TD 07 – Convergence de variables aléatoires et (encore) partiel de l'an dernier (corrigé)

Exercice 1. *Suite de bits aléatoires*

On se donne X_i une suite infinie de bits aléatoires non biaisés.

1. Montrer que presque sûrement tout mot fini apparaît dans la suite X_i.

☞ Soit w un mot fini. On définit A_j (pour $j \geq 1$) l'événement "$X_{kj+1} \cdots X_{k(j+1)} = w$". On a $\mathbf{P}\{A_j\} = \frac{1}{2^k}$. Par indépendance des X_i (et lemme de groupement par paquets), les événements A_j sont indépendants. On a alors

$$\begin{aligned}\mathbf{P}\{w \text{ n'apparaît pas dans la suite } X_i\} &\leq \mathbf{P}\{\cap_{1\leq j\leq\infty}\overline{A_j}\} \\ &= \lim_{N\to\infty} \mathbf{P}\{\cap_{1\leq j\leq N}\overline{A_j}\} \\ &= \lim_{N\to\infty}\left(\frac{2^k-1}{2^k}\right)^N \\ &= 0\end{aligned}$$

Donc w apparaît dans la suite X_i avec probabilité 1.

On note maintenant Y_w l'événement "le mot w apparaît dans la suite X_i". On a vu que pour tout w fini, $\mathbf{P}\{Y_w\} = 1$. De plus, il y a un nombre dénombrable de mots finis, i.e. un nombre dénombrables d'événements Y_w. On en déduit que $\mathbf{P}\{\cap_w Y_w\} = 1$ (en passant par le complémentaire c'est plus propre : $\mathbf{P}\{\cup_w \overline{Y_w}\} \leq \sum_w \mathbf{P}\{\overline{Y_w}\} = 0$).

2. En déduire que la presque sûrement tout mot fini apparaît une infinité de fois dans la suite X_i.

☞ Tout mot fini est le sous-mot d'une infinité de mots fini distincts. Donc si une séquence infinie contient tout mot fini, elle contient tout mot fini une infinité de fois (car elle contiendra tous les sur-mots d'un mot fixé, et qu'il y en a une infinité).

Exercice 2. *Algorithme probabiliste pour calculer la médiane*

On étudie un algorithme probabiliste[1] pour déterminer la médiane d'un ensemble $E = \{x_1, \dots, x_n\}$ de n nombres réels en temps $O(n)$. On rappelle que m est une médiane de E si au moins $\lceil n/2 \rceil$ des élements de E sont inférieurs ou égaux à m, et au moins $\lceil n/2 \rceil$ des élements de E sont supérieurs ou égaux à m. Pour simplifier on suppose n impair (ce qui fait que la médiane est unique) et on suppose aussi que les éléments de E sont tous distincts.

Voici comment fonctionne l'algorithme

(a) Soit $(Y_i)_{1\leq i\leq n}$ une suite de v.a. i.i.d. de loi de Bernoulli de paramètre $n^{-1/4}$. On considère le sous-ensemble aléatoire de E défini par $F = \{x_i : Y_i = 1\}$. Si card $F \leq \frac{2}{3}n^{3/4}$ ou card $F \geq 2n^{3/4}$ on répond «`ERREUR 1`».

(b) On trie F et on appelle d le $\lfloor \frac{1}{2}n^{3/4} - \sqrt{n} \rfloor$ème plus petit élément de F, et u le $\lfloor \frac{1}{2}n^{3/4} - \sqrt{n} \rfloor$ème plus grand élément de F.

(c) On détermine le rang de d et de u dans E (l'élément minimal a rang 1, l'élément maximal a rang n), que l'on note respectivement r_d et r_u. Si $r_d > n/2$ ou $r_u < n/2$ on répond «`ERREUR 2`».

(d) On note $G = \{x_i \in E : d < x_i < u\}$. Si card $G \geq 4n^{3/4}$ on répond «`ERREUR 3`».

(e) On trie G et on renvoie le $(\lceil n/2 \rceil - r_d)$ème élement de G.

1. Justifier pourquoi l'algorithme retourne la médiane en temps $O(n)$ lorsqu'il ne répond pas de message d'erreur.

☞ Si aucun message d'erreur n'est renvoyé, l'algorithme s'exécute en temps $O(n)$; en effet la génération des (Y_i) prend un temps $O(n)$, le tri de F et G prend un temps $O(m \log m)$ pour $m = O(n^{3/4})$, et la détermination de r_d, de r_u et de G nécessite $O(n)$ comparaisons. De plus, l'absence de message d'erreur numéro 2 garantit que la médiane est dans l'intervalle $[d, u]$, donc dans G.

1. Remarque : il existe un algorithme déterministe de même performance

2. Montrer que pour $i \in \{1, 2, 3\}$, on a

$$\lim_{n\to\infty} \Pr\left(\text{l'algorithme retourne «}\texttt{ERREUR}\ i\text{»}\right) = 0.$$

Pour simplifier l'analyse et éviter d'écrire des symobles $\lfloor\cdot\rfloor$ ou $\lceil\cdot\rceil$, on pourra supposer implicitement que des nombres tels que $\sqrt{n}$, $\frac{1}{2}n^{3/4}$, ... sont des entiers

☞

1. Pour l'erreur 1 : comme card $F = Y_1 + \cdots + Y_n$ a la loi $B(n, n^{-1/4})$, on a par l'inégalité de Chernoff II

$$\mathbb{P}(\text{card } F \geq 2n^{3/4}) \leq \exp(-n^{3/4}/3), \quad \mathbb{P}(\text{card } F \leq \frac{2}{3}n^{3/4}) \leq \exp(-n^{3/4}/18).$$

2. Pour l'erreur 2 : on note E^- l'ensemble des éléments de E inférieurs où égaux à la médiane, et on remarque que $r_d > n/2$ équivaut à card $(F \cap E^-) < \frac{1}{2}n^{3/4} - \sqrt{n}$. La v.a. card $(F \cap E^-)$ suit la loi $B(\lceil n/2 \rceil, n^{-1/4})$ (notons μ sa moyenne) donc par l'inégalité de Chernoff II

$$\mathbb{P}(\text{card } (F \cap E^-) < \frac{1}{2}n^{3/4} - \sqrt{n}) \leq \mathbb{P}(\text{card } (F \cap E^-) \leq (1 - 2n^{-1/4})\mu) \leq \exp(-\mu\sqrt{n}) \to 0$$

Un argument symétrique traite le cas de $r_u > n/2$ et considérant E^+ l'ensemble des éléments de E supérieurs où égaux à la médiane

3. Pour l'erreur 3 : si card $G \geq 4n^{3/4}$, alors ou bien card $(G \cap E^-) \geq 2n^{3/4}$ ou bien card $(G \cap E^+) \geq 2n^{3/4}$; ces deux événements ayant même probabilité, il suffit de montrer que $\mathbb{P}(\text{card } (G \cap E^-) \geq 2n^{3/4}) \to 0$. On remarque que si card $(G \cap E^-) \geq 2n^{3/4}$, alors $r_d \leq \frac{n}{2} - 2n^{3/4}$ et donc l'ensemble F contient au moins $\frac{1}{2}n^{3/4} - \sqrt{n}$ parmi les $\frac{n}{2} - 2n^{3/4}$ plus petits éléments de E. La probabilité de ce dernier événement est $\mathbb{P}(X \geq (1+\varepsilon)\mathbf{E}[X])$, où $X \sim B(\frac{n}{2} - 2n^{3/4}, n^{-1/4})$ et $\varepsilon = \frac{\sqrt{n}}{n^{3/4}/2 - 2\sqrt{n}} = O(n^{-1/4})$. Une dernière application de l'inégalité de Chernoff II permet de conclure que la probabilité considérée tend vers 0.

Exercice 3. *Approximation de Poisson*

On se place dans le modèle *Balls and Bins* où l'on jette m balles au hasard dans n paniers. Le problème est que les v.a. X_i représentant le nombre de balles dans le i-ème panier ne sont pas indépendantes (intuitivement, car $X_1 + \cdots + X_n = m$). On voudrait approximer le modèle *Balls and Bins* par le modèle *Approximation de Poisson*, dans lequel $Y_1, \ldots Y_n$ sont des variables aléatoires indépendantes qui suivent chacune une loi de Poisson de moyenne $\mu = m/n$ (la variable Y_i est donc pensée pour être la version simplifiée de X_i).

1. Montrer que $Y = \sum_{i=1}^n Y_i$ suit une loi de Poisson dont on précisera le paramètre.

   ☞ Facile par récurrence sur n : la somme de v.a. de Poisson indépendantes Y_i de paramètre μ_i est une v.a. de Poisson de param $\sum_i \mu_i$. Voir par exemple Lemma 5.2 dans MU, p. 96. ($\mathbf{P}\{Y_i = j\} = e^{-\mu}\mu^j/j!$)

2. Montrer que la distribution de $(Y_1, \ldots, Y_n)$ conditionnée au fait que $Y = m$ est la même que la distribution de $(X_1, \ldots, X_n)$.

   Note : on peut en fait obtenir un résultat légèrement plus général. Si $(X_1, \ldots, X_n)$ représente la charge de n paniers après avoir lancé au hasard k balles, et que les Y_i sont n v.a. indépendantes suivant chacune une loi de Poisson de paramètre m/n, alors la distribution de $(Y_1, \ldots, Y_n)$ conditionnée au fait que $Y = k$ est la même que la distribution de $(X_1, \ldots, X_n)$, indépendamment de la valeur de m.

   ☞ voir Thm 5.6 dans MU 5.4, p. 100

   Fixing $k_1 \ldots k_n$ summing to k, the probability that all $X_i = k_i$ is

$$\frac{k!}{k_1! \ldots k_n!} \frac{1}{n^k}$$

   Now, as Y_i are independent and follow a Poisson law, the probability that all $Y_i = k_i$ knowing that their sum $Y = k$ is

$$\frac{\prod_i e^{-m/n}\mu^{k_i}/k_i!}{e^{-m}m^k/k!} = \frac{k!}{k_1! \ldots k_n!} \frac{1}{n^k}$$

3. Soit f une fonction sur n variables, à valeurs réelles positives ou nulles. Prouver que

$$\mathbf{E}\left[f(X_1, \ldots, X_n)\right] \leq e\sqrt{m}\mathbf{E}\left[f(Y_1, \ldots, Y_n)\right] .$$

   On pourra prouver comme étape intermédiaire que $m! < e\sqrt{m}\left(\frac{m}{e}\right)^m$.

   ☞ voir Thm 5.7 dans MU 5.4, p. 101

   Inequality : Since $\ln x$ is concave we have

$$\int_{i-1}^{i} \ln x dx \geq \frac{1}{2}\left(\ln(i-1) + \ln i\right)$$

Then

$$\int_{1}^{n} \ln = n \ln n - 1 \geq \ln(n!) - \frac{1}{2} \ln n$$

Then exponentiate.
Then,

$$\mathbf{E}\left[f(Y_1, \ldots, Y_n)\right] \geq \mathbf{E}\left[f(X_1, \ldots, X_n)\right] \mathbf{P}\left\{Y = m\right\}$$

And use the inequality

4. En déduire le corollaire suivant : soit $\mathcal{E}$ un événement qui dépend de la charge des paniers. Supposons que $\mathcal{E}$ arrive avec probabilité p dans l'*Approximation de Poisson*, c'est-à-dire si la charge des paniers est $(Y_1, \ldots, Y_n)$. Alors $\mathcal{E}$ arrive avec probabilité au plus $pe\sqrt{m}$ dans le modèle *Balls and Bins*, c'est-à-dire si la charge des paniers est $(X_1, \ldots, X_n)$.

   ☞ Voir corollaire 5.9, MU p. 102

5. *Application :* On jette n balles dans n paniers selon le modèle *Balls and Bins*. Montrer qu'avec probabilité au moins $1 - 1/n$ (pour n assez grand), la charge maximale est $\geq \ln n / \ln \ln n$.

   ☞ Lemma 5.12 MU p. 103
   Poisson with $m = n$: probability that no bin exceed charge M is $(1 - \frac{1}{eM!})^n \leq e^{-n/(eM!)}$

Exercice 4. *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. $\mathbf{P}\left\{X_n = 1\right\} = 1 - p_n$ et $\mathbf{P}\left\{X_n = 0\right\} = p_n$).

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

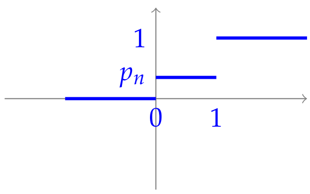

Figure 1 – Fonction de répartition de X_n

   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.

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 $\mathbf{P}\left\{|X_n - X_{n+1}| \geq \varepsilon\right\} = \mathbf{P}\left\{X_n \neq X_{n+1}\right\} = 1/2$ 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 $\mathbf{P}\left\{|X_n - X| \geq \varepsilon\right\} \to 0$. Par inégalité triangulaire, cela implique en particulier que $\mathbf{P}\left\{|X_n - X_{n+1}| \geq 2\varepsilon\right\} \to 0$. Prenons $2\varepsilon = 1/2$, on a alors $\mathbf{P}\left\{|X_n - X_{n+1}| \geq 2\varepsilon\right\} = \mathbf{P}\left\{X_n \neq X_{n+1}\right\} \geq p_n$. En effet, une fois X_{n+1} fixé, on a $\mathbf{P}\left\{X_n \neq X_{n+1}\right\} = p_n$ si $X_{n+1} = 1$ et $\mathbf{P}\left\{X_n \neq X_{n+1}\right\} = 1 - p_n$ 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 $\varepsilon > 0$

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

   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 suffisante (nécessaire ce sera la semaine prochaine) 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 SecondBorelCantelli), 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.