alors

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

Soient $k_2, \ldots, k_{j+1}$ des entiers. Puisque l'événement «$Z_Z = k_2$, …, $Z_j = k_j$» est réunion disjointe de tels événements $H(\overline{x})$, on en déduit que

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

ce qui est la probabilité qu'une variable aléatoire de loi $\mathsf{G}(\frac{N+1-j}{N})$ égale k_j. Le résultat en découle en écrivant

$$\mathbf{P}(Z_2 = k_2, \ldots, Z_N = k_N) = \prod_{j=2}^{N} \mathbf{P}(Z_j = k_j | Z_2 = k_2, \ldots, Z_{j-1} = k_{j-1}). \qquad \square$$

On a peut donc écrire

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

la variable aléatoire Y 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 $\mathbf{P}(Y = 0) = \frac{n-1}{1000}$, 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, pout 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 ordre 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 $X_1, \dots, X_n$ 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 16 février

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

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

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 ε^{tX} et en justifiant les calculs à l'aide du théorème de convergence dominée : *si une suite (Z_n) de variable 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

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

si $|t| < |\ln(1-p)|$ et la loi géométrique admet donc des moments de tous les ordres.

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 $\mathsf{B}(p)$.

Théorème (Inégalité de Chernoff I, cas symétrique)**.** *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*

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

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

Démonstration. Réalisons X comme $X_1 + \cdots + X_n$ où les (X_i) sont i.i.d. de loi $\mathsf{B}(1/2)$. Il est utile de remplacer $\{0,1\}$ par $\{-1,1\}$ pour que la moyenne soit 0. On pose donc $Y_i = 2X_i - 1$; les variables aléatoires (Y_i) sont i.i.d. de loi uniforme sur $\{-1,1\}$. On pose aussi $Y = Y_1 + \cdots + Y_n = 2X - n$. On a donc

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

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

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

$$\mathbf{P}(Y \geqslant 2a) = \mathbf{P}(\exp(tY) \geqslant \exp(2ta) \leqslant e^{-2ta}\,\mathbf{E}[e^{tY}] = e^{-2ta} M_Y(t).$$

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 $M_Y(t) \leqslant \exp(nt^2/2)$. On a donc

$$\mathbf{P}(Y \geqslant 2a) \leqslant e^{nt^2/2-2ta}$$

et on optimise sur t en choisissant la valeur $t = 2a/n$, 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

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

$$\mathbf{P}(X \geqslant \frac{3}{4}n) \leqslant \frac{4}{n} \; textnormalparl'ingalitdeTchebychev$$

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

La variance de X est $n/4$ et son écart-types est $\sigma = \sqrt{n}/2$. On peut donc réécrire l'inégalité de Chernoff sous la forme

$$\mathbf{P}(X \geqslant \mu + t\sigma) \leqslant e^{-t^2/2}.$$

Cette inégalité est extrêmment précise. On verra plus tard que le membre de gauche converge (par le théorème central limite) vers

$$\frac{1}{\sqrt{2\pi}} \int_t^\infty \exp(-x^2/2)\,\mathrm{d}x$$

et cette quantité est équivalente à $\exp(-t^2/2)/t\sqrt{2\pi}$ lorsque t tend vers l'infini : l'exposant dans l'exponentielle donné par l'inégalité de Chernoff est optimale.

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, cas général)**.** *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes, avec $X_k \sim \mathsf{B}(p_k)$. On pose $X = X_1 + \cdots + X_n$ et $\mu = \mathbf{E}[X] = p_1 + \cdots + p_n$. Pour tout $\varepsilon > 0$, on a pour tout $\varepsilon > 0$,*

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

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

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

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

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 $t_1 = \ln(1+\varepsilon)$ pour obtenir

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

Enfin, on conclut par l'inégalité suivante.

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

Pour la deuxième inégalité, on peut suppose $0 < \varepsilon < 1$. On écrit alors, pour $t < 0$ à déterminer

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

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

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

et conclut grâce au lemme suivant.

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

Enfin, les deux lemmes se peuvent se démontrer par de banales études de fonction. □

Les bornes données par l'inégalité de Chernoff dans la cas général ne sont pas symétriques.

Remarque. Si $0 < \varepsilon < 1$, alors on $\frac{\varepsilon^2}{2+\varepsilon} \geqslant \frac{\varepsilon^2}{3}$ et on peut donc écrire pour $0 < \varepsilon < 1$

$$\mathbf{P}(X \geqslant (1+\varepsilon)\mu) \leqslant \exp\left(-\frac{\varepsilon^2}{3}\mu\right),$$

et donc les bornes pour les déviations en-dessous ou au-dessus de la moyenne sont de même nature.

Remarque. Si $\varepsilon > 1$, alors $\mathbf{P}(X \leqslant (1-\varepsilon)\mu) = 0$ et majorer cette quantité n'a que peu d'intérêt. Nous allons maintenant voir pourquoi il n'est pas possible d'obtenir une majoration du type

$$\mathbf{P}(X \geqslant (1+\varepsilon)\mu \leqslant \exp(-c\varepsilon^2\mu),$$

où $c > 0$ est une constante, qui soit valable pour tout $\varepsilon > 0$. Le problème survient pour des grandes valeurs de ε.

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

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

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

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

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 $t > 0$,*

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

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.

Fin cours #4 du 23 février

2.4 Applications des inégalités de Chernoff

2.4.1 Estimation de paramètre

On souhaite estimer la proportion p d'individus dans une population P qui partagent une caractéristique donnée (opinion politique, présence d'un gène, ...) à partir d'échantillons, c'est-à-dire de la connaissance de n individus $I_1, \ldots, I_n$ choisis indépendamment selon la loi uniforme sur P.

Les variables aléatoires $X_1, \ldots, X_n$ définies par

$$X_i = \begin{cases} 1 & \text{si l'individu } I_i \text{ possède la caractéristique étudiée} \\ 0 & \text{sinon} \end{cases}$$

sont i.i.d. de loi $\mathsf{B}(p)$. Étant donnés $\delta > 0$ et $\gamma > 0$, on souhaite définir une variable aléatoire $\tilde{p} = \tilde{p}(X_1, \ldots, X_n)$ de telle sorte que l'intervalle aléatoire $[\tilde{p} - \delta, \tilde{p} + \delta]$ vérifie la condition

$$\mathbf{P}(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) > 1 - \gamma$$

quelle que soit la valeur de $p \in [0, 1]$.

On choisit pour cela $\tilde{p} = \frac{1}{n}(X_1 + \cdots + X_n)$. On a alors, en utilisant l'inégalité de Chernoff II

$$\begin{aligned} \mathbf{P}(p \notin [\tilde{p} - \delta, \tilde{p} + \delta]) &= \mathbf{P}(p \notin [\tilde{p} - \delta, \tilde{p} + \delta]) \\ &= \mathbf{P}(X_1 + \cdots + X_n < np(1 - \delta/p) + \mathbf{P}(X_1 + \cdots + X_n < np(1 - \delta/p) \\ &\leqslant 2 \exp\left(-\frac{\delta^2/p^2}{2 + \delta/p} pn\right) \\ &= 2 \exp\left(-\frac{\delta^2}{2p + \delta} n\right) \\ &\leqslant 2 \exp\left(-\frac{\delta^2}{3} pn\right) \end{aligned}$$

On en déduit que si $n \geqslant 3 \log(2/\gamma)/\delta^2$, alors, *indépendamment de la taille de la population*, la condition

$$\mathbf{P}(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) > 1 - \gamma$$

est vérifiée. On dit que l'on a déterminé un *intervalle de confiance non asymptotique* pour le paramètre p.

2.5 Partage équilibré

Soit A une matrice $n \times m$ à coefficients dans $\{0, 1\}$, avec $m \gg n$. 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 du 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'invidivus 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, avec $n \gg k$. Lorsque les tâches sont attribuées au hasard (uniformément, indépendamment), quel serveur a la charge maximale ?

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 $M = \max(X_1, \ldots, X_k)$.

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

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

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. $\square$

Chapitre 3

Quelques modèles probabilistes importants

3.1 Le processus de branchement

On modélise un arbre généalogique aléatoire. Le paramètre est une mesure de probabilités sur $\mathbf{N}$, appelée *loi de branchement* et notée μ par la suite.

À la génération 0, la population est constituée d'un unique individu. Cet individu donne naissance à un nombre aléatoire d'enfants distribué selon la loi μ, qui forment la génération 1. Chaque de ces individus suit le même principe, et les variables aléatoires déterminant le nombre d'enfants de chaque individus sont supposés indépendants. On se demande alors à quelle condition sur μ la population a une chance non nulle de survivre éternellement.

Comme le nombre d'individus à une génération donnée n'est pas a priori borné, il est commode de se donner une famille $(\xi_{n,i})_{n\geqslant 0, i\geqslant 1}$ de variables aléatoires i.i.d. de loi μ. La variable $\xi n, i$ donnera le nombre d'enfants du $i$ème individu de la $n$ème génération.

Le nombre Z_n d'individus nés à la $n$ème génération est alors donné $Z_0 = 1$ et la formule de récurrence

$$Z_{n+1} = \sum_{i=0}^{Z_n} \xi_{n,i}.$$

On considère les événements complémentaires A (survie) et B (extinction) donnés par

$$A = \text{« } \forall n \in \mathbf{N},\ Z_n > 0 \text{ »},\ B = \text{« } \exists n \in \mathbf{N},\ Z_n = 0 \text{ »}$$

Théorème. *On suppose $\mu \neq \delta_1$ et on pose*

$$m = \sum_{k=0}^{\infty} k\mu(k) = \mathbf{E}[\xi_{0,1}].$$

1. *Si $m \leqslant 1$, alors $\mathbf{P}(A) = 0$.*
2. *Si $m > 1$, alors $\mathbf{P}(A) > 0$.*

De plus, $\mathbf{P}(A)$ est la plus petite solution dans $[0,1]$ à l'équation $g(s) = s$, où

$$g(s) = \mathbf{E}\left[s^{\xi_{0,1}}\right] = \sum_{k=0}^{\infty} \mu(k) s^k.$$

Remarquons que g est une série entière de rayon de convergence $\geqslant 1$ qui vérifie $g(1) = 1$.
Définissons pour $n \in \mathbf{N}$ la fonction $g_n = \underbrace{g \circ g \circ \cdots \circ g}_{n \text{ fois}}$.

Lemme. *Pour tout $n \geqslant 1$, on a pour $s \in [0,1]$*

$$g_n(s) = \mathbf{E}\left[s^{Z_n}\right].$$

Démonstration. On démontre le résultat par récurrence sur n. La formule est vraie pour $n = 1$. On écrit pour $s \in [0,1]$

$$\begin{aligned}\mathbf{E}\left[s^{Z_{n+1}}\right] &= \mathbf{E}\left[\sum_{k=0}^{\infty} s^{\xi_{n,1}+\cdots+\xi_{n,k}} \mathbf{1}_{\{Z_n=k\}}\right] \\ &= \sum_{k=0}^{\infty} \mathbf{E}\left[s^{\xi_{n,1}+\cdots+\xi_{n,k}} \mathbf{1}_{\{Z_n=k\}}\right],\end{aligned}$$

l'échange étant justifié car toutes les quantités impliquées sont positives. Par le lemme des coalitions, les variables aléatoires $Z_n, \xi_{n,1}, \ldots, \xi_{n,k}$ sont indépendantes et donc

$$\mathbf{E}\left[s^{Z_{n+1}}\right] = \sum_{k=0}^{n} \mathbf{P}(Z_n = k) g_n(s)^k$$

en utilisant l'hypothèse de récurrence au rang n. On a donc $\mathbf{E}\left[s^{Z_{n+1}}\right] = g(g_n(s))$, ce qui conclut la preuve. □

Preuve du théorème. On pose $u_n = \mathbf{P}(Z_n = 0)$. La suite (u_n) est croissante et vérifie $u_0 = 0$ et $u_{n+1} = g(u_n)$. Comme la fonction g est croissante, continue et vérifie $g(1) = 1$, la suite (u_n) converge vers la plus petite solution dans $[0,1]$ à l'équation $g(s) = s$.

Remarquons que $g'(s) = \mathbf{E}[\xi_{0,1}] = m$. Si $m > 1$, alors pour ε assez petit on a $g(1-\varepsilon) < 1-\varepsilon$. Comme par ailleurs $g(0) = \mu(0) \geqslant 0$, le théorème des valeurs intermédiaires impliques g a un point fixe dans $[0, 1-\varepsilon]$.

Si $m \leqslant 1$, comme la fonction g est strictement convexe (puisque $\mu \neq \delta_1$), elle est strictement au-dessus de sa tangente au point 1. On en déduit que $g(1-t) > 1-t$ pour tout $t \in]0,1]$. □

Fin cours #5 du 8 mars

3.2 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 $V = \{1, \ldots, n\}$ 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 $\{1, \ldots, n\}$. 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.
— 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. □

On peut en réalisé 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 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

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

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

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

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

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

d'où le résultat. □

Voici à quoi ressemblent les composantes connexes d'un graphe aléatoire de loi $\mathsf{G}_{n,p}$ (les énoncés suivants sont vraies avec probabilité tendant vers 1 quand $n \to \infty$).

— Si $p \leqslant (1-\varepsilon)/n$, alors toutes les composantes connexes ont taille $O(\log n)$
— Si $(1+\varepsilon)/n \leqslant p \leqslant (1-\varepsilon)\log(n)/n$, il existe une (unique) composante connexe «géante» de taille $\Theta(n)$.
— Si $p \geqslant (1+\varepsilon)\log(n)/n$, le graphe est connexe.

3.3 Le processus de Poisson

On dit qu'une variable aléatoire X suit la loi $\mathsf{E}(\lambda)$, loi exponentielle de paramètre $\lambda > 0$, si elle a pour densité

$$f_\lambda : x \mapsto \lambda e^{-\lambda x} \mathbf{1}_{\{x \geqslant\}}.$$

On a alors, pour $t \geqslant 0$, la valeur $\mathbf{P}(X \geqslant t) = \int_t^\infty f_\lambda(x)\,\mathrm{d}x = e^{-\lambda t}$. On calcule aussi que $\mathbf{E}[X] = 1/\lambda$ et $\mathbf{Var}(X) = 1/\lambda^2$.

Les lois exponentielles sont l'analogue en temps continu des lois géométriques. En particulier, si une variable aléatoire $X \sim \mathsf{E}(\lambda)$ vérifie la propriété d'absence de mémoire en temps continu : pour tous $s, t > 0$

$$\mathbf{P}(X > s + t | X > t) = \mathbf{P}(X > s).$$

La famille des lois exponentielles est stable par minimum.

Proposition. *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes, avec $X_i \sim \mathsf{E}(\lambda_i)$. Soit $Y = \min(X_1, \ldots, X_n)$. Alors $Y \sim \mathsf{E}(\lambda_1 + \cdots + \lambda_n)$.*

Démonstration. Il suffit de faire le cas $n = 2$ et de faire ensuite une preuve par récurrence puisque $\min(X_1, \ldots, X_n) = \min(\min(X_1, \min(X_2, \ldots, X_n))$. Par indépendance,

$$\mathbf{P}(Y \geqslant t) = \mathbf{P}(X_1 \geqslant t, X_2 \geqslant t) = \mathbf{P}(X_1 \geqslant t)\mathbf{P}(X_2 \geqslant t) = e^{-\lambda_1 t} e^{-\lambda_2 t} = e^{-(\lambda_1+\lambda_2)t}. \quad \square$$

On peut maintenant introduire le processus de Poisson. Soit $(X_j)_{j \geqslant 1}$ une suite de variables aléatoires i.i.d. de loi $\mathsf{E}(\lambda)$. On pose $T_0 = 0$ et pour $j \in \mathbf{N}$

$$T_j = X_1 + \ldots X_j$$

puis $N_t = \max\{k \ : \ T_k \leqslant t\}$. La famille de variables aléatoires $(N_t)_{t \geqslant 0}$ s'appelle un processus de Poisson de paramètre λ.

Ce processus est utilisé pour modéliser le temps d'apparition d'événements dans un contexte chaotique. Par exemple, les temps de passage d'un autobus à un arrêt donné sont en théorie déterministes, mais en pratique en cas de forte circulation, il peut être plus réaliste de les modéliser par un processus de Poisson. Le temps de passage du $j$ème bus est T_j et le nombre de passages entre les instants 0 et t est donné par N_t. Le paramètre λ est la fréquence de passage des autobus.

Théorème. *Soit $(N_t)_{t \geqslant 0}$ un processus de Poisson d'intensité $\lambda > 0$. Alors*

1. *Pour tout $t \geqslant 0$, la variable aléatoire N_t suit la loi $\mathsf{P}(\lambda t)$*
2. *Le processus (N_t) est «à accroissements indépendants», c'est-à-dire que pour tous $s_1 < t_1 < s_2 < t_2 < \cdots < s_k < t_k$, les variables aléatoires $(N_{t_i} - N_{s_i})_{1 \leqslant i \leqslant k}$ sont indépendantes et de loi $\mathsf{P}(\lambda(t_i - s_i))$.*

Démonstration. Puisque $N_t \geqslant n \iff T_n \leqslant t$, on peut déduire la loi de N_t de celle de T_n. On calcule la densité de la variable aléatoire T_n comme étant

$$t \mapsto \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t} \mathbf{1}_{\{t \geqslant 0\}}.$$

Pour cela on utilise la formule de convolution : si X de densité f_X et Y de densité f_Y sont indépendantes, alors $X + Y$ a pour densité $f_X \star f_Y := z \mapsto \int f_X(x) f_Y(z - x)\,\mathrm{d}x$. C'est l'analogue continu de la formule

$$\mathbf{P}(X + Y = n) = \sum_{k+l=n} \mathbf{P}(X = k)\mathbf{P}(Y = l)$$

pour des variables discrètes. Si f_n est la densité de T_n, alors $f_1(t) = \lambda e^{-\lambda t}$. De $f_n(t) = \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t} \mathbf{1}_{\{t \geqslant 0\}}$ on calcule par récurrence pour $t \geqslant 0$

$$\begin{aligned} f_{n+1}(t) = f_n \star f_1(t) &= \int_0^t \frac{\lambda^n u^{n-1}}{(n-1)!} e^{-\lambda u} \lambda e^{-\lambda(t-u)}\,\mathrm{d}u \\ &= \frac{\lambda^{n+1} e^{-\lambda t}}{(n-1)!} \int_0^t u^{n-1}\,\mathrm{d}u = \frac{\lambda^{n+1} e^{-\lambda t} t^n}{n!} \end{aligned}$$

On écrit alors, pour $n \geqslant 1$,

$$\begin{aligned} \mathbf{P}(N_t = n) &= \mathbf{P}(N_t \geqslant n) - \mathbf{P}(N_t \geqslant n+1) \\ &= \mathbf{P}(T_n \leqslant t) - \mathbf{P}(T_{n+1} \leqslant t) \\ &= \int_0^t \left(\frac{\lambda^n x^{n-1}}{(n-1)!} e^{-\lambda x} - \frac{\lambda^{n+1} x^n}{n!} e^{-\lambda x} \right) \mathrm{d}x \end{aligned}$$

La fonction intégrée est la dérivée de $x \mapsto e^{-\lambda x} \frac{(\lambda x)^n}{n!}$ et donc $\mathbf{P}(N_t = n) = e^{\lambda t}(\lambda t)^n / n!$, qui est la probabilité qu'une variable aléatoire de loi $\mathsf{P}(\lambda t)$ égale n.

La deuxième partie découle par récurrence sur k du fait que pour tout $s > 0$, la famille de variables aléatoires $(N_{s+t} - N_s)_{t \geqslant 0}$ est un processus de Poisson qui est indépendant de N_t. C'est une conséquence de la propriété d'absence de mémoire des lois exponentielles ; nous n'écrivons pas les détails qui sont fastidieux. □

Expliquons maintenant ce qu'on appelle le paradoxe de l'autobus en conservant les notations précédentes. Pour $t > 0$ fixé, on a $t \in [T_{N_t}, T_{N_t+1}]$. Dans la modélisation des temps de massage d'un autobus, pour un observateur qui arrive a un instant t fixé, T_{N_t} est l'instant de passage du bus précédent et T_{N_t+1} est l'instant de passage du prochain bus. Soit $U_t = t - T_{N_t}$ (le temps depuis lequel le dernier bus est passé) et $V_t = T_{N_{t+1}} - t$ (le temps à atteindre depuis le prochain bus).

Lemme. *Pour t fixé, les variables aléatoires U_t et V_t sont indépendantes, $V_t \sim \mathsf{E}(\lambda)$ et $U_t \sim \min(E, t)$ où $E \sim \mathsf{E}(\lambda)$*

Démonstration. Si $0 < u < t$ et $0 < v$, alors

$$P(U_t > u, V_t > v) = \mathbf{P}(N_{t+v} - N_{t-u} = 0) = \mathbf{P}(\mathsf{P}(\lambda(v+u)) = 0) = e^{-\lambda(u+v)}. \qquad \square$$

On a $\mathbf{E}[V_t] = 1/\lambda$ et pour t grand $\mathbf{E}[U_t] = 2/\lambda$. L'intervalle moyen entre deux bus est donc de $2/\lambda$. C'est à comparer au cas déterministe où les intervalles entre deux bus sont tous exactement de longueur $1/\lambda$.

C'est le paradoxe de l'autobus :

— En moyenne, le temps entre deux passages d'autobus est de $1/\lambda$

— En moyenne, le temps observé entre deux passages d'autobus est de $2/\lambda$

Le paradoxe s'explique par le fait que les intervalles longs ont plus de chances d'être observés que les intervalles courts. Pour l'observateur, ce qu'il observe en moyenne dans le modèle du processus de Poisson équivaut au pire cas du processus déterministe.

Fin cours #6 du 22 mars

Chapitre 4

Convergence des variables aléatoires

4.1 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

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

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 1) par longueur de mot, 2) 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.*

(dessin pour illustrer la différence entre la loi forte et la loi faible).

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 probabilités (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 reste 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).

4.2 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

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

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.

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. □

Les lois gaussiennes sont omniprésentes dans l'étude des phénomènes de grande dimension, à cause du théorème central limite. Voici un exemple un peu différent. Considérons le problème suivant : 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 $S = \{x \in \mathbf{R}^n : \|x\| = 1\}$ (où l'on note $\|x\| = \left(x_1^2 + \cdots + x_n^2\right)^{1/2}$ la norme euclidienne) selon la «mesure de probabilité uniforme». Une idée naïve est de choisir un $y_1, \ldots, y_n$ 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. Une application des inégalités de Hoeffding (exercice : écrire les détails) montre que l'événement E a probabilité exponentiellement petite, donc que cette méthode prend un temps exponentiel !

La bonne méthode est de choisir $(z_1, \ldots, z_n)$ i.i.d. de loi $\mathsf{N}(0, 1)$; alors le vecteur renormalisé $\frac{z}{\|z\|}$ est de loi uniforme sur S et cette méthode est beaucoup plus efficace que la précédente !

Fin cours # 7 du 29 mars

On appelle vecteur gaussien standard dans $\mathbf{R}^n$ un n-uplet de variables aléatoires indépendantes de loi $\mathsf{N}(0, 1)$. Si $X = (X_1, \ldots, X_n)$ est un vecteur gaussien standard, alors pour toute matrice orthogonale $O \in O_n(\mathbf{R})$, le vecteur $Y = OX$ est aussi un vecteur gaussien standard. Cette propriété d'invariance par rotation découle du fait que X a pour «densité» la fonction

$$(x_1, \ldots, x_n) \mapsto \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\|x_1\|^2 + \cdots + \|x_n\|^2}{2}\right)$$

qui est une fonction de la norme euclidienne $\|x\|$, et cette dernière est invariante par rotation.

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 un exemple de la méthode probabiliste ; on en verra d'autres au prochain chapitre.

Soit $X = (X_1, \ldots, X_k)$. Un vecteur gaussien standard dans $\mathbf{R}^k$. La loi de $\|X\|^2 = X_1^2 + \ldots X_k^2$ s'appelle *loi du chi-deux à p 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 $Z = X_1^2 + \cdots + X_k^2$ avec $(X_1, \ldots, X_k)$ un vecteur gaussien standard. On a, pour tout $0 < \lambda < 1/2$

$$\begin{aligned}
\mathbf{P}(Z \geqslant (1+\varepsilon)k) &\leqslant \exp(-\lambda(1+\varepsilon)k)\, \mathbf{E}\left[\exp(\lambda X_1^2 + \cdots + \lambda X_n^2)\right] \\
&= \exp(-\lambda(1+\varepsilon)k)\, \mathbf{E}\left[\exp(\lambda X_1^2)\right]^k
\end{aligned}$$

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 $(1+\varepsilon)\exp(-\varepsilon) \leqslant \exp(-(\varepsilon^2 - \varepsilon^3)/2)$.

Chapitre 5

La méthode probabiliste : examples

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.

5.1 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 exite 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 quelle proportion des clauses peut être satisfaite. 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. □

5.2 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}.

5.3 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. *Pour tout n, il existe $A_n \in \{0,1\}^{n \times n}$ telle que*

$$\min_{b \in \{-1,1\}^n} \|A_n b\|_\infty = \Omega(\sqrt{n})$$

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

Lemme. *Il existe une constante $c > 0$ telle que, 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*

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

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. Écrivons $X_i = 2Y_i + 1$, ce qui fait que les variables aléatoires (Y_i) sont i.i.d. de loi uniforme sur $\{-1,1\}$. Les variables aléatoires (Z_i) définies par $Z_i = b_i Y_i$ sont aussi i.i.d. de loi uniforme sur $\{-1,1\}$. On a alors, en posant $B = b_1 + \cdots + b_n$,

$$\sum_{i=1}^n b_i X_i = B + 2\sum_{i=1}^n Z_i$$

puis

$$\mathbf{P}\left(\left|\sum_{i=1}^n b_i X_i\right| \leqslant c\sqrt{n}\right) = \mathbf{P}\left(S \in \left[-\frac{B}{2} - \frac{c\sqrt{n}}{2}, -\frac{B}{2} + \frac{c\sqrt{n}}{2}\right]\right)$$

La quantité de droite vaut 2^n fois la somme de $\lfloor c\sqrt{n}/2 \rfloor$ coefficients binomiaux. Puisque la fonction $k \mapsto \binom{n}{k}$ est croissante pour $k \leqslant n/2$ et décroissante pour $k \geqslant n/2$, cette quantité est maximale en choisissant les coefficients binomiaux autour les plus proches la valeur $k = n/2$, ce qui correspond au cas $B = 0$. Par le théorème central limite,

$$\lim_{n\to\infty} \mathbf{P}\left(S \in \left[-\frac{B}{2} - \frac{c\sqrt{n}}{2}, -\frac{B}{2} + \frac{c\sqrt{n}}{2}\right]\right) = \int_{-c/2}^{c/2} \exp(-x^2)\,\mathrm{d}x$$

et cette quantité peut être rendue $\leqslant 1/2$ en choisissant la constante c suffisamment petite. Ceci prouve le lemme pour n suffisamment grand et les petites valeurs de n peuvent être incluses à ajustant la valeur de c si besoin. □

Fin cours # 8 du 5 avril et fin du programme pour l'examen partiel