A la recherche d'indices dans un labyrinthe

Angèle RAZET

Je remercie Quentin BERGER d'avoir encadré ce mémoire.

Table des matières

1 Etude du problème de la clique cachée

1.1 Présentation du problème

Pour donner une idée de ce que l'on cherche à modéliser et de l'intuition sur les objets que l'on considèrera, commençons par étudier le problème de la clique cachée dans sa version gaussienne.
Pour cela, on se donne $n \in \mathbb{N}^*$, $1 \leq k \leq n$ un entier et $\mu > 0$ un réel.

Nous allons étudier G, le graphe complet non-orienté à n sommets. On note A l'ensemble de ses arêtes. Dans le problème de la clique cachée, on pondère aléatoirement les arêtes de G de deux manières, c'est-à-dire qu'on se donne deux vecteurs aléatoires $X^0 = (X^0_a)_{a\in A}$ et $X^1 = (X^1_a)_{a\in A}$ de $\mathbb{R}^A$. Voici leur loi, que l'on notera respectivement $\mathbb{P}_0$ et $\mathbb{P}_1$:

— **loi de** $(X^0_a)_{a\in A}$ **:** $\forall a \in A$, $X^0_a \sim \mathcal{N}(0,1)$;
— **loi de** $(X^1_a)_{a\in A}$ **:** on choisit au hasard une k-clique K ; si $a \in K$, $X^1_a \sim \mathcal{N}(\mu, 1)$ sinon $X^1_a \sim \mathcal{N}(0,1)$.

On remarque que $\mathbb{P}_0$ *et* $\mathbb{P}_1$ *sont des mesures de probabilité de densité par rapport à la mesure de Lebesgue* λ. *On pose* $\mathbb{P}_0 = f_0\lambda$ *et* $\mathbb{P}_1 = f_1\lambda$, *où* f_0 *et* f_1 *sont des fonctions continues strictement positives.*

Définiton (Test) : Un test est une fonction T de $\mathbb{R}^A$ dans $\{0,1\}$ telle que $T^{-1}(0)$ est $\mathbb{P}_0$-mesurable et $\mathbb{P}_1$-mesurable.

Dans le problème de la clique cachée, on pose $R(T) = \mathbb{P}_0(T(X^0) = 1) + \mathbb{P}_1(T(X^1) = 0)$ le risque associé à un test T. Le but est de minimiser ce risque.

Donnons maintenant une idée de ce que cela pourrait modéliser. L'environnement sur lequel on travaille est le graphe G, les données sont les pondérations des arêtes. Supposons que cet environnement est contaminé avec probabilité $0 < p < 1$ et que lorsque les données sont contaminées, la pondération des arêtes suit la loi $\mathbb{P}_1$ (environnement bruité et contaminé), sinon elle suit la loi $\mathbb{P}_0$ (environnement bruité). Ainsi le vecteur aléatoire $X = (X_a)_{a\in A}$ généré suit une loi $\mathbb{P} = (pf_1 + (1-p)f_0)\lambda$.
On veut, à chaque pondération $x \in \mathbb{R}^A$ du graphe, dire si elle est généré par un environnement contaminé ou non. C'est le rôle du test. Cependant, les variables suivant des lois normales, chaque pondération peut avoir été générée avec et sans contamination des données. On note C l'évenement « les données sont contaminées ». On cherche donc à mesurer l'erreur faite. Supposons que la pondération générée par X est $x \in \mathbb{R}^A$ et que le test affirme que cette pondération n'est pas issue d'une contamination. Alors la probabilité que le test se trompe est : $\mathbb{P}(C|X = x) = \frac{pf_1(x)}{pf_1(x)+(1-p)f_0(x)}$. C'est une fonction de X qui dépend de T, on la note g_T. On peut alors mesurer son espérance sous la loi $\mathbb{P}$, on a directement $\mathbb{E}(g_T(X)) = (1-p)\mathbb{P}_0(T(X) = 1)+p\mathbb{P}_1(T(X) = 0)$. C'est la probabilité moyenne que le test se trompe. Etant donné que $\mathbb{P}_0(T(X) = 1)$ et $\mathbb{P}_1(T(X) = 0)$ sont des termes positifs, faire tendre vers 0 le risque $R(T)$ est equivalent à faire tendre vers 0, $\mathbb{E}(g_T(X))$.

Remarque : On peut imaginer que plus k et μ sont grands, plus la détection d'une contamination est aisée.

1.2 Quelques résultats

Les resultats suivants sont admis, ils sont énoncés pour donner une idée de que l'on cherche à obtenir. Des démonstrations sont données dans une cadre plus général dans le cours de G. Lugosi [3]. Cette généralité permet de noter l'impact de la symétrie du problème sur les bornes obtenues.
On pose $K = \binom{k}{2}$, le nombre d'arêtes d'un k-clique.

1.2.1 Bornes inférieures

Soit $\delta > 0$. Pour $\mu \geq \sqrt{\frac{8n}{K^2}\log\frac{2}{\delta}}$, le test $T = \mathbf{1}_{\sum_{a\in A} X_a \geq \frac{\mu K}{2}}$ vérifie $R(T) \leq \delta$.
Remarque : Comme on l'avait supposé, la borne est proportionnelle à $\frac{1}{k}$, à $\sqrt{n}$ et environ à $\sqrt{\log\delta}$.

1.2.2 Borne supérieure

Soit $\delta > 0$. Si $\mu \leq \sqrt{\frac{1}{K}\log\left(1 + \frac{4n(1-\delta)^2}{K^2}\right)}$, alors tout test T vérifie, $R(T) \geq \delta$.

2 Treillis et arbres binaires

Dans cette partie, nous allons étudier en détail les graphes que sont les treillis T_m et les arbres binaires B_m de longueur finie $m \in \mathbb{N}^*$.
Rappelons la définition d'un treillis :

— les sommets sont les $(i,j) \in \mathbb{Z}^2$ où $0 \leq i \leq m$ et $-i \leq j \leq i$ et $i = j[2]$;

— les arêtes sont les $\{(i,j),(i+1,j+1)\}$ et les $\{(i,j),(i+1,j-1)\}$.

2.1 Cadre et résultats

Cadre :

— Les vecteurs seront les sommets du graphe. La classe $\mathcal{C}$ est l'ensemble des chemins de longueur m ;

— la loi $\mathbb{P}_0$ est telle que : $X_i \sim \mathcal{N}(0,1)$ pour tout $i \in [\![1,n]\!]$: c'est l'hypothèse $\mathcal{H}_0$;

— pour $p \in \mathcal{C}$ la loi $\mathbb{P}_{1,p}$ est telle que : $X_i \sim \mathcal{N}(\mu,1)$ pour tout $i \in p$ sinon $X_i \sim \mathcal{N}(0,1)$;

— π une distribution sur $\mathcal{C}$: c'est l'hypothèse $\mathcal{H}_1$;

— on considèrera deux risques :

1. $\gamma(T) = \mathbb{P}_0(T(X)=1) + \max\limits_{p\in\mathcal{C}} \mathbb{P}_{1,p}(T(X)=0)$;
2. $\gamma_\pi(T) = \mathbb{P}_0(T(X)=1) + \mathbb{E}_\pi\, \mathbb{P}_{1,p}(T(X)=0)$.

Remarque : on notera $\mathbb{P}_1(T(X)=0) = \mathbb{E}_\pi\, \mathbb{P}_{1,p}(T(X)=0)$ *ou* $\max\limits_{p\in\mathcal{C}} \mathbb{P}_{1,p}(T(X)=0)$ *suivant que le risque est* γ_π *ou* γ.

On cherche à déterminer pour une suite (μ_m) (indexée donc par la longueur des graphes) s'il existe une suite de tests (T_m) telle que $\gamma_{(\pi)}(T_m) \xrightarrow[n\to\infty]{} 0$. On dira alors que (T_m) est un succès.

Voici les résultats dont nous donnerons une démonstration partielle ou complète. Les deux premiers résultats concernent les treillis et le dernier les arbres binaires.

Théorème 1 : On prend π la distribution uniforme sur les chemins et γ_π le risque. Alors :

(i) si $\mu_m m^{\frac{1}{4}} \xrightarrow[n\to\infty]{} +\infty$ alors il existe une suite de tests qui est un succès ;

(ii) si $\mu_m m^{\frac{1}{4}} \xrightarrow[n\to\infty]{} 0$ aucune suite de tests n'est un succès.

Théorème 2 : On prend γ le risque. Alors :

(i) si $\mu_m \log(m)^{\frac{1}{2}} \xrightarrow[n\to\infty]{} +\infty$ alors il existe une suite de tests qui est un succès ;

(ii) si $\mu_m \log(m)(\log\log m)^{\frac{1}{2}} \xrightarrow[n\to\infty]{} 0$ aucune suite de tests n'est un succès.

Remarque : étant donné que $\gamma_\pi \leq \gamma$ *quelque soit* π*, il est cohérent que les bornes de ce deuxième théoreme soient plus grandes.*

Théorème 3 : On prend π la distribution uniforme sur les chemins et γ_π le risque. Alors :

(i) si $\mu_m \geq \sqrt{2\log(2)}$ alors il existe une suite de tests qui est un succès ;

(ii) si $\mu_m < \sqrt{2\log(2)}$ aucune suite de tests n'est un succès.

2.2 Preuves des resultats

2.2.1 Idées et résultats généraux

— **Test minimisant :** On pose $L(x) = \frac{dP_1}{d\mathbb{P}_0}$. On note T^* le test ainsi défini : $T^*(x) = 1 \Leftrightarrow L(x) \geq 1$. Ainsi pour tout test T, $\gamma_{(\pi)}(T) \geq \gamma_{(\pi)}(T^*)$ et on pose $R^* = \gamma_{(\pi)}(T^*)$.

Preuve :

$$\begin{aligned}\gamma_{(\pi)}(T) - R^* &= \int_{R^n} \mathbf{1}_{T(x)=1}\, d\mathbb{P}_0 + \int_{R^n} \mathbf{1}_{T(x)=0}\, d\mathbb{P}_1 - \int_{R^n} \mathbf{1}_{T^*(x)=1}\, d\mathbb{P}_0 - \int_{R^n} \mathbf{1}_{T^*(x)=0}\, d\mathbb{P}_1 \\ &= \left(\int_{R^n} \mathbf{1}_{T(x)=1}\mathbf{1}_{T^*(x)=0}\, d\mathbb{P}_0 - \int_{R^n} \mathbf{1}_{T(x)=1}\mathbf{1}_{T^*(x)=0}\, d\mathbb{P}_1\right) \\ &+ \left(\int_{R^n} \mathbf{1}_{T(x)=0}\mathbf{1}_{T^*(x)=1}\, d\mathbb{P}_1 - \int_{R^n} \mathbf{1}_{T(x)=0}\mathbf{1}_{T^*(x)=1}\, d\mathbb{P}_0\right)\end{aligned}$$

Par définiton de T^*, on remarque que les deux parenthèses sont positives. Ainsi $\gamma_{(\pi)}(T) - R^* \geq 0$.

— **Première expression de R^* :** On a $\gamma_{(\pi)}(T^*) = 1 - \frac{1}{2}\,\mathbb{E}_0\,|1 - L(X)|$.
Preuve :

$$\begin{aligned}\gamma_{(\pi)}(T^*) &= \int_{R^n} \mathbf{1}_{T^*(x)=1}\, d\mathbb{P}_0 + \int_{R^n} \mathbf{1}_{T*(x)=0}\, d\mathbb{P}_1 \\ &= \int_{R^n} \left(\mathbf{1}_{T^*(x)=1} + \mathbf{1}_{T^*(x)=0}L(x)\right)\, d\mathbb{P}_0 \\ &= 1 + \int_{R^n} \left(\mathbf{1}_{T^*(x)=1} - \mathbf{1}_{T^*(x)=1}L(x)\right)\, d\mathbb{P}_0 \\ &= 1 - \int_{R^n} (L(x) - 1)\, \mathbf{1}_{L(x)-1\geq 0}\, d\mathbb{P}_0\end{aligned}$$

Or $\int_{R^n} (1 - L(x))\, d\mathbb{P}_0 = 0$. Ainsi $\gamma_{(\pi)}(T^*) = 1 - \frac{1}{2}\,\mathbb{E}_0\,|1 - L(X)|$.

— **Seconde expression de R^* :** On a $\gamma_{(\pi)}(T^*) = 1 - \mathbb{E}_0\,((1 - L(X))_+)$.
Preuve : La preuve utilise le résultat précédent.
On remarque de plus que $\mathbb{E}_0\,(1 - L(X)) = 0$. Or $\mathbb{E}_0\,(1 - L(X)) = \mathbb{E}_0\,((1 - L(X))_+) - \mathbb{E}_0\,((1 - L(X))_-)$. Ainsi, $\mathbb{E}_0\,|1 - L(X)| = \mathbb{E}_0\,((1 - L(X))_+) + \mathbb{E}_0\,((1 - L(X))_-) = 2\,\mathbb{E}_0\,((1 - L(X))_+)$. D'où le resultat attendu.

— **Majoration de $R*$:** La définition de L implique que $\int_{R^n} L(x)\, d\mathbb{P}_0 = 0 = 1$. De plus, par l'inégalité de Jensen, on a :

$$\begin{aligned}\mathbb{E}_0\,|1 - L(X)| &\leq \sqrt{\mathbb{E}_0\,((1 - L(X))^2)} \\ &= \sqrt{\mathbb{E}_0\,(1 - 2L(X) + L(X)^2)} \\ &= \sqrt{1 + \mathbb{E}_0\,(L(X)^2)}\end{aligned}$$

— **Calcul de $\mathbb{E}_0\,(L(X)^2)$ (pour $\mathbb{P}_1 = \mathbb{E}_\pi\,\mathbb{P}_{1,p}$) :** Nous allons montrer que $\mathbb{E}_0\,(L(X)^2) = \mathbb{E}_{\pi\otimes\pi}(e^{\mu^2|Z|})$.

$$L(X = (x_1, ..., x_n)) = \mathbb{E}_\pi e^{\mu X_p - m\mu^2}$$

Donc :

$$\mathbb{E}_0\,(L(X)^2) = \mathbb{E}_{\pi\otimes\pi}\mathbb{E}_0\,(e^{\mu(X_p + X_{p'} - 2m\mu^2}$$

Or ,

$$\begin{aligned}\mathbb{E}_0\, e^{\mu(X_p + X_{p'})} &= \mathbb{E}_0\, e^{\mu(\sum_{v\in p\ p'} X_v + \sum_{v\in p'\ p} X_v + 2\sum_{v\in p\cap p'} X_v))} \\ &= (\mathbb{E}_0\, e^{\mu X})^{2(m-Z)}(\mathbb{E}_0\, e^{2\mu X})^Z \\ &= (e^{\mu^2})^{2(k-Z)}(e^{2\mu^2})^Z \\ &= e^{\mu^2 Z}.\end{aligned}$$

2.2.2 Preuve du théorème 1

Borne inférieure : On suppose que : $\mu_m m^{\frac{1}{4}} \xrightarrow[n\to\infty]{} +\infty$. On peut donc se donner une suite $(h_m)_{m\in\mathbb{N}^*}$, telle que $\mu_m h_m^{-\frac{1}{2}} m^{\frac{1}{4}} \xrightarrow[n\to\infty]{} +\infty$ et $h_m \xrightarrow[n\to\infty]{} +\infty$.
L'introduction de $(h_m)_{m\in\mathbb{N}^*}$ permet de se donner une petite marge dans les calculs qui feront converger des termes vers 0.

Désormais, on note N_m le nombre de sommets dont l'ordonnée en valeur absolue est inférieure à $h_m\sqrt{m}$. On a donc $N_m \underset{+\infty}{\sim} h_m m^{\frac{3}{2}}$. On définit ainsi $\mathcal{P}_m$ qui est l'ensemble des chemins dont tous les sommets sont dans N_m puis $S_m = \sum_{v\in\mathcal{P}_m} X_v$ qui est une variable aléatoire.

On va considérer les tests T_m tels que $T_m = \mathbf{1}_{S_m \geq \frac{\mu m}{2}}$. On veut montrer que cette série de test est un succès.
L'idée est de majorer le risque par une somme de trois termes comme ci-dessous :
$\gamma_\pi \leq \mathbb{P}(\mathcal{N}(0, N_m) \geq \frac{\mu_m m}{2}) + \mathbb{P}(\mathcal{N}(\mu_m m, N_m) < \frac{\mu_m m}{2}) + \pi(p \notin \mathcal{P}_m)$.
On va maintenant montrer que chacun de ces trois termes tend vers 0.

Majoration de $\mathbb{P}(\mathcal{N}(0, N_m) \geq \frac{\mu_m m}{2})$: par l'inégalité de Bienaymé-Tchebichev, on a :
$\mathbb{P}(\mathcal{N}(0, N_m) \geq \frac{\mu_m m}{2}) \leq \frac{4N_m}{m^2\mu_m^2} \sim \left(\frac{4\sqrt{h_m}}{\mu_m m^{\frac{1}{4}}}\right)^2 \xrightarrow[n\to\infty]{} 0.$

La majoration de $\mathbb{P}(\mathcal{N}(\mu_m m, N_m) < \frac{\mu_m m}{2})$ se fait de la même manière.

Majoration de $\pi(p \notin \mathcal{P}_m)$: Ici, on fait l'analogie entre la distribution uniforme sur les chemins et le résultat de la marche aléatoire de paramètre $\frac{1}{2}$. On utilise l'existence d'une bijection entre les marches telles que $S_m > h_m\sqrt{m}$ et celles qui ne sont pas dans $\mathcal{P}_m$ mais telles que $S_m \leq h_m\sqrt{m}$. L'union de ces deux ensembles de chemins est disjointes et égale à $\mathcal{P}_m$.
Ainsi, en utilisant l'inégalité de Bienaymé-Tchebichev, on a $\pi(p \notin \mathcal{P}_m) = \mathbb{P}(|S_m| > h_m\sqrt{m}) \leq \frac{m}{4h_m^2 m} \xrightarrow[n\to\infty]{} 0$.

Borne supérieure : Nous allons monter que si $\mu_m m^{\frac{1}{4}} \xrightarrow[n\to\infty]{} 0$ alors $R_m^* \xrightarrow[n\to\infty]{} 1$.
On utlise les resulats établis dans la partie précédente. Le but est donc de montrer que $\mathbb{E}(\exp \mu^2|Z|)) \xrightarrow[n\to\infty]{} 1$.
A nouveau, on fait l'analogie avec le processus de marche aléatoire. On identifie donc cette espérence avec $\mathbb{E}(e^{\mu^2 Z_m})$ où Z_m est le nombre de retours en zéro de la marche aléatoire.

Lemme : Notons τ le temps de premier retour en zéro. On a $\mathbb{P}(\tau \geq m) \geq \frac{C}{\sqrt{m}}$ pour un certain $C > 0$.

Preuve : On pose $a_n = \mathbb{P}(S_n = 0) = \frac{\binom{n}{k}}{2^n}$ et $b_n = \mathbb{P}(\tau = n)$ $(b_0 = 0)$ pour tout $n \in \mathbb{N}$. On considère $A(x)$ et $B(x)$ les séries réelles génératrices associées.
On remarque que pour $n \geq 1$, $a_n = \sum_{k=0}^n b_k a_{n-k}$. Ainsi, $A(x) = A(x)B(x) + 1$. Or on peut caculer explicitement A : $A(x) = \frac{1}{\sqrt{1-x^2}}$. Donc $B(x) = 1 - \sqrt{1-x^2}$ et $b_{2n} = \frac{1}{2n-1}\binom{2n}{n}4^{-n}$ et $b_{2n+1} = 0$.
On peut calculer $\mathbb{P}(\tau \geq m) = \sum_{k,2k\geq m} \frac{1}{2k-1}\binom{2k}{k}4^{-k} \geq \int_m^\infty \frac{1}{4t^{\frac{3}{2}}}\,dt \geq \frac{C}{\sqrt{m}}$.

On a désormais :

$$\begin{aligned}
\mathbb{E}(e^{\mu^2 Z_m)}) &= \sum_{k=0}^{m-1} \mathbb{P}(Z_m = k)\, e^{\mu^2 k} \\
&\underset{Abel}{=} 1 - \mathbb{P}(Z_m \geq m)\, e^{\mu^2 m} + \sum_{k=1}^{m} \mathbb{P}(Z_m \geq k)\, (e^{\mu^2 k} - e^{\mu^2 (k-1)}) \\
&= 1 - \mathbb{P}(Z_m \geq m)\, e^{\mu^2 m} + (1 - e^{-\mu^2}) \sum_{k=1}^{m} \mathbb{P}(Z_m \geq k)\, e^{\mu^2 k}
\end{aligned}$$

Or :

$$\begin{aligned}\mathbb{P}(Z_m \geq k) &\leq (\mathbb{P}(\tau \leq m))^k \\ &\leq (1 - \frac{C}{\sqrt{m}})^k \leq e^{-\frac{kC}{\sqrt{m}}}\end{aligned}$$

Ainsi , en se donnant $(\epsilon_m)_{m\in\mathbb{N}}$ telle que $\mu_m \leq \epsilon_m m^{\frac{-1}{4}}$ et $\epsilon_m \xrightarrow[n\to\infty]{} 0$:

$$\begin{aligned}(1 - e^{-\mu^2}) \sum_{k=1}^{m} \mathbb{P}(Z_m \geq k)\, e^{\mu^2 k} &\leq \mu^2 \sum_{k=1}^{m} e^{\mu^2 k - \frac{kC}{\sqrt{m}}} \\ \underset{Somme\ de\ Riemann}{\leq} \epsilon_m \int_0^{\infty} e^{-tC_1}\, dt &\xrightarrow[n\to\infty]{} 0\end{aligned}$$

On peut donc conclure que $\mathbb{E}(e^{\mu^2 Z_m)}) \xrightarrow[n\to\infty]{} 1$.

2.2.3 Preuve du théorème 2 :

On rappelle que pour ce théorème, le risque à minimiser est γ et non plus γ_π.

Borne inférieure : Cette preuve ne fait intervenir aucun nouvel outil. Elle consiste à montrer que la série de tests définie par $T_m = \mathbf{1}_{S_m \geq \frac{\mu_m}{2}}$ où $S_m = \sum_{i=1}^{m} \frac{1}{i} \sum_{j\leq i} X_{ij}$ est un succès.
Pour ce faire, on considère comme pour la preuve précédente la loi de S_m sous $\mathbb{P}_0$ et sous $\mathbb{P}_{1,p}$ et on se ramène à des lois normales.

Borne supérieure : Le but est ici de minorer γ, et pour ce faire on va s'appuyer sur le fait que $\gamma \geq \gamma_\pi$ quelle que soit la distribution π envisagée. Le but est comme ci-dessus de montrer que $\mathbb{E}_{\pi\otimes\pi}(e^{\mu^2 Z_m}) \xrightarrow[n\to\infty]{} 1$ pour une certaine distribution π. Au lieu de construire une distribution, nous allons passer par un processus aléatoire. La démonstration de l'existence d'un tel processus est difficile, elle introduit la notion de prédictabilité.

Définiton (Processus aléatoire) : Un processus aléatoire à temps discret est une suite de variables aléatoire non-nécessairement indépendantes.

Définition (Prédictabilité) : Soit (S_n) un processus aléatoire, la prédictabilité de (S_n) est définie par :

$$PRE_S(k) = \sup_{x\in\mathbb{Z}, n\in\mathbb{N},\, x_0,\ldots,x_n} \mathbb{P}(S_{n+k} = x \mid S_0 = x_0, .., S_n = x_n) \tag{10}$$

Le but est donc de mesurer la visibilité maximale qu'on a à une certaine distance k. Voici deux lemmes que nous admettrons.

Lemme 1 : Soit $(aj)_{j\in\mathbb{N}}$, une suite de réels positifs. On suppose que $\sum_{j\in\mathbb{N}} a_j < 1$. Alors il existe un processus aléatoire $(S_n)_{n\in\mathbb{N}}$ où $\mid S_{n+1} - S_n \mid = 1$ tel que $PRE_S(k) \leq \frac{20}{ka\left\lfloor \log_2(\frac{k}{2}) \right\rfloor}$.

Lemme 2 : Soit $(S_n)_{0\leq n\leq m}$ un processus aléatoire. On suppose que $\sum_{1\leq k\leq \left\lfloor \frac{m}{B} \right\rfloor} PRE_S(k) \leq \theta < 1$, alors quel que soit $(v_n)_{0\leq n\leq m}$ un chemin, $\mathbb{P}(\mid S \cap v \mid \geq k) \leq B\theta^{\frac{k}{B}}$.

On choisit donc pour tout $m \in \mathbb{N}$ un processus aléatoire $S^{(m)}$ généré par le lemme 1 associé à la suite $a_j(m) = \frac{1}{3log_2(m)} \mathbf{1}_{j \leq \log_2(m)}$. Ainsi, on a par le premier lemme que $PRE_S(k) \leq \frac{60\ \log_2(m)}{k}$.
On pose $B_m = 120\ \frac{(\log(m))^2}{\log(2)}$. Ainsi, en prévison de l'application du second lemme, on a :

$$\sum_{1 \leq k \leq \lfloor \frac{m}{B} \rfloor} PRE_S(k) = \sum_{1 \leq k \leq \lfloor \frac{m}{B} \rfloor} \frac{60\ \log_2(m) \log(2)}{120\ k (\log(m))^2}$$
$$= \frac{1}{2\ \log(m)} \sum_{1 \leq k \leq \lfloor \frac{m}{B} \rfloor} \frac{1}{k} < \frac{1}{2}$$

On peut maintenant calculer la limite de $\mathbb{E}(e^{\mu^2 Z_m})$.

Remarquons pour commencer que pour tout $K \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{E}(e^{\mu^2 Z_m}) &= \sum_{k \geq 1} e^{\mu_m^2 k}\, \mathbb{P}(N_m = k) \\
&= \sum_{k=1}^{K-1} e^{\mu_m^2 k}\, \mathbb{P}(N_m = k) \ + \sum_{k \geq K} e^{\mu_m^2 k}\, (\mathbb{P}(N_m \geq k) - \mathbb{P}(N_m \geq k+1)) \\
&\leq e^{\mu_m^2 K} + (1 - e^{-\mu_m^2}) \sum_{k \geq K} \mathbb{P}(N_m \geq k)\, e^{\mu_m^2 k}
\end{aligned}$$

On pose $K = K_m = \frac{2B_m \log(B_m)}{\log(2)}$ et $a_m = e^{\mu_m^2 - \frac{\log(2)}{B_m}}$. Voici quelques résultats utiles pour la fin du calcul :
— $a_m < 1$;
— $\mu_m^2 B_m \xrightarrow[n \to \infty]{} 0$
— $\mu_m^2 K_m \xrightarrow[n \to \infty]{} 0$;
— $\frac{1}{1-a_m} \leq \frac{1}{1 - e^{-\frac{\log(2)}{B_m}}} = O(B_m)$
— $e^{-\log(2) \frac{K_m}{B_m}} = \frac{1}{B_m}$

$$\begin{aligned}
e^{\mu_m^2 K} + (1 - e^{-\mu_m^2}) \sum_{k \geq K} \mathbb{P}(N_m \geq k)\, e^{\mu_m^2 k} &\leq e^{\mu_m^2 K_m} + (1 - e^{-\mu_m^2}) \sum_{k \geq K} e^{\mu_m^2 k + \log(B_m) - \log(2) \frac{k}{B_m}} \\
&\leq e^{\mu_m^2 K_m} + B_m (1 - e^{-\mu_m^2})\, a_m^{K_m \frac{1}{1-a_m}} \\
&\leq e^{\mu_m^2 K_m} + \mu_m^2 B_m^2 e^{K_m \mu_m^2} e^{-\frac{K_m \log(2)}{B_m}} \\
&= e^{\mu_m^2 K_m} + \mu_m^2 B_m e^{K_m \mu_m^2} \xrightarrow[n \to \infty]{} 1
\end{aligned}$$

Cela achève la démonstration du théorème.

2.2.4 Preuve du théorème 3

On étudie maintenant le cas des arbres binaires.

Borne inférieure : On pose $M_m = \max \{X_p \mid p \in P_m\}$, où P_m est l'ensemble des chemins de longueur m et X_p est la somme des valeurs des noeuds du chemin p. Le test T_m sera défini par la valeur de M_m.

On calcule : $\mathbb{P}_0(X_p > m\sqrt{2\ \log(2)}) = \mathbb{P}(\mathcal{N}(0,m) > m\sqrt{2\ \log(2)}) = \mathbb{P}(\mathcal{N}(0,1) > \sqrt{2m\ \log(2)})$.

Or on sait que : $\mathbb{P}(\mathcal{N}(0,1) > t) \leq \frac{e^{-\frac{t^2}{2}}}{t\sqrt{2\pi}}$.

Ainsi, $\mathbb{P}_0(X_p > m\sqrt{2\ \log(2)}) \leq \frac{1}{2^m\ \sqrt{4\pi m \log(2)}}$.

En sachant que $|\ \mathcal{P}_m\ | = 2^m$, on peut calculer la limite en $m \to +\infty$ de $\mathbb{P}_0(M_m \geq m\sqrt{2\ \log(2)})$:

$$\begin{aligned}\mathbb{P}_0(M_m > m\sqrt{2\ \log(2)}) &\leq \sum_{p\in\mathcal{P}_m} \mathbb{P}_0(X_p > m\sqrt{2\ \log(2)}) \\ &\leq 2^m\ \frac{1}{2^m\ \sqrt{4\pi m \log(2)}} \\ &\xrightarrow[m\to\infty]{} 0.\end{aligned}$$

Montrons mainteant que $\mathbb{P}_1(M_m > m\sqrt{2\ \log(2)}) \xrightarrow[m\to\infty]{} 1$.

Nous allons séparer 2 cas :

— si $\mu > \sqrt{2\ \log(2)}$, alors pour $p \in \mathcal{C}_m$:

$$\begin{aligned}\mathbb{P}_1(M_m > m\sqrt{2\ \log(2)}) &\geq \mathbb{P}_{1,p}(M_m > m\sqrt{2\ \log(2)}) \\ &= \mathbb{P}(\mathcal{N}(0, m\mu) > m\sqrt{2\ \log(2)}) \\ &\xrightarrow[m\to\infty]{} 1\end{aligned}$$

— si $\mu = \sqrt{2\ \log(2)}$:

$$\begin{aligned}\mathbb{P}_1(M_m > m\sqrt{2\ \log(2)}) &\geq \mathbb{P}_{1,p}(M_m > m\sqrt{2\ \log(2)}) \\ &= \frac{1}{2}\end{aligned}$$

Ces calculs montrent que pour $\mu_m = \mu > \sqrt{2\ \log(2)}$, la série de tests $T_m = T = \mathbf{1}_{M_m > m\sqrt{2\ \log(2)}}$ est un succès. Cependant, le dernier calcul montre que ce test n'est pas suffisamment "fin" pour couvir le cas où $\mu_m = \mu = \sqrt{2\ \log(2)}$.

Pour traiter ce cas, l'idée est d'introduire une série d'indices $(m_n)_{n\in\mathbb{N}}$ telle que :

— $\sum_{n\in\mathbb{N}} \mathbb{P}_0(M_{m_n} > m\sqrt{2\ \log(2)}) < \infty$

— $\sum_{n\in\mathbb{N}} \mathbb{P}_1(M_{m_n} > m\sqrt{2\ \log(2)}) = \infty$

On vérifie que la série $(m_n)_{n\in\mathbb{N}}$ où $m_n = e^{n!}$ convient, c'est celle qu'on utilisera plus tard. Le lemme de Borel-Cantelli nous indique que $\mathbb{P}_0(M_{m_n} > m_n\sqrt{2\ \log(2)}$ pour une infinité de $n) = 0$ et $\mathbb{P}_1(M_{m_n} > m_n\sqrt{2\ \log(2)}$ pour une infinité de $n) = 1$. Le test que l'on va prendre va donc compter le nombre de fois où $M_{m_n} > m_n\sqrt{2\ \log(2)}$.

Pour cela, on a besoin d'un peu plus de précison quant à la fréquence minimale à laquelle $M_{m_n} > m_n\sqrt{2\ \log(2)}$ sous l'hypothèse 1. On remarque donc que pour tout $p \in \mathcal{C}_m$:

$\liminf_{k\to\infty} \frac{1}{k}\ \sum_{n=0}^{k}\ \mathbf{1}_{X_p(m_n) > m_n\sqrt{2\ \log(2)}} \geq \frac{1}{2}\ \ \mathbb{P}_{1,p}\ -\ p.s.$

Voici la preuve de ce résulat :

$$\begin{aligned}\frac{1}{k}\sum_{n=0}^{k} \mathbf{1}_{X_p(m_n) > m_n\sqrt{2\ \log(2)}} &\geq \frac{1}{k}\ \sum_{n=0}^{k} (\ \mathbf{1}_{X_p(m_n) > m_n\sqrt{2\ \log(2)}} - \mathbf{1}_{X_p(m_{n-1}) < 0}\ ) \\ &\geq \frac{-1}{k}\ \sum_{n=0}^{k} \mathbf{1}_{X_p(m_{n-1})<0} + \sum_{n=0}^{k} \mathbf{1}_{X_p(m_n) - X_p(m_{n-1}) > m_n\sqrt{2\ \log(2)}}\end{aligned}$$

Or $P_{1,p}(X_p(m_n) < 0) \leq 2^{-m_n}$, donc $\sum_{k\in\mathbb{N}} P_{1,p}(X_p(m_n) < 0) < \infty$.
Alors $\frac{1}{k}\sum_{n=0}^{k} \mathbf{1}_{X_p(m_{n-1})<0} \xrightarrow[k\to\infty]{} 0$ $\mathbb{P}_{1,p}$ $-$ $p.s.$

De plus,
$\mathbb{P}_{1,p}(X_p(m_n) - X_p(m_{n-1}) > m_n\sqrt{2\ \log(2)}) = \mathbb{P}(\mathcal{N}((e^{n!} - e^{(n-1)!})\sqrt{2\ \log(2)},\ e^{n!} - e^{(n-1)!}) > e^{n!}\sqrt{2\ \log(2)})$.
On remarque que l'écart-type qui est de $\sqrt{(e^{n!} - e^{(n-1)!}}) \sim e^{\frac{n!}{2}}$ est bien plus grand que la différence entre la borne et l'espérance qui est de $e^{(n-1)!}$.
Ainsi, $\mathbb{P}_{1,p}(X_p(m_n) - X_p(m_{n-1}) > m_n\sqrt{2\ \log(2)}) \xrightarrow[m\to\infty]{} \frac{1}{2}$. En rassemblant l'ensemble de ces résultats, on en déduit l'inégalité voulue.

On considrère alors le test T_m qui renvoie 0 si pour $0 \leq n \leq m$, Il y a moins de $\frac{k}{4}$ "n", tels que $M_n > m_n\sqrt{2\ \log(2)}$. Les resultats précédent ont permis de montrer le succès de cette série de tests.

Borne supérieure : Pour cette preuve, on va utiliser la seconde expression de R^*. Le but est de montrer que $\mathbb{E}_0\left((1 - L_m)_+\right) \xrightarrow[m\to\infty]{} 0$ (on rappelle que la dépendance en m est dans presque toutes les variables, y compris dans L ici). Les variables L_m n'étant pas définies sur les mêmes espaces, nous allons introduire la notion de martingale et donner les résultats dont nous aurons besoin sans démonstration.

Définiton (Filtration) : Une filtration sur $(\Omega, \mathcal{F})$ est une suite croissante de tribus $(\mathcal{F}_n)_{n\geq 0}$ incluses dans $\mathcal{F}$.

Définition (Martingale) : Soit $(\mathcal{F}_n)_{n\geq 0}$ une filtration. Une suite de variables aléatoires $(X_n)_{n\geq 0}$ est une martingale par rapport à la filtration $(\mathcal{F}_n)_{n\geq 0}$ si pour tout $n \in \mathbb{N}$:

— X_n est $\mathcal{F}_n$ - mesurable ;

— $\mathbb{E}(M_{n+1}|\mathcal{F}_n) = M_n$.

Définiton (Uniforme intégrabilité) : Une suite $(X_n)_{n\geq 0}$ de v.a dans $L^1(\Omega, \mathcal{F})$ est uniformément intégrable si :

$$\lim_{a\to\infty}\left(\sup_{n\geq 0}\mathbb{E}(X_n \mid \mathbf{1}_{\{|X_n|>a\}}\right) = 0$$

Théorème : Soit $(M_n)_{n\geq 0}$ une martingale. Il y a equivalence entre :

— (i) $(M_n)_{n\geq 0}$ est une martingale uniformément intégrable.

— (ii) $(M_n)_{n\geq 0}$ converge p.s et dans L^1 vers une v.a $M_\infty \in L^1$.

On vérifie donc que sous la loi $\mathbb{P}_0$, L_m est une v.a. adaptée à la filtration $\mathcal{F}_m = \{v \,|\, |v| \leq m-1\}$ ($|v|$ est la distance du sommet v à la racine de l'arbre) et on pose $\mathcal{F}_\infty = \bigcup_{m\geq 1}\mathcal{F}_m$ (c'et la plus petite tribu qui contient tous les $\mathcal{F}_m$). On vérifie que la suite de v.a. $(L_m)_{m\geq 1}$ est une martingale. La condition $\mu_m < \sqrt{2\ \log(2)}$ permet d'obtenir l'uniforme convergence de la martingale. Il existe donc $L_\infty \in L^1(\Omega, \mathcal{F})$ telle que $L_m \xrightarrow[m\to\infty]{p.s.} L_\infty$. On a alors :

$$R^*_m = \mathbb{E}_0\left((1 - L_m)_+\right) \xrightarrow[m\to\infty]{} \mathbb{E}_0\left((1 - L_\infty)_+\right)$$

De plus $\mathbb{E}_0\left(L_\infty\right) = 1$. Donc $\mathbb{P}_0(L_\infty = 0) < 1$. Donc,

$$\lim_{m\to\infty} R^*_m > 0.$$

Cela achève la preuve du troisième théorème.

2.2.5 Commentaires

Nous avons vu que les résultats obtenus diffèrent d'une part d'un risque à l'autre et d'autre part, d'un graphe à l'autre. En effet, dans le cas du treillis, le résultat où le risque envisagé est γ_π (où π est la distribution

uniforme sur les chemins) est bien plus fin que le resultat où l'on considère le risque γ. De plus, la géométrie du graphe semble jouer un rôle très important puisque dans le cas de l'arbre binaire, on a montré qu'il y avait un effet de seuil pour le risque γ_π en la valeur $\sqrt{2\ \log(2)}$. De plus, les techniques de démonstration ont chaque fois utilisé les spécificités de la géométrie de chaque graphe.

3 Applications

Voici quelques exemples d'applications de ces résultats et des idées de ce qu'ils pourraient modéliser :

— Contamination d'un réseau d'eau : On pourait imaginer que l'on modélise un réseau de canalisations par un graphe, et que des contrôles de qualité d'eau sont faits à chaque noeud. Il y a une valeur moyenne de ce que l'on cherche à mesurer pour laquelle l'eau est considérée comme saine, puis si elle est contaminée, alors la moyenne est plus basse ou plus élevée. Alors des résultats comme ci-dessus permettrait de vérifier avec grande fiabilité si l'eau à été contaminée.

— Détection d'un signal au milieu de bruit : la modélisation est présentée dans l'article [2].

— Etudes des polymères dirigés en physique statistique : En effet, on étudie des treillis de dimension 2 ou 3 où l'on assossie à chaque noeud une v.a. qui suit une loi normale d'espérance nulle et de variance 1. Cela donne une v.a. globale qu'on note ω. On s'intéresse au comportement de la fonction de partion qui est une fonction de ω : $Z_{m,\beta} = \frac{1}{(2d)^m} \sum_S \prod_{i=0}^m e^{\beta_i \omega_{i,s_i} - \frac{\beta_i^2}{2}}$ (S est l'ensemble des chemins). On cherche à savoir quand est-ce-que la v.a. $Z_{m,\beta}$ tend vers 0 en probabilité. Cela revient à chercher une suite d'événements $(A_m)_{m\geq 0}$ tels que $\mathbb{P}(A_n) \xrightarrow[m\to\infty]{} 1$ et $\mathbb{E}(Z_{m,\beta}\mathbf{1}_{A_m}) \xrightarrow[m\to\infty]{} 0$. Pour tout événement A on note $\mathbb{P}_1(A) = \mathbb{E}(Z_{m,\beta}\mathbf{1}_A)$, on reconnait alors le cadre des résultats présentés, et la recherche d'une suite de tests T_m est equivalente à la recherche d'une suite d'événements $(A_m)_{m\geq 0}$.

4 Bibliographie

[1] L. Addario-Berry, N. Broutin, L. Devroye, G. Lugosi, On combinatorial testing problems. Ann. Stat., Vol. 38 no 5, 2010.

[2] E. Arias-Castro, E.J. Candès, H. Helgason, O. Zeitouni, Searching for a trail of evidence in a maze. Ann. Stat., Vol. 36 no 4, 2008.

[3] G. Lugosi, Lectures on combinatorial statistics. 47th Probability Summer School, Saint- Flour, 2017.

[4] A. Joulin, Martingales et Applications.

5 Illustrations

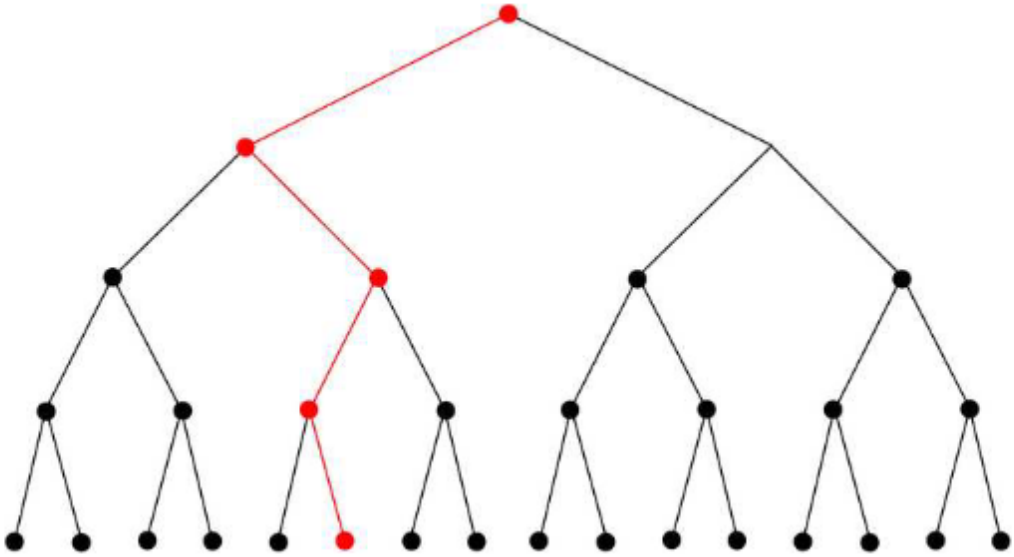

FIGURE 1 : Un chemin dans un arbre binaire.

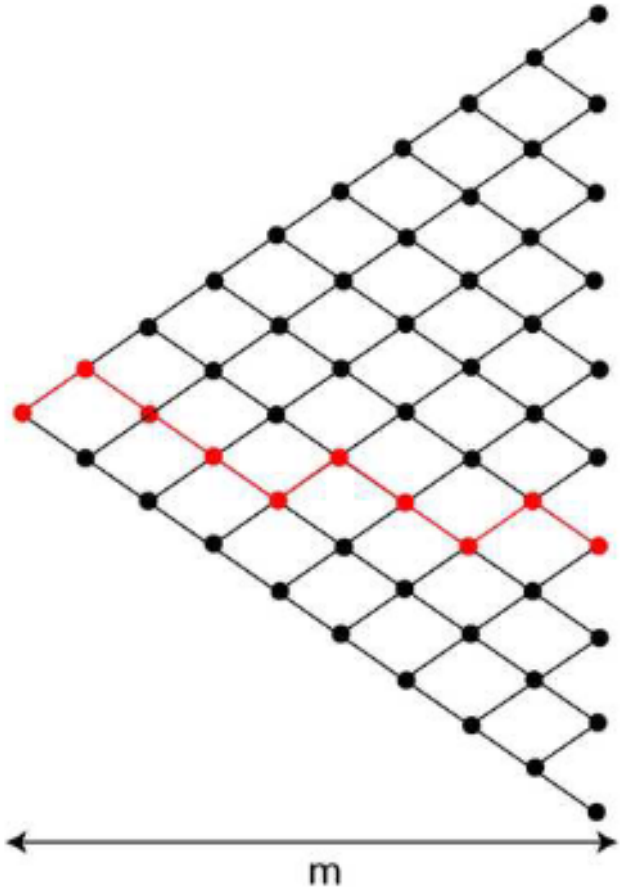

FIGURE 2 : Un chemin dans un treillis.