

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 $D \in M_n(\mathbf{F}_2)$ une matrice non nulle et $x \in \mathbf{F}_2^n$ choisi uniformément au hasard. Alors

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

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 $(x_1, \dots, x_{n-1})$ fixés, il y a probabilité $\frac{1}{2}$ (sur le choix de x_n) que $(Dx)_1 \neq 0$. $\square$

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

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

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

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

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 $\text{mincut}(G) \leq \text{mincut}(G/e)$ 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*

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

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

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

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 $1 \leq i \leq n-2$, 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 $\geq k$ et donc $|E| \geq \frac{kn}{2}$; on a donc $\mathbf{P}(A_1) \leq \frac{2}{n}$.

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

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

De la même manière, on a

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

$$\vdots$$

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

On a donc

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

Lorsque les événements $A_1, A_2, \dots, A_n$ sont réalisés, la coupe produite par l'algorithme de KARGER est la coupe C . Ceci conclut la preuve du lemme. $\square$

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 $\mathbf{P} : \mathcal{F} \rightarrow [0, 1]$ qui à 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 ≥ 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 $\mathbf{P}(A) \leq \mathbf{P}(B)$. En effet, par σ -additivité (appliquée à une suite d'événements dont tous sauf deux sont vides) on a $\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B \setminus A) \geq \mathbf{P}(A)$. 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*

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

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,

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

d'où le résultat. $\square$

Une question naturelle : pourquoi ne pas toujours prendre $\mathcal{F} = \mathcal{P}(\Omega)$? Quel intérêt y a-t-il à 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* ($A \perp B$) 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

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

et donc la probabilité de A «ne dépend pas» de B . Voilà 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 14 septembre 2026

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 $A_1, \dots, A_n$ 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

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

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 $[0, 1[$. 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.

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 $X : \Omega \rightarrow \mathbf{R}$ telle que, pour tous $a < b$ réels l'ensemble $\{a \leq X \leq b\} = X^{-1}([a, b])$ 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 $X : \Omega \rightarrow E$ 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 $X : \Omega \rightarrow \mathbf{R}$ ne sont pas importants.

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

1. On peut prendre $\Omega = \{1, \dots, 6\}$, $X : \Omega \rightarrow \mathbf{R}$ 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 $f_X : \mathbf{R} \rightarrow \mathbf{R}^+$ continue par morceaux vérifiant $\int_{-\infty}^{\infty} f_X(s) ds = 1$, il existe une variable aléatoire X dont la loi vérifie, pour tout $a < b$

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

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 $\{X_i \leq t_i\}$ 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 $f_{\alpha} : \mathbf{R}^{I_{\alpha}} \rightarrow \mathbf{R}$ 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 $X_n : \Omega \rightarrow \mathbf{R}$ 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 $x_1, \dots, x_n$, 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*

$$\mathbf{E}[Y] = \sum_{k=1}^{\infty} \mathbf{P}(Y \geq k).$$

En effet, $\mathbf{P}(Y \geq k) = \sum_{n=k}^{\infty} \mathbf{P}(Y = n)$ 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

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

Fin cours # 2 du 16 septembre 2026

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 $h : \mathbf{R} \rightarrow \mathbf{R}$, on a

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

dès lors que l'intégrale a un sens.

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

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

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

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

(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 $n \rightarrow \infty$.

Démonstration. Soit $S = \{x_1, \dots, x_n\}$ avec $x_1 < x_2 < \dots < x_n$. 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 $\{x_i, x_{i+1}, \dots, x_j\}$, 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 $\mathbf{B}(p)$ est la loi $p\delta_1 + (1-p)\delta_0$. Une variable aléatoire X a pour loi $\mathbf{B}(p)$, ce qu'on note $X \sim \mathbf{B}(p)$, si et seulement si $\mathbf{P}(X = 1) = p$ et $\mathbf{P}(X = 0) = 1 - p$. La loi $\mathbf{B}(\frac{1}{2})$ est la loi d'un bit aléatoire.

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

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

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

Si $Y \sim \mathbf{G}(p)$, 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 $\mathbf{G}(p)$. 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 $\mathbf{P}(Y > k) = (1-p)^k$. $\square$

Exercice. Soient $Y_1 \sim \mathbf{G}(p_1)$ et $Y_2 \sim \mathbf{G}(p_2)$ 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 $(X_n)_{n \geq 1}$ 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 $1 \leq j \leq N$ 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 $T_2 - 1 \sim \mathbf{G}(\frac{N-1}{N})$. 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 $1 < j \leq N$ sont indépendantes. De plus Z_j suit la loi $\mathbf{G}(\frac{N+1-j}{N})$.

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

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

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 « $Z_2 = k_2, \dots, Z_j = k_j$ » est réunion disjointe de tels événements $H(\bar{x})$). $\square$

On peut donc écrire la variable aléatoire

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

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 sûr 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 ≥ 0 . Alors, pour tout $a > 0$,*

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}.$$

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

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 $a > 0$,

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

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 $1 \leq p \leq q$. Alors pour toute variable aléatoire X on a*

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

Démonstration. Puisque l'inégalité à montrer se réécrit en $\mathbf{E}[|Y|] \leq (\mathbf{E}[|Y|^r])^{1/r}$ 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 $x \geq 0$,

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

d'où on déduit en prenant l'espérance

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

puis $\mathbf{E}[|X|] \leq 1$. □

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 aléatoires 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 $|\mathbf{Cov}(X, Y)| \leq \sigma(X)\sigma(Y)$.

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*

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

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 $a > 0$,*

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

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

Fin cours #3 du 23 septembre