TD 04 – Inégalités de Chernoff (1) (corrigé)

Exercice 1. *Comparer les bornes*

On a un dé équilibré à 6 côtés. On lance le dé n fois et on note X le nombre de fois où le dé tombe sur 6. Soit q la probabilité de l'événement $X \geq n/4$.

1. Comparer les bornes supérieures que l'on obtient en utilisant l'inégalité de Markov, l'inégalité de Chebyshev, et une borne de Chernoff (la plus adaptée).

☞ La variable X est une somme de n variables de Bernouilli de paramètre $1/6$. On a donc $\mathbf{E}[X] = \frac{n}{6}$ et $\mathbf{Var}[X] = \frac{5n}{36}$. On obtient :

Markov : $\mathbf{P}\{X \geq n/4\} \leq \frac{\mathbf{E}[X]}{n/4} = 2/3$.

Chebyshev : $\mathbf{P}\left\{X \geq \frac{n}{4}\right\} = \mathbf{P}\left\{X - \frac{n}{6} \geq \frac{n}{12}\right\} \leq \frac{\mathbf{Var}[X]}{(n/12)^2} = \frac{20}{n}$.

Chernoff (en utilisant la variante #2 où les X_i sont des variable aléatoires indépendantes à valeurs dans $[0,1]$: $\mathbf{P}\{X \geq (1+\varepsilon)\mathbf{E}[X]\} \leq \exp\left(-\frac{\varepsilon^2}{2+\varepsilon}\mathbf{E}[X]\right)$) :

$$\mathbf{P}\left\{X \geq \frac{n}{4}\right\} = \mathbf{P}\left\{X \geq \left(1+\frac{1}{2}\right)\mathbf{E}[X]\right\} \leq \exp\left(-\frac{n}{60}\right) .$$

Exercice 2. *Chernoff Interval*

— Soit X une variable aléatoire quelconque avec $0 \leq X \leq 1$ et $\mathbb{E}[X] = p$. Considérons la variable aléatoire $Y \in \{0,1\}$ telle que $\Pr(Y=1) = p$.Montrer que pour tout $\lambda > 0$,

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \mathbb{E}\left[e^{\lambda Y}\right].$$

Indice : on pourra utiliser la convexité de la fonction exponentielle.

— En utilisant ce fait, montrer que la borne de Chernoff vue en cours reste valable si l'on remplace l'hypothèse $X_i \in \{0,1\}$ par $X_i \in [0,1]$.

☞ La fonction $x \to e^{\lambda x}$ est convexe donc pour $x \in [0,1]$, $e^{\lambda x} \leq (1-x)e^0 + xe^\lambda = (1-x) + xe^\lambda$.
En particulier,

$$\mathbf{E}\left[e^{\lambda X}\right] \leq \mathbf{E}\left[(1-X) + Xe^\lambda\right] = (1-p) + pe^\lambda = \mathbf{E}\left[e^{\lambda Y}\right] .$$

Maintenant, on a tous les outils pour reprendre la preuve de la borne de Chernoff : soit X_i des variables aléatoires indépendantes à valeurs dans $[0,1]$ avec $\mathbf{E}[X_i] = p_i$. Alors

$$\mathbf{E}\left[e^{\lambda X_i}\right] \leq (1-p_i) + p_i e^\lambda = 1 + p_i(e^\lambda - 1) \leq e^{p_i(e^\lambda - 1)} .$$

Soit $X = \sum X_i$ et $\mu = \mathbf{E}[X] = \sum_{i=1}^n p_i$. Alors de même que dans la preuve originale,

$$\mathbf{E}\left[e^{\lambda X}\right] = \prod_i \mathbf{E}\left[e^{\lambda X_i}\right] \leq \prod_i e^{p_i(e^\lambda - 1)} = e^{\mu(e^\lambda - 1)}$$

On applique Markov à $e^{\lambda X}$:

$$\begin{aligned}\mathbf{P}\{X \geq (1+\delta)\mu\} &= \mathbf{P}\left\{e^{\lambda X} \geq e^{\lambda(1+\delta)\mu}\right\} \\ &\leq \frac{\mathbf{E}\left[e^{\lambda X}\right]}{e^{\lambda(1+\delta)\mu}} \\ &\leq \frac{e^{\mu(e^\lambda - 1)}}{e^{\lambda(1+\delta)\mu}}\end{aligned}$$

En posant $\lambda = \ln(1+\delta) > 0$, on obtient

$$\mathbf{P}\{X \geq (1+\delta)\mu\} \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu .$$

Exercice 3. *Calcul de 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 $\mathcal{O}(n)$. On rappelle que m est une médiane de E si au moins $\lceil n/2 \rceil$ des

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

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

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

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

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

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 $\mathsf{P}(\text{card } (G \cap E^-) \ge 2n^{3/4}) \to 0$. 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 $\mathsf{P}(X \ge (1+\varepsilon)\mathbf{E}[X])$, 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 4. *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 $\mathbf{E}[X^2] \le 3$, alors $\mathbf{P}\{X \ge 1/4\} \ge \gamma$. »

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

☞ On écrit

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

Par l'inégalité de Cauchy–Schwarz, $\mathsf{E}[X\mathbf{1}_{X\ge 1/4}] \le \sqrt{\mathsf{E}[X^2]\mathsf{P}(X \ge 1/4)} \le \sqrt{3}\sqrt{\mathsf{P}(X \ge 1/4)}$. 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 $\max_c F(a, b, c) \geq \frac{n^{3/2}\gamma}{2}$.