TD 06 (corrigé)

Exercice 1. *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 $(X_1, \dots, X_n)$ 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 2. *Arrondi*

Soit U un ensemble à n éléments. On appelle recouvrement de U un ensemble $\mathcal{S} = (S_1, \cdots, S_m)$ de parties de U qui vérifie $\bigcup S_i = U$. Étant donné $\mathcal{S}$ un recouvrement de U, on note $\mathrm{OPT}(\mathcal{S})$ le cardinal minimal d'un sous-ensemble de $\mathcal{S}$ qui est encore un recouvrement de U.

1. Expliquer rapidement pourquoi $\mathrm{OPT}(\mathcal{S})$ est la solution du problème d'optimisation suivant :

$$\text{Minimiser } \sum_{i=1}^m x_i \text{ sous les contraintes } x_i \in \{0,1\} \text{ et } \sum_{i=1}^m x_i \mathbf{1}_{S_i} \geq \mathbf{1}_U \tag{1}$$

☞ Les sous-ensembles de $\mathcal{S}$ sont en bijections avec $\{0,1\}^m$. On a donc, pour tout sous-ensemble $\mathcal{T} \subseteq \mathcal{S}$, $\#\mathcal{T} = \sum x_i$ avec x_i qui vaut 1 si et seulement si $S_i \in \mathcal{T}$.

De plus, $\mathcal{T}$ est un recouvrement de U si et seulement si pour tout $y \in U$, on a $\mathbf{1}_{\mathcal{T}}(y) = 1 = \mathbf{1}_U(y)$. Or, $\mathbf{1}_{\mathcal{T}}(y) = 1$ si et seulement si $\sum_{i=1}^m x_i \mathbf{1}_{S_i}(y) \geq 1$ (car y doit être dans au moins un S_i qui est gardé dans $\mathcal{T}$, et il peut être dans plusieurs S_i différents).

On considère le problème suivant qui est une relaxation de (1) :

$$\text{Minimiser } \sum_{i=1}^m z_i \text{ sous les contraintes } z_i \in [0,1] \text{ et } \sum_{i=1}^m z_i \mathbf{1}_{S_i} \geq \mathbf{1}_U \tag{2}$$

Alors que le problème (1) est NP-difficile, le problème (2) peut être résolu en temps polynomial par les méthodes de programmation linéaire.

2. Soit k un entier, et $(z_i)_{1\leq i\leq m}$ qui minimisent (2). Soient $(X_{i,j})_{1\leq i\leq m, 1\leq j\leq k}$ des variables aléatoires indépendantes vérifiant $\mathbf{P}\{X_{i,j} = 1\} = z_i$, $\mathbf{P}\{X_{i,j} = 0\} = 1 - z_i$. On définit un sous-ensemble $\mathcal{T} \subset \mathcal{S}$ par la condition :

$$S_i \in \mathcal{T} \iff \exists j \in \{1, \ldots, k\} \, : \, X_{i,j} = 1\,.$$

Montrer que :

$$\mathbf{E}\left[\#\mathcal{T}\right] \leq k \, \mathrm{OPT}(\mathcal{S})\,.$$

☞ Par linéarité de l'espérance, on a :

$$\begin{aligned}
\mathbf{E}\left[\#\mathcal{T}\right] &= \sum_{i=1}^{m} \mathbf{P}\left\{S_i \in \mathcal{T}\right\} \\
&= \sum_{i=1}^{m} \mathbf{P}\left\{X_{i,1} + \ldots + X_{i,k} \geq 1\right\} \\
&\leq \sum_{i=1}^{m} \mathbf{E}\left[X_{i,1} + \ldots + X_{i,k}\right] \\
&\leq k \times \sum_{i=1}^{m} z_i\,.
\end{aligned}$$

Puis, les (z_i) minimisent (2), qui est une relaxation de (1), donc $\sum z_i \leq \mathrm{OPT}(\mathcal{S})$.

3. Déterminer une valeur de $c > 0$ telle que, si on pose $k = \lfloor c \log n \rfloor$, on ait :

$$\mathbf{P}\left\{\mathcal{T} \text{ est un recouvrement de } U\right\} \geq 1 - \frac{1}{n}\,.$$

☞ Comme d'habitude, avec la borne de l'union, c'est plus facile de partir de :

$$\mathbf{P}\left\{\mathcal{T} \text{ n'est pas un recouvrement de } U\right\} \leq \sum_{y \in U} \mathbf{P}\left\{y \text{ n'apparait pas dans } \mathcal{T}\right\}\,.$$

On note A_y l'événement $\{y \notin \bigcup_{T \in \mathcal{T}} T\}$. Alors on a :

$$A_y = \left\{ \sum_{i \text{ t.q. } y \in S_i} \sum_{j=1}^{k} X_{i,j} = 0 \right\}$$

Tous les $X_{i,j}$ sont indépendants et suivent une loi de Bernouilli, la somme des paramètres est donc $\mu := \sum_{i \text{ t.q. } y \in S_i} k \times z_i$, et on a (en supposant U non vide, mais sinon tout cela n'est pas très intéressant) $\sum z_i \geq 1$, donc $\mu \geq k$.
Avec Chernoff II appliquée avec $\epsilon = 1$, on obtient

$$\mathbf{P}\left\{A_y\right\} \leq \exp(-\frac{\mu}{3}) \leq \exp(-\frac{k}{3}) \leq \exp(-\frac{c\log(n)}{3}) = n^{-c/3}\,.$$

En particulier, pour $c = 6$, on a bien $\mathbf{P}\left\{\mathcal{T} \text{ n'est pas un recouvrement de } U\right\} \leq \frac{1}{n}$.
Cas $\delta = 1$. On peut ne pas utiliser Chernoff et avoir une version + directe. On a

$$\mathrm{P}(A_y) = \mathrm{P}(\bigcap_{i : y \in S_i} \bigcap_{j=1}^{k} \{X_{i,j} = 0\}) \tag{3}$$
$$= \prod_{i : y \in S_i} \prod_{j=1}^{k} \mathrm{P}(X_{i,j} = 0) \tag{4}$$
$$= \prod_{i : y \in S_i} \prod_{j=1}^{k} (1 - z_i) \tag{5}$$
$$\leq \prod_{i : y \in S_i} \prod_{j=1}^{k} e^{-z_i} \tag{6}$$
$$= e^{-\sum_{i : y \in S_i} \sum_{j=1}^{k} z_i} \tag{7}$$
$$= e^{-\mu} \leq e^{-k} \leq e^{-(c \ln n + 1)} \tag{8}$$
$$= n^{-c} e^{-1} \tag{9}$$

et borne de l'union :

$$\mathrm{P}(\bigcup_y A_y) \leq n \mathrm{P}(A_y) \tag{10}$$
$$= n^{1-c} e^{-1} \tag{11}$$

On peut prendre $c > 2$ qui fonctionne pour n pas trop petit.

Exercice 3. *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 :

$\mathcal{Q}_r$: « 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) :

$$2e^{-\frac{\varepsilon'2}{2+\varepsilon'}\mathbf{E}[X(r)]} \leq 2e^{-\frac{\varepsilon2}{2r[P]+\varepsilon}m} \leq 2e^{-\frac{\varepsilon2}{2+\varepsilon}m}.$$

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,

$$\mathbf{P}\{\exists r \text{ s.t. } |r[P] - X(r)| \geq \varepsilon\} \leq \sum_{i=1}^{k} \mathbf{P}\{|r_i[P] - X(r_i)| \geq \varepsilon\} \leq k2e^{-\frac{\varepsilon2}{2+\varepsilon}m} .$$

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

$$n^4 2e^{-\frac{\varepsilon2}{2+\varepsilon}m} \leq \delta ,$$

ce qui est possible pour

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

Exercice 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 $\mathrm{E}[N] = \sum \mathrm{P}(A_i) = 2n(1-p)^n$.

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 :

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

☞

i. Si $c > 1$, on a

$$\begin{aligned} \mathsf{E}[N] &= 2n(1-p)^n \\ &= 2n\left(1 - \frac{c \log n}{n}\right)^n \\ &= 2n \exp(n \log(1 - \frac{c \log n}{n})) \to 0 \end{aligned}$$

et donc $\mathsf{P}(N \geq 1) \leq \mathsf{E}[N] \to 0$.

ii. 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 (la première partie tend vers 0 avec un développement limité, les deux autres chacune vers 1/2). 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 $2C^2 > 1$.