TD 06 – Inégalité de Chernoff (Suite) (corrigé)

Exercice 1. *Sondage*

Nous sommes en période de campagne BDE à l'ENS de Lyon et nous voulons faire un sondage d'opinion pour estimer la proportion p de la population normalienne souhaitant voter pour la Zicliste. Supposons que l'on interroge n personnes choisies uniformément et indépendamment au hasard, et que chacune d'elle réponde par "Oui, je souhaite voter pour eux" ou "Non, je ne suis souhaite pas voter pour eux". Étant donnés $\theta > 0$ et $0 < \delta < 1$, on souhaite trouver une estimation $\overline{X}$ de p telle que

$$\mathbf{P}\left\{|\overline{X} - p| \leq \theta\right\} > 1 - \delta\,.$$

Par exemple pour $1 - \delta = 0.95$, on pourra ainsi dire que le sondage a une précision de θ à 95%.

1. Que choisir comme estimation $\overline{X}$ de p ?

   ☞ Soit $X_i = 1$ si la $i$ème personne interrogée est d'accord, et 0 zéro. On pose ensuite $X = \sum_{i=1}^{n} X_i$ et $\overline{X} = 1/n \cdot X$. Alors $\mathbf{E}[X] = pn$ et $\mathbf{E}\left[\overline{X}\right] = p$.

2. Combien de personnes doit-on interroger pour que l'estimation $\overline{X}$ vérifie nos conditions ? Autrement dit, donner une borne inférieure sur n en termes de θ et δ. On remarquera que cette borne ne dépend pas de la taille de la population totale.

   ☞ On utilise la borne de Chernoff 'two-sided' sur X :

$$\mathbf{P}\left\{|\overline{X} - p| \geq \varepsilon p\right\} = \mathbf{P}\left\{|X - pn| \geq \varepsilon pn\right\} \leq 2\exp\left(-\frac{\varepsilon^2}{2+\varepsilon} \cdot pn\right).$$

   Ensuite on pose $\varepsilon = \theta/p$ pour avoir $\varepsilon p = \theta$, et en ré-injectant :

$$\mathbf{P}\left\{|\overline{X} - p| \geq \varepsilon p\right\} \leq 2\exp\left(-\frac{\theta^2/p^2}{2+\theta/p} \cdot pn\right) = 2\exp\left(-\frac{\theta^2}{2p+\theta} \cdot n\right).$$

   Essayons maintenant de borner ceci par δ. On a :

$$\frac{\theta^2}{2p+\theta} \geq \frac{\theta^2}{2+\delta}$$

   et donc

$$\mathbf{P}\left\{|\overline{X} - p| \geq \theta\right\} \leq 2\exp\left(-\frac{\theta^2}{2+\theta} \cdot n\right)\,.$$

   Enfin :

$$\begin{aligned}\delta \geq 2\exp\left(-\frac{\theta^2}{2+\theta} \cdot n\right) &\Leftrightarrow \exp\left(\frac{\theta^2}{2+\theta} n\right) \geq \frac{2}{\delta}\\ &\Leftrightarrow \frac{\theta^2}{2+\theta} n \geq \ln\frac{2}{\delta}\\ &\Leftrightarrow n \geq \frac{2+\theta}{\theta^2}\ln\frac{2}{\delta}\end{aligned}$$

3. Calculer la valeur de n obtenue grâce à votre borne pour les paramètres $\theta = 0.2$ et $1 - \delta = 95\%$.

   ☞ On obtient $n \geq 203$.

Exercice 2. *blackjack*

Vous êtes le croupier dans une partie de blackjack au Gala de l'ENS de Lyon, et vous soupçonnez un joueur de tricher en comptant les cartes. En effet, sur les quelques premières mains que vous venez de le voir jouer, il gagne 55% du temps (alors, que, sans tricher, la probabilité de gagner un main est 1/2). Cependant, vous voulez attendre d'avoir un peu plus de certitude avant de démasquer le joueur.

1. On suppose que le joueur continue de gagner 55% du temps. Combien de mains devez-vous le laisser jouer avant d'être sûr à 90% qu'il triche ?

☞ On pose $\varepsilon = 0.05$ et $\delta = 0.1$. Supposons que le joueur joue n mains et on note $X_i = 1$ si le joueur a gagné la i-ème main, 0 sinon. On pose $X = \sum i = 1^n X_i$. On va dénoncer le tricheur donc on veut se tromper avec proba au plus δ. On se trompe si le joueur ne trichait pas, auquel cas X_i vaut 1 avec proba 1/2, et donc $\mathbf{E}[X] = n/2$. On veut donc que, avec ces hypothèses, $\mathbf{P}\{X/n - 1/2 \geq \varepsilon\} \leq \delta$. Or $\mathbf{P}\{X/n - 1/2 \geq \varepsilon\} = \mathbf{P}\{X - \mathbf{E}[X] \geq n\varepsilon\}$ donc on peut appliquer Hoeffding et on obtient :

$$\mathbf{P}\{X/n - 1/2 \geq \varepsilon\} = \mathbf{P}\{X - \mathbf{E}[X] \geq n\varepsilon\} \leq e^{-\frac{2\varepsilon^2 n^2}{n}} = e^{-2\varepsilon^2 n} .$$

On veut que ceci soit $\leq \delta$ donc il suffit de prendre $n \geq \frac{1}{2\varepsilon^2}\log(1/\delta) \approx 460$ mains.

(Attention, correction à vérifier)

Exercice 3. *Sous-gaussiennes*

Une variable aléatoire X est dite *sous-gaussienne* de paramètre σ si elle vérifie l'inégalité suivante :

$$\forall \lambda \in \mathbb{R}, \mathbf{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2 \sigma^2}{2}} .$$

Soit X_1 et X_2 deux variables aléatoires indépendantes, respectivement sous-gaussiennes de paramètre σ_1 et σ_2.

1. Montrez que $X_1 + X_2$ est sous-gaussienne de paramètre $\sqrt{\sigma_1^2 + \sigma_2^2}$. Montrez que cX_1 est $|c|\sigma_1$-sous-gaussienne pour tout $c \in \mathbb{R}$.

☞ On a, par indépendance : $\mathbf{E}\left[e^{\lambda(X_1+X_2)}\right] \leq \mathbf{E}\left[e^{\lambda X_1}\right]\mathbf{E}\left[e^{\lambda X_2}\right] \leq e^{\frac{\lambda^2\sigma_1^2}{2}} e^{\frac{\lambda^2\sigma_2^2}{2}} = e^{\frac{\lambda^2(\sigma_1^2+\sigma_2^2)}{2}}$. Donc $X_1 + X_2$ est $\sqrt{\sigma_1^2+\sigma_2^2}$-sous-gaussienne. De même $\mathbf{E}\left[e^{\lambda c X_1}\right] \leq e^{\frac{\lambda^2 c^2 \sigma_1^2}{2}}$ donc cX_1 est $|c|\sigma_1$-sous-gaussienne.

2. Montrez que si X est une variable aléatoire σ-sous-gaussienne, alors pour tout $t \geq 0$,

$$\mathbf{P}\{X \geq t\} \leq e^{-\frac{t^2}{2\sigma^2}}$$

.

☞ On a $\mathbf{P}\{X \geq t\} = \mathbf{P}\left\{e^{\lambda X} \geq e^{\lambda t}\right\} \leq \frac{\mathbf{E}\left[e^{\lambda X}\right]}{e^{\lambda t}} \leq e^{\frac{\lambda^2\sigma^2}{2} - \lambda t}$. On choisit $\lambda = t/\sigma^2$ et on obtient le résultat.

3. Soit μ un réel et $X_1, \ldots, X_n$ des variables aléatoires telles que les $X_i - \mu$ sont indépendantes et σ-sous-gaussiennes. Soit $\delta \in [0, 1]$. Montrez qu'avec probabilité au moins $1 - \delta$, on a

$$\mu \leq \hat{\mu} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}} ,$$

où $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$.

☞ En utilisant le résultat de la question 1, $\sum_n X_i - \mu$ est $\sigma\sqrt{n}$-sous-gaussienne donc, $\hat{\mu} - \mu = \frac{1}{n}\sum_n X_i - \mu$ est $\frac{\sigma\sqrt{n}}{n} = \sigma/\sqrt{n}$-sous-gaussienne. On applique la question précédente pour obtenir

$$\mathbf{P}\{\hat{\mu} - \mu \geq t\} \leq e^{-\frac{nt^2}{2\sigma^2}}$$

. On choisit $t = \sqrt{\frac{\sigma^2\log(1/\delta)}{2n}}$ pour obtenir le résultat.

Exercice 4. *Algorithme probabiliste pour calculer la médiane*

On étudie un algorithme probabiliste[1] pour déterminer la médiane d'un ensemble $E = \{x_1, \ldots, x_n\}$ 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`».

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

(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.

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

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

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

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

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 $\mathsf{P}(\text{card } (G \cap E^-) \geq 2n^{3/4}) \to 0$. 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 $\mathsf{P}(X \geq (1+\varepsilon)\mathbf{E}[X])$, 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 5. *Graphe Aléatoire Bipartite*

Soit $0 < p < 1$ et $n \in \mathbb{N}^*$. On définit un graphe aléatoire non orienté $H_{2n,p}$ de la manière suivante. On se donne une famille $\{X_{i,j} \,:\, 1 \leq i \leq n,\ n+1 \leq j \leq 2n\}$ de v.a. i.i.d. de loi de Bernoulli de paramètre p. On pose alors $H_{2n,p} = (V, E)$, avec $V = \{1, \ldots, 2n\}$ et

$$E = \{(i,j) \,:\, X_{i,j} = 1\} \subset \{1, \ldots, n\} \times \{n+1, \ldots, 2n\}.$$

1. Quelle est la loi du nombre d'arêtes de $H_{2n,p}$?

☞ Le nombre d'arêtes de $H_{2n,p}$ suit la loi $B(n^2, p)$.

2. Quelle est l'espérance du nombre de sommets isolés de $H_{2n,p}$?

☞ Soit N le nombre de sommets isolés. Si A_i est l'événement « le sommet i est isolé », on a par linéarité de l'espérance, on a $\mathsf{E}[N] = \sum \mathsf{P}(A_i) = 2n(1-p)^n$.

3. Dans cette question on pose $p = c\log(n)/n$ pour un nombre réel $c > 0$.

1. Montrer que si $c > 1$, alors

$$\lim_{n\to\infty} \mathsf{P}(H_{2n,p} \text{ a un sommet isolé}) = 0.$$

2. Montrer que si $c < 1$, alors

$$\lim_{n\to\infty} \mathsf{P}(H_{2n,p} \text{ a un sommet isolé}) = 1.$$

☞

1. Si $c > 1$, on a $\mathsf{E}[N] = 2n\exp(n\log(1 - \frac{c\log n}{n})) \to 0$ et donc $\mathsf{P}(N \geq 1) \leq \mathsf{E}[N] \to 0$.
2. Si $c < 1$, on calcule

$$\mathsf{E}[N^2] = \sum_{i,j=1}^{2n} \mathsf{P}(A_i \cap A_j) = 2n(1-p)^n + 2n(n-1)(1-p)^{2n} + 2n^2(1-p)^{2n-1}$$

d'où il vient que $\mathsf{E}[N^2]/\mathsf{E}[N]^2$ tend vers 1. On utilise l'inégalité de Tchebychev pour conclure que

$$\mathsf{P}(N=0) = \mathsf{P}(\mathsf{E}[N] - N \geq \mathsf{E}[N]) \leq \mathsf{P}(|N - \mathsf{E}[N]| \geq \mathsf{E}[N]) \leq \frac{\mathbf{Var}\,[N]}{\mathsf{E}[N]^2} = \frac{\mathsf{E}[N^2]}{\mathsf{E}[N]^2} - 1 \to 0.$$

4. Dans cette question on pose $p = 1/2$. Montrer qu'il existe une constante $C > 0$ telle que

$$\lim_{n\to\infty} \mathsf{P}\left(\text{tous les sommets de } H_{2n,p} \text{ ont un degré inférieur à } \frac{n}{2} + C\sqrt{n\log n}\right) = 1.$$

☞ Le degré d_i du sommet i suit la loi $B(n, 1/2)$. Par l'inégalité de Chernoff I, on a donc

$$\mathsf{P}(d_i \geq \frac{n}{2} + a) \leq \exp(-2a^2/n).$$

Ainsi, par la borne de l'union,

$$\mathsf{P}(\max_i d_i \geq \frac{n}{2} + a) \leq 2n\exp(-2a^2/n).$$

Cette quantité tend vers 0 si $a = C\sqrt{n\log n}$ avec $2C^2 > 1$.