TD 08 – Convergence de variables aléatoires (corrigé)

Exercice 1. *Second theorème de Borel-Cantelli*

L'objectif de cet exercice est de montrer le second théorème de Borel-Cantelli. Il donne une réciproque du theorème de Borel-Cantelli vu en cours, dans le cas où les événements sont indépendants. Soit $(A_n)_{n\in\mathbb{N}}$ une suite d'événements *indépendants* de probabilité p_n. On suppose que la somme $\sum_n p_n$ diverge. L'objectif de cet exercice est de montrer qu'alors, presque sûrement, une infinité d'événements A_n se réalisent.

1. Exprimer l'événement "une infinité d'événements A_n se réalisent" en terme d'unions et d'intersections des événements A_n.

☞ Soit $\omega \in \Omega$ un éléments de l'espace de probabilité. Alors ω appartient à l'événement "une infinité d'événements A_n se réalisent" si et seulement si ω appartient à une infinité de A_n, i.e. $\omega \in \cap_{k\geq 0} \cup_{n\geq k} A_n$. Donc l'événement "une infinité d'événements A_n se réalisent" n'est autre que l'événement $\cap_{k\geq 0} \cup_{n\geq k} A_n$ (aussi appelé $\limsup A_n$).

2. Soit $B_{k,\ell}$ l'événement $\cap_{k\leq n\leq \ell}\overline{A_n}$. Montrer que pour tout k fixé, $\lim_{\ell\to\infty} \mathbf{P}\left\{B_{k,\ell}\right\} = 0$. *Indice : on pourra utiliser l'inégalité $1 + x \leq e^x$ pour tout $x \in \mathbb{R}$.*

☞ Par indépendance des A_n (et donc indépendance de leur complémentaire, cf exercice " complément des indépendants"), on a $\mathbf{P}\{B_{k,\ell}\} = \prod_{n=k}^{\ell}(1-p_n)$. En utilisant l'indice, on a alors $\mathbf{P}\{B_{k,\ell}\} \leq \prod_{n=k}^{\ell} e^{-p_n} = e^{-\sum_{n=k}^{\ell} p_n}$. Mais par hypothèse, la somme des p_n diverge, donc pour tout k fixé, $\lim_{\ell\to\infty}\sum_{n=k}^{\ell} p_n = +\infty$. On conclut que $\lim_{\ell\to\infty}\mathbf{P}\{B_{k,\ell}\} = 0$.

3. On note $B_k = \cap_{n\geq k}\overline{A_n}$. En déduire que $\mathbf{P}\{\cup_k B_k\} = 0$.

☞ Il suffit de montrer que $\mathbf{P}\{B_k\} = 0$ pour tout k. On aura alors $\mathbf{P}\{\cup_k B_k\} \leq \sum_k \mathbf{P}\{B_k\} = 0$. Mais $\mathbf{P}\{B_k\} = \mathbf{P}\{\cap_{n\geq k}\overline{A_n}\} = \lim_{\ell\to\infty}\mathbf{P}\{B_{k,\ell}\}$ (car les événements $B_{k,\ell}$ sont décroissants et leur intersection est égale à B_k). On conclut avec la question précédente que $\mathbf{P}\{B_k\} = 0$.

4. Conclure que $\mathbf{P}\{$"une infinité d'événements A_n se réalisent"$\} = 1$.

☞ On a vu à la première question que l'événement "une infinité d'événements A_n se réalisent" est en fait égal à $\cap_{k\geq 0}\cup_{n\geq k} A_n$. Le complémentaire de cet événement est donc $\cup_{k\geq 0}\cap_{n\geq k}\overline{A_n} = \cup_{k\geq 0} B_k$. On a donc bien $\mathbf{P}\{\cap_{k\geq 0}\cup_{n\geq k} A_n\} = 1 - \mathbf{P}\{\cup_{k\geq 0} B_k\} = 1$ d'après la question précédente.

5. *Application.* Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables de Bernoulli indépendantes de paramètre $\mathbf{P}\{X_n = 1\} = p_n = 1/n$. Montrer que presque sûrement la suite X_n contient un nombre infini de '1', mais seulement un nombre fini de '11'.

☞ Commençons par montrer que presque sûrement, la suite X_n contient un nombre infini de '1'. On note A_n l'événement "$X_n = 1$". On a $\mathbf{P}\{A_n\} = 1/n$, et donc $\sum_n \mathbf{P}\{A_n\}$ diverge. D'après le second théorème de Borel-Cantelli (les A_n sont indépendants car les X_n le sont), on a donc $\mathbf{P}\{$"une infinité d'événements A_n se réalisent"$\} = 1$, ce qui est équivalent à dire que presque sûrement la suite X_n contient une infinité de 1.

Montrons maintenant que presque sûrement la suite X_n ne contient qu'un nombre fini de '11'. On utilise cette fois le théorème de Borel-Cantelli vu en cours. Soit C_n l'événement "$X_n = X_{n+1} = 1$". Par indépendance de X_n et X_{n+1}, on a $\mathbf{P}\{C_n\} = 1/(n^2+n) \leq 1/n^2$. (Remarque : on n'a pas que les C_n sont indépendants, mais l'indépendance n'est pas nécessaire pour utiliser le théorème de Borel-Cantelli dans ce sens.) Donc la somme $\sum_n \mathbf{P}\{C_n\}$ converge. D'après le théorème de Borel-Cantelli, on conclut que presque sûrement, seuls un nombre fini d'événements C_n sont réalisés. C'est-à-dire, presque sûrement il n'y a qu'un nombre fini de '11' dans la suite des X_n.

Comme l'intersection de deux événement presque sûrs est aussi presque sûre, on conclut que presque sûrement la suite X_n contient un nombre infini de '1', mais seulement un nombre fini de '11'.

Exercice 2. *Conditions de convergence*

Soit X_n une suite infinie de variables de Bernoulli indépendantes de paramètres $1 - p_n$, avec $0 \leq p_n \leq 1/2$ (i.e. $\mathbf{P}\{X_n = 1\} = 1 - p_n$ et $\mathbf{P}\{X_n = 0\} = p_n$).

1. Donner une condition nécessaire et suffisante pour que la suite X_n converge en distribution.

☞ Supposons que les variables X_n convergent en distribution vers une variable X. Les fonctions de répartitions F_{X_n} des variables X_n sont comme sur la Figure 1. En particulier, elles sont continues en $1/2$, et pour tout n, on a $F_{X_n}(1/2) = p_n$. Notons $p = F_X(1/2)$. Par définition de la convergence en distribution, on a $\lim_{n\to\infty} p_n = p$ (en particulier, les p_n convergent).

Supposons à l'inverse que les p_n convergent vers une constante p. Comme $[0,1]$ est fermé et les p_n vivent dans $[0,1]$, on en déduit que $p \in [0,1]$. Définissons X la variable de Bernoulli de paramètre p. Alors, on a bien, pour tout $x \neq \{0,1\}$, $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$, i.e. X_n converge en distribution vers X.

On conclut que X_n converge en distribution ssi p_n converge.

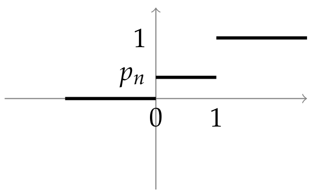

FIGURE 1 – Fonction de répartition de X_n

2. Donner une condition nécessaire et suffisante pour que la suite X_n converge en probabilité.

☞ Comme la convergence en probabilité implique la convergence en distribution, on sait qu'une condition nécessaire est que les p_n converge. Mais ce n'est pas une condition suffisante. Supposons par exemple que $p_n = 1/2$ pour tout n. Alors les p_n sont bien convergents, mais, si je prend $\varepsilon = 1/2$, j'ai $\mathbf{P}\{|X_n - X_{n+1}| \geq \varepsilon\} = \mathbf{P}\{X_n \neq X_{n+1}\} = 1/2$ par indépendance des X_n. En particulier, cette quantité ne tend pas vers zero, donc les X_n ne peuvent pas converger en probabilité. Le problème ici est que les X_n suivent bien la même loi, mais comme ils sont indépendants, rien ne nous assure que leurs valeurs seront proches.

Reprenons notre condition nécessaire. Supposons que X_n converge en probabilité vers X. Alors, pour tout $\varepsilon > 0$, on a $\mathbf{P}\{|X_n - X| \geq \varepsilon\} \to 0$. Par inégalité triangulaire, cela implique en particulier que $\mathbf{P}\{|X_n - X_{n+1}| \geq 2\varepsilon\} \to 0$. Prenons $2\varepsilon = 1/2$, on a alors $\mathbf{P}\{|X_n - X_{n+1}| \geq 2\varepsilon\} = \mathbf{P}\{X_n \neq X_{n+1}\} \geq p_n$. En effet, une fois X_{n+1} fixé, on a $\mathbf{P}\{X_n \neq X_{n+1}\} = p_n$ si $X_{n+1} = 1$ et $\mathbf{P}\{X_n \neq X_{n+1}\} = 1 - p_n$ si $X_{n+1} = 0$. Dans tous les cas, cette probabilité est supérieur à p_n, car on a choisi $p_n \leq 1/2$. On en déduit donc que $p_n \to 0$.

Supposons maintenant $p_n \to 0$, et notons X la variable aléatoire valant toujours 1. On a, pour tout $\varepsilon > 0$

$$\mathbf{P}\{|X_n - X| \geq \varepsilon\} = \mathbf{P}\{X_n = 0\} = p_n \to 0.$$

On en conclut que X_n converge en probabilité vers X.

On a donc que X_n converge en probabilité ssi p_n tend vers 0 (avec la contrainte $p_n \leq 1/2$).

3. Donner une condition nécessaire et suffisante pour que la suite X_n converge presque sûrement.

☞ On a vu que si la suite X_n converge presque sûrement, alors elle converge vers 1 (car elle converge en probabilité). On veut donc montrer que $\mathbf{P}\{X_n \to 1\} = 1$, quitte à faire quelques hypothèses supplémentaires sur les p_n. On sait, d'après le lemme de Borel-Cantelli que si $\sum_n p_n$ converge, alors avec probabilité 1, un nombre fini de variables X_n valent 0 (car les événements " $X_n = 0$" ont probabilité p_n). Mais dire qu'un nombre fini de variables X_n valent 0 est équivalent à dire que X_n converge vers 1 (car les variables X_n sont à valeur dans $\{0, 1\}$). On en déduit donc que si $\sum_n p_n$ converge, alors X_n converge vers 1 presque sûrement.

Pour la réciproque, on utilise le second théorème de Borel-Cantelli (cf exercice "second théorème de Borel-Cantelli"), qui dit que si les X_n sont indépendants et $\sum p_n$ diverge, alors, avec probabilité 1, il existe une infinité de X_n valant 0. En particulier, X_n ne peux pas converger vers 1. On en déduit donc que si X_n converge presque sûrement, alors $\sum_n p_n$ converge.

On a donc que X_n converge presque sûrement ssi $\sum_n p_n$ converge.

Exercice 3. *Calcul de limite*

L'objectif de cet exercice est de montrer l'égalité suivante :

$$\lim_n \exp(-n) \sum_{k=0}^{n} \frac{n^k}{k!} = \frac{1}{2}$$

On considère $(X_n)_{n \geq 0}$ une suite de variables aléatoires indépendantes de loi de Poisson de paramètre 1. On note $S_n = \sum_{k=0}^{n} X_k$.

1. Montrer que pour tout $n \geq 0$, $\mathbf{P}\{S_n \leq n\} = \exp(-n) \sum_{k=0}^{n} \frac{n^k}{k!}$. ☞ S_n est une somme de variables aléatoires indépendante de loi de Poisson de paramètre 1, donc suit une loi de Poisson de paramètre n.

2. Conclure en utilisant le Théorème Central Limite.

☞ D'après le Théorème Central Limite, on a :

$$\mathbf{P}\{()\, S_n \leq n) = \mathbf{P}\{()\, \frac{S_n - n}{\sqrt{n}} \leq 0) \xrightarrow[n \to +\infty]{(d)} \Phi(0) = \frac{1}{2}$$

Exercice 4. *Théorème de Mycielski*

Recall that the *chromatic number* $\chi(G)$ is the smallest number of colors needed to color the vertices of G such that any two adjacent vertices have different colors. Clearly, graphs with large cliques have a high chromatic number, but the opposite is not true. The goal of this exercise is th prove Mycielski's theorem, which states that for any integer $k \geq 2$, there exists a graph G such that G contains no triangles and $\chi(G) \geq k$.

1. Fix $0 < \varepsilon < \frac{1}{3}$ and let G be a random graph on n vertices where each edge appears independently with probability $p = n^{\varepsilon-1}$. Show that when n tends to infinity, the probability that G has more than $n/2$ triangles tends to 0.

   ☞ The expected number of triangles is less than n^3p^3. By Markov, G has more than $n/2$ triangles with a probability $< \frac{n^3p^3}{n/2} \to 0$.

2. Let $\alpha(G)$ be the size of the largest *independent set* of G (A set of vertices X is *independent* if there is no edge between any two vertices of X in G). Show that $\chi(G) \geq n/\alpha(G)$.

   ☞ By definition of χ, there is a coloring of G with χ colors, which is also a partition of $V(G)$ into subsets such that each subset is independent. Hence, the carnality of each subset is at most α. This implies $\chi\alpha \geq n$.

3. Let $a = 3n^{1-\varepsilon}\ln n$. Show that when n tends to infinity,

$$\mathbb{P}(\alpha(G) < a) \to 1.$$

   Deduce that there exists n and G of size n such that G has at most $n/2$ triangles and $\alpha(G) < a$.

   ☞ α exceeds a with proba at most

$$\binom{n}{a}(1-p)^{\binom{2}{a}} < n^a e^{-p\frac{1}{2}a(a-1)} < n^a n^{-\frac{3}{2}(a-1)} \to 0.$$

4. Let G be such a graph. Let G' be a graph obtained from G by removing a minimum number of of vertices so that G' does not contain any triangle. Show that

$$\chi(G') > \frac{n^\varepsilon}{6\ln n}$$

   and conclude the proof of Mycielski's Theorem.

   ☞

$$\chi > |G'|/\alpha > \frac{n/2}{3n^{1-\varepsilon}\ln n} > \frac{n^\varepsilon}{6\ln n}$$

Exercice 5. *Filtres de Bloom*

[Disclaimer : l'exercice parle de choses vues en cours que vous n'avez en fait pas vu. Mais ça n'a aucune importance.]

Rappelez-vous les tables de hachage vues en cours et reprenons l'exemple de l'interdiction des mots de passe trop simples. On dispose d'un ensemble F de mots de passe interdits, et l'on veut stocker F de manière intelligente pour pouvoir, à chaque fois qu'un utilisateur choisi un nouveau mot de passe, vérifier si ce mot de passe est admissible. Dans le premier exemple du cours (*Chain Hashing*), on cherche à minimiser le temps d'une requête pour savoir si $x \in F$. Dans le deuxième exemple du cours (*Bit Strings/Fingerprints*), on cherche à minimiser l'espace de stockage de F, quitte à ce que certaines requêtes produisent un faux positif (i.e. répond que $x \in F$ alors que $x \notin F$).

On va s'intéresser ici à un troisième exemple appelés *filtre de Bloom* qui permet d'obtenir un meilleur compromis entre espace de stockage et taux de faux positifs. Un filtre de Bloom est un tableau A à n cases, initialement remplies à 0. On dispose de k fonctions de hachage indépendantes $h_1, \ldots, h_k$ à valeurs dans $\{1, \ldots, n\}$. On suppose comme à l'accoutumée pour les fonctions de hachage, que chaque h_i associe à n'importe quel élément de l'univers un nombre choisi uniformément au hasard dans $\{1, \ldots, n\}$. Soit $F = \{f_1, \ldots, f_m\}$ l'ensemble des m mots interdits. L'étape de pré-processing est la suivante : pour chaque $f \in F$, et pour chaque $i \leq k$, on met $A[h_i(f)]$ à 1 (si cette case était déjà à 1, on ne la touche pas). Supposons maintenant que l'on ait une requête du type $s \in ?F$. On répond de la manière suivante : si tous les $A[h_i(s)]$ valent 1 pour $1 \leq i \leq k$, alors on répond $s \in F$. Sinon, on répond $s \notin F$. On vérifie facilement qu'il est impossible d'obtenir un faux-négatif.

1. Soit X le nombre de cases de A dans lesquelles il reste un 0 après le pré-processing. Quelle est l'espérance de X/n ?

☞ Considérons une case donnée $A[\ell]$. A chaque hachage $h_i(f_j)$, la probabilité que $h_i(f_j) = \ell$ (i.e. que $A[\ell]$ passe ou repasse à 1) est $1/n$. Donc la probabilité que $A[\ell] = 0$ après les km hachages est $\left(1 - \frac{1}{n}\right)^{km}$. Cette probabilité correspond à l'espérance de la proportion de cases à 0 (car on multiplie puis on divise par n).

2. Soit $p = e^{-km/n}$. Dans cette question, on suppose pour simplifier que X est égal à pn. Quelle est la probabilité P d'un faux positif ? Comment choisir k pour minimiser P, et qu'obtient-on comme valeur de P ?

☞ Un faux positif se produit pour un mot s si les k hachages $h_i(s)$ (pour $1 \leq i \leq k$) tombent sur des cases contenant toutes un 1. Si i est fixé, la probabilité que $h_i(s)$ tombe sur une case contenant un 0 est $pn/n = p$, dont on obtient au total : $P = (1-p)^k = (1 - e^{-km/n})^k$
Or, comme $p = e^{-km/n}$, on a $\ln p = -km/n$ donc $k = -\ln(p)n/m$ et donc $P = (1-p)^k = e^{k \ln(1-p)} = e^{-\ln(p)\ln(1-p)n/m}$. On voit que cette expression est minimisée pour $p = 1/2 \to$ on remarque donc que l'on obtient une probabilité de faux-positif optimale lorsque la moitié des bits du tableau sont à 1, autrement dit lorsque le filtre de Bloom ressemble à un tableau aléatoire.
On obtient $P = (2^{-\ln 2})^{n/m} \approx (0.61)^{n/m}$.

3. Justifier pourquoi il a semblé raisonnable de supposer, par simplification, que $X = pn$. Plus exactement, utiliser l'approximation de Poisson pour borner $\mathbf{P}\{|X - np| \geq \varepsilon n)\}$, et commenter.

☞ On utilise la question 3 de l'exercice "Approximation de Poisson", qui nous dit que si X_i est la charge réelle de n paniers dans lesquels on a jeté km balles, et si Y_i est la charge dans l'approximation de Poisson (i.e. chaque Y_i est une variable de Poisson de paramètre km/n indépendantes), alors pour toute fonction f à valeurs positives ou nulles

$$\mathbf{E}[f(X_1, \dots, X_n)] \leq e\sqrt{km}\,\mathbf{E}[f(Y_1, \dots, Y_n)].$$

On va prendre $f(x_1, \dots, x_n) = 0$ si le nombre de variables x_i égales à 0 est dans $[np - n\varepsilon, np + n\varepsilon]$ et 1 sinon. la fonction f est bien à valeur positives ou nulles. On rappelle que X est le nombre de variables X_i égales à zéro. On définit de la même manière Y qui est égale au nombre de Y_i valant 0. Alors $f(X_1, \dots, X_n)$ est la fonction indicatrice de l'événement "$|X - np| > n\varepsilon$". Donc $\mathbf{E}[f(X_1, \dots, X_n)] = \mathbf{P}\{|X - np| > n\varepsilon\}$. De même, on a $\mathbf{E}[f(Y_1, \dots, Y_n)] = \mathbf{P}\{|Y - np| > n\varepsilon\}$. Majorons cette dernière probabilité. Comme les Y_i sont indépendantes et de même loi, la variable Y suit une loi binomiale de paramètres $(n, \mathbf{P}\{Y_i = 0\})$. Or $\mathbf{P}\{Y_i = 0\} = e^{-km/n} = p$. Donc $\mathbf{E}[Y] = np$. En utilisant l'inégalité de Chernoff II, on obtient $\mathbf{P}\{|Y - np| > n\varepsilon\} \leq 2e^{-\frac{\varepsilon^2 n}{2p+\varepsilon}}$. Et donc, en revenant aux X_i (en en utilisant le résultat de l'approximation de Poisson), on a

$$\mathbf{P}\{|X - np| > n\varepsilon\} \leq 2e \cdot \sqrt{km} \cdot e^{-\frac{\varepsilon^2 n}{2p+\varepsilon}}.$$

Si on fixe comme à la question précédente $p = 1/2$, alors $km = n\ln(2)$ et la quantité ci-dessus tend vers 0 pour tout choix de $\varepsilon > 0$. Donc l'approximation faite à la question précédente est justifiée.