TD 05 – Inégalité de Chernoff: applications (corrigé)

Exercice 1. *Tester la pièce*

1. On considère une pièce qui tombe sur pile avec probabilité p. Combien de fois doit-on lancer la pièce pour déterminer p à ± 0.1 avec probabilité au moins 0.9 ?

☞ Soit $X_i = 1$ si le i-ième lancer donne pile, 0 sinon. Soit $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$. On veut approximer p par $\overline{X_n}$, en choisissant soigneusement n. Notre but est donc d'obtenir :

$$\mathbf{P}\left\{|\overline{X_n} - p| \leq 0.1\right\} \geq 0.9$$

ou, autrement dit, $\mathbf{P}\left\{|\overline{X_n} - p| \geq 0.1\right\} \leq 0.1$.

Or, $\mathbf{E}[X_i] = p$ et $\mathbf{Var}[X_i] = p(1-p)$ car il s'agit de variables de Bernouilli, donc $\mathbf{E}\left[\overline{X_n}\right] = np/n = p$ et $\mathbf{Var}\left[\overline{X_n}\right] = \frac{1}{n^2}\sum_{i=1}^n \mathbf{Var}[X_i] = \frac{p(1-p)}{n}$. Donc, en utilisant l'inégalité de Chebyshev,

$$\begin{aligned}
\mathbf{P}\left\{|\overline{X_n} - p| \geq 0.1\right\} &\leq \frac{\mathbf{Var}[X]}{0.1^2} \\
&\leq \frac{p(1-p)}{0.01n} \\
&\leq \frac{100}{n}\frac{1}{4} \text{ car } p(1-p) \leq 1/4 \\
&\leq \frac{25}{n}
\end{aligned}$$

Il suffit donc d'avoir $25/n \leq 0.1$, autrement dit $n \geq 250$.

Exercice 2. *BucketSort*

Le tri par seaux est un algorithme de tri très simple dont la complexité moyenne est linéaire. Les hypothèses sont les suivantes : nous avons $n = 2^m$ nombres à trier, tirés uniformément et indépendamment sur $\{0, \dots, 2^k - 1\}$ avec $k \geq m$ (k est supposé connu). L'algorithme procède en deux étapes : d'abord, il effectue un pré-tri en jetant selon certaines règles les n éléments dans n seaux ; ensuite, il appelle un algorithme de tri simple (en temps quadratique, tel que tri par insertion ou tri par sélection) dans chaque seau. Enfin, il concatène dans l'ordre les listes triées obtenues dans chaque seau. Pour que l'algorithme soit exact, il faut que le pré-tri soit fait de telle sorte que tous les éléments du seau i soient inférieurs à tous les éléments du seau j, pour $i < j$.

1. Donner une façon très simple de faire le pré-tri tout en respectant la condition énoncée ci-dessus. On veut que le choix du seau pour un élément x soit effectué en temps constant (on suppose ici que les opérations arithmétiques peuvent être effectuées en temps constant).

☞ On suppose que les seaux sont numérotés de 0 à $n-1$. L'élément x peut être vu comme un nombre en binaire écrit sur k bits. On regarde les m bits de poids forts (ce qui revient à diviser par 2^{k-m}), cela nous donne un nombre s_x entre 0 et $n-1 \to$ on met x dans le seau numéro s_x. Ainsi, on est sûrs que $x < y$ si $s_x < s_y$, donc la condition d'exactitude du pré-tri est respectée.

2. Soit X_i la v.a. comptant le nombre d'éléments dans le seau i après le pré-tri. Quelle loi suit X_i ?

☞ X_i suit une loi binomiale de paramètres $(n, 1/n)$, car les n éléments d'entrée sont choisis indépendamment et uniformément.

3. Prouver que la complexité en moyenne est $\mathcal{O}(n)$.

☞ Le pré-tri coûte un temps constant pour chaque élément, donc un temps linéaire en n au total. Ensuite, le tri du seau numéro i coûte $c(X_i)^2$ pour une certaine constante c. Or, comme X_i suit une loi binomiale $(n, 1/n)$, on a (avec $p = 1/n$, $\mathbf{E}[X_i] = np$, $\mathbf{Var}[X_i] = np(1-p)$) :

$$\mathbf{E}\left[X_i^2\right] = \mathbf{E}[X_i]^2 + \mathbf{Var}[X_i] = (np)^2 + np(1-p) = np(np + 1 - p) = 1 \cdot (1 + 1 - \frac{1}{n}) = 2 - \frac{1}{n} < 2 .$$

Donc l'espérance du temps passé dans la deuxième étape est au plus

$$\mathbf{E}\left[\sum_{i=0}^{n-1} c(X_i)^2\right] = c\sum_{i=0}^{n-1}\mathbf{E}\left[X_i^2\right] \leq 2cn .$$

L'espérance du temps d'exécution totale est donc linéaire en n.

Exercice 3. *Interrupteurs*

Partie I :

1. Montrer qu'il existe une constante $\gamma > 0$ rendant l'énoncé suivant vrai : si une v.a. positive X vérifie $\mathsf{E}[X] = 1$ et $\mathsf{E}[X^2] \leq 3$, alors $\mathsf{P}(X \geq 1/4) \geq \gamma$.
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 $\mathsf{P}(X_i = 1) = \mathsf{P}(X_i = -1) = \frac{1}{2}$. On pose $Y = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$. Calculer $\mathsf{E}[Y^2]$ et $\mathsf{E}[Y^4]$ et en déduire que

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

☞ 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,

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

Partie II :

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 $a_{ij}b_i c_j = 1$. On considère la quantité

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

qui est le nombre d'ampoules allumées moins le nombre d'ampoules éteintes. Enfin, deux joueurs jouent au jeu suivant : le joueur 1 choisit la position des interrupteurs (a_{ij}), puis le joueur 2 choisit la position des interrupteurs (b_i) et (c_j). Le joueur 1 veut minimiser $F(a, b, c)$ et joueur 2 veut le maximiser. On considère donc

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

3. Montrer que $V(n) = O(n^{3/2})$ en considérant le cas où le joueur 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. Le joueur 2 applique la stratégie suivante : il 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 à l'aide de la question I.2 et en déduire que $V(n) = \Omega(n^{3/2})$.

☞ 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. *Estimer l'intersection avec un rectangle*

Let $P \subset \mathbb{Z}^2$ of size n. Our objective is to be able to answer quickly queries of the form what is the fraction of points in P that are in the rectangle $r = [a_1, b_1] \times [a_2, b_2]$? We write $r[P] = \frac{|P \cap r|}{n}$ for this fraction. We consider a simple data structure to approximate $r[P]$ efficiently for any query r. The data structure is just a random subset $S \subset P$ of size m. On query r, the estimate for $r[P]$ we output is $\frac{|S \cap r|}{m}$. The structure S defines an ε-approximation if for all queries r, we have $|r[P] - \frac{|S \cap r|}{m}| \leq \varepsilon$.

1. What m should we take to obtain an ε-approximation with probability $1 - \delta$?

☞ 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 5. *Chernoff Bound Interval*

Let X be an arbitrary random variable with $0 \leq X \leq 1$ and $\mathbf{E}[X] = p$. Consider the random variable $Y \in \{0,1\}$ with $\mathbf{P}\{Y = 1\} = p$. Show that for any $\lambda > 0$, $\mathbf{E}\left[e^{\lambda X}\right] \leq \mathbf{E}\left[e^{\lambda Y}\right]$. *Hint : you may want to use the convexity of the exponential function.*
Using this fact, show that the Chernoff bound we saw in class still holds if we replace the condition $X_i \in \{0,1\}$ by $X_i \in [0,1]$.

☞ 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 $e^{\lambda X}$:

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

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

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