TD 05 – Inégalité de Chernoff: applications (corrigé)

Exercice 1. *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 $\mathcal{O}(n)$. 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 $(Y_i)_{1\le i\le n}$ 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 $F \le \frac{2}{3}n^{3/4}$ ou card $F \ge 2n^{3/4}$ 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 $G \ge 4n^{3/4}$ 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 $\mathcal{O}(n)$ 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

$$\mathrm{P}(\text{card } F \ge 2n^{3/4}) \le \exp(-n^{3/4}/3), \quad \mathrm{P}(\text{card } F \le \frac{2}{3}n^{3/4}) \le \exp(-n^{3/4}/18).$$

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

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

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 $G \ge 4n^{3/4}$, alors ou bien card $(G \cap E^-) \ge 2n^{3/4}$ ou bien card $(G \cap E^+) \ge 2n^{3/4}$; ces deux événements ayant même probabilité, il suffit de montrer que $\mathrm{P}(\text{card } (G \cap E^-) \ge 2n^{3/4}) \to 0$. On remarque que si card $(G \cap E^-) \ge 2n^{3/4}$, alors $r_d \le \frac{n}{2} - 2n^{3/4}$ 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 $\mathrm{P}(X \ge (1+\varepsilon)\mathbf{E}[X])$, 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 2. *Estimer l'intersection avec un rectangle*

Soit $P \subset \mathbb{Z}^2$ un ensemble de n coordoonées. On veut répondre (rapidement) à des questions du type :

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

$\mathcal{Q}_r$: « Quel est la proportion de points de P qui se situent dans le rectangle $r = [a_1, b_1] \times [a_2, b_2]$? »

Pour n'importe quel rectangle r, on note $\mathbf{r}[P] = \frac{|P \cap r|}{n}$ la proportion recherchée.
Pour estimer $\mathbf{r}[P]$ de façon efficace, on considère $S \subseteq P$ un sous-ensemble de taille m de P (choisi aléatoirement) et on renvoie $\mathbf{r}[S] = \frac{|S \cap r|}{m}$.
On dit que S est une ε-*approximation* si pour tout r, on a $|\mathbf{r}[P] - \mathbf{r}[S]| \leq \varepsilon$.

1. Avec quelle taille m obtient-on une ε-approximation avec une probabilité $1 - \delta$?
Indication : on peut considérer un ensemble S dont l'espérance de la taille est m, *plutôt que de taille exactement m.*

☞ On va prendre un sample S dont *l'espérance de la taille* est m (plutôt que taille exactement m).
Pour $p \in P$, soit X_p une variable aléatoire de Bernouilli de paramètre m/n, et l'on définit S par la relation suivante : si $X_p = 1$ alors $p \in S$, et si $X_p = 0$ alors $p \notin S$. Fixons un rectangle r et soit $X(r) = \sum_{p \in R} X_p = |S \cap R|$ de telle sorte que $X(r)/m$ soit notre estimateur. Alors $\mathbf{E}[X(r)] = \sum_{p \in R} \mathbf{P}\{p \in S\} = \sum_{p \in R} \mathbf{P}\{m/n\} = mr[P]$. On peut donc appliquer Chernoff à $X(r)$ car :

$$\mathbf{P}\{|X(r)/m - r[P]| \geq \varepsilon\} = \mathbf{P}\{|X(r) - mr[P]| \geq \varepsilon m\} = \mathbf{P}\{|X(r) - \mathbf{E}[X(r)]| \geq \varepsilon/r[P] \cdot \mathbf{E}[X(r)]\} \leq 2e^{-\frac{\varepsilon'^2}{2+\varepsilon'}\mathbf{E}[X(r)]} .$$

avec $\varepsilon' = \varepsilon/r[P]$, ce qui donne (en utilisant $r[P] \leq 1$ pour la dernière inégalité) :

$$2e^{-\frac{\varepsilon'^2}{2+\varepsilon'}\mathbf{E}[X(r)]} \leq 2e^{-\frac{\varepsilon^2}{2r[P]+\varepsilon}m} \leq 2e^{-\frac{\varepsilon^2}{2+\varepsilon}m}.$$

Cette inégalité est vraie pour un rectangle r fixé, mais nous avons besoin d'une Union-Bound sur tous les rectangles. Or, il y a une infinité de rectangles possibles dans $\mathbb{Z}^2$, donc nous devons être un peu plus malin. Il faut remarquer que si r et r' sont des rectangles pour lesquels $P \cap r = P \cap r'$, alors $r[P] = r'[P]$ et l'estimation sera la même, donc l'erreur sur l'un sera exactement la même que l'erreur sur l'autre. En d'autre termes, on veut trouver un certain nombre de rectangle $r_1, r_2, \ldots, r_k$ tels que pour tout rectangle r de $\mathbb{Z}^2$, il existe i tel que $P \cap r = P \cap r_i$. Ainsi,

$$\mathbf{P}\{\exists r \text{ s.t. } |r[P] - X(r)| \geq \varepsilon\} \leq \sum_{i=1}^{k} \mathbf{P}\{|r_i[P] - X(r_i)| \geq \varepsilon\} \leq k2e^{-\frac{\varepsilon^2}{2+\varepsilon}m} .$$

Montrons maintenant qu'on peut obtenir $k = n^4$: pour chaque 4-uplet des points de P $((x_1, y_1)(x_2, y_2), (x_3, y_3), (x_4, y_4))$, on définit un rectangle $r_i = [x_1, x_2] \times [y_3 y_4]$. Cet ensemble de n^4 rectangles a bien la propriété demandée car si r est un rectangle quelconque, on peut "pousser" sa limite verticale gauche le plus à droite possible jusqu'à rencontrer un point de P, auquel cas on s'arrête de "pousser". On fait de même pour les quatre côtés du rectangle (on "pousse" vers l'intérieur jusqu'à rencontrer un point de P), et on tombe sur un r_i pour lequel $r \cap P = r_i \cap P$.
En résumé, nous voulons m tel que

$$n^4 2e^{-\frac{\varepsilon^2}{2+\varepsilon}m} \leq \delta ,$$

ce qui est possible pour

$$m \geq \frac{2+\varepsilon}{\varepsilon^2}(4 \ln n + \ln 2 - \ln \delta) = \Omega(\ln n) .$$

Exercice 3. *Interrupteurs*

1. Montrer qu'il existe une constante $\gamma > 0$ rendant l'énoncé suivant vrai :

   « Si une v.a. positive X vérifie $\mathbf{E}[X] = 1$ et $\mathbf{E}[X^2] \leq 3$, alors $\mathbf{P}\{X \geq 1/4\} \geq \gamma$. »

   Indication : définir la variable aléatoire $Y = \mathbf{1}_{X \geq 1/4}$ et se ramener à l'inégalité de Cauchy-Schwarz. $\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$

   ☞ On écrit

   $$1 = \mathsf{E}[X] = \mathsf{E}[X\mathbf{1}_{X<1/4}] + \mathsf{E}[X\mathbf{1}_{X\geq 1/4}] \leq \frac{1}{4} + \mathsf{E}[X\mathbf{1}_{X\geq 1/4}].$$

   Par l'inégalité de Cauchy–Schwarz, $\mathsf{E}[X\mathbf{1}_{X\geq 1/4}] \leq \sqrt{\mathsf{E}[X^2]\mathsf{P}(X \geq 1/4)} \leq \sqrt{3}\sqrt{\mathsf{P}(X \geq 1/4)}$. On obtient la minoration voulue pour $\gamma = 3/16$.

2. Soient $(X_1, \ldots, X_n)$ des v.a. i.i.d. vérifiant $\mathbf{P}\{X_i = 1\} = \mathbf{P}\{X_i = -1\} = \frac{1}{2}$.
   On pose $Y = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$. Calculer $\mathbf{E}[Y^2]$ et $\mathbf{E}[Y^4]$ et en déduire que :

   $$\mathbf{E}[|X_1 + \cdots + X_n|] \geq \frac{\gamma}{2}\sqrt{n}.$$

   ☞ On a $\mathsf{E}[Y^2] = \frac{1}{n} \cdot \mathbf{Var}[Y] = \frac{1}{n} \cdot \sum_i \mathbf{Var}[X_i] = 1$ (par inédependance). On a ensuite

   $$\mathsf{E}[Y^4] = \frac{1}{n^2} \cdot \sum_{i,j,k,l=1}^{n} \mathsf{E}[X_i X_j X_k X_l].$$

   L'indépendance des X_i et le fait que $\mathsf{E}[X_i] = 0$ implique $\mathsf{E}[X_i X_j X_k X_l] = 0$ dès qu'un indice apparaît une unique fois parmi $\{i, j, k, l\}$. Les seuls termes non nuls sont ceux où $i = j = k = l$ ou $i = j \neq k = l$ ou $i = k \neq j = l$ ou $i = l \neq j = k$. On a donc

   $$\mathsf{E}[Y^4] = 1/n^2(n + 3n(n-1)) = 3 - 2/n \leq 3.$$

   On applique la question précédente à $X = Y^2$, d'où $\mathsf{P}(Y^2 \geq 1/4) = \mathsf{P}(|X_1 + \cdots + X_n| \geq \frac{\sqrt{n}}{2}) \geq \gamma$. Enfin,

   $$\mathsf{E}[|X_1 + \cdots + X_n|] \geq \frac{\sqrt{n}}{2}\mathsf{P}\left(|X_1 + \cdots + X_n| \geq \frac{\sqrt{n}}{2}\right) \geq \frac{\gamma\sqrt{n}}{2}.$$

On considère une grille $n \times n$ d'ampoules ainsi que 3 séries d'interrupteurs : des interrupteurs $a = (a_{ij})_{1\leq i,j\leq n}$ associés à chaque ampoule, des interrupteurs $b = (b_i)_{1\leq i\leq n}$ associés à chaque ligne et des interrupteurs $c = (c_j)_{1\leq j\leq n}$ associés à chaque colonne. Chaque interrupteur prend la valeur -1 ou 1. L'ampoule en position (i, j) est allumée si et seulement si $a_{ij} \times b_i \times c_j = 1$. On considère la quantité

$$\mathbf{F}(a, b, c) = \sum_{i,j=1}^{n} a_{ij} b_i c_j$$

qui est le nombre d'ampoules allumées moins le nombre d'ampoules éteintes.
Deux joueuses jouent au jeu suivant :

1. la joueuse 1 choisit la position des interrupteurs (a_{ij}),
2. puis la joueuse 2 choisit la position des interrupteurs (b_i) et (c_j).

La joueuse 1 veut minimiser $\mathbf{F}(a, b, c)$ et la joueuse 2 veut le maximiser. On considère donc :

$$\mathbf{V}(n) = \min_{a\in\{-1,1\}^{n\times n}} \max_{b,c\in\{-1,1\}^n} \mathbf{F}(a, b, c).$$

3. Montrer que $\mathbf{V}(n) = \mathcal{O}(n^{3/2})$ en considérant le cas où la joueuse 1 joue au hasard.

   ☞ Soit $(a_{ij})_{1\leq i,j\leq n}$ des v.a. i.i.d. de loi uniforme sur $\{-1, 1\}$. Quel que soit le choix de b et c, on a

   $$\mathsf{P}(F(a, b, c) \geq t) \leq \exp(-t^2/2n^2)$$

   par l'inégalité de Chernoff (en effet, $F(a, b, c)$ est la somme de n^2 v.a. de loi uniforme sur $\{-1, 1\}$). Par la borne de l'union,

   $$\mathsf{P}(\max_{b,c} F(a, b, c) \geq t) \leq 4^n \exp(-t^2/2n^2).$$

   Lorsque $t > \sqrt{2n^3 \log 4}$, cette probabilité est < 1 et donc $\mathsf{P}(\max_{b,c} F(a, b, c) < t) > 0$: il existe donc un choix de a tel que $\max_{b,c} F(a, b, c) < t$, d'où $V(n) = O(n^{3/2})$.

4. La joueuse 2 applique la stratégie suivante : elle choisit b au hasard, puis ensuite choisit c de façon à allumer le maximum de lampes. Estimer le nombre moyen de lampes allumées par cette stratégie *(indication : utiliser la question 2)* et en déduire que $\mathbf{V}(n) = \Omega(n^{3/2})$.

☞ Fixons $a = (a_{i,j})$ et choisissons (b_i) i.i.d. de loi uniforme sur $\{-1, 1\}$. On a alors

$$\max_c F(a,b,c) = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} b_j \right|.$$

En utilisant la linéarité de l'espérance, le fait que $(b_j)_j$ et $(a_{ij}b_j)_j$ ont même loi et la question I.2, il vient

$$\mathsf{E} \max_c F(a,b,c) = n\mathsf{E} \left| \sum_{j=1}^{n} b_j \right| \geq \frac{n^{3/2}\gamma}{2}.$$

En particulier, pour tout choix de a, il existe b tel que $\max_c F(a,b,c) \geq \frac{n^{3/2}\gamma}{2}$.

Exercice 4. *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 $\{X_{i,j} \mid 1 \leq i \leq n,\ n+1 \leq j \leq 2n\}$ 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

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

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 $c > 0$.

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

   $$\lim_{n\to\infty} \mathbf{P}\left\{H_{2n,p} \text{ a un sommet isolé}\right\} = 0.$$

   ii. Montrer que si $c < 1$, alors :

   $$\lim_{n\to\infty} \mathbf{P}\left\{H_{2n,p} \text{ a un sommet isolé}\right\} = 1.$$

☞

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

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

☞ 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 $2C^2 > 1$.