Correction - AP 9

Graphes aléatoires d'Erdös-Renyi
Distance en variation et couplage

Exercice 1 (Graphes aléatoires d'Erdös-Renyi - ESSEC 2022 sujet 0)

1. Le nombre maximal d'arêtes correspond au nombre total de paires de sommets soit :

$$\binom{n}{2} = \frac{n(n-1)}{2}.$$

2. Un exemple de fonction qui convient :

```python
def listAdj(S,p):
    L=[[] for k in range(len(S))]
    for i in range(len(S)-1):
        for j in range(i+1,len(S)):
            if rd.random()<p:
                L[i].append(S[j])
                L[j].append(S[i])
    return(L)
```

3. On a :

$$D_k = \sum_{i=1}^{k-1} T_{i,k} + \sum_{i=k+1}^{n} T_{k,i}.$$

Donc D_k est la somme de $(n-1)$ variables de Bernoulli indépendantes de paramètre p d'où D_k suit la loi binomiale de paramètres $n-1$ et p.

4. (a) Comme X_k vaut 1 si le sommet est isolé, 0 sinon, $\displaystyle\sum_{k=1}^{n} X_k$ est égal au nombre de sommets isolés dans le graphe, c'est-à-dire Z_n.

Par indépendance des $T_{i,k} \hookrightarrow \mathcal{B}(p)$,

$$E(X_k) = P(X_k = 1) = P(D_k = 0) = P\left(\bigcap_{i\neq k}(T_{i,k} = 0)\right) = (1-p)^{n-1}$$

Par linéarité de l'espérance, $E(Z_n) = n(1-p)^{n-1}$.

(b) On a :

$$Z_n^2 = \left(\sum_{k=1}^{n} X_k\right)^2 = \sum_{1\leq i,j\leq n} X_iX_j = \sum_{1\leq i=j\leq n} X_iX_j + \sum_{1\leq i<j\leq n} X_iX_j + \sum_{1\leq j<i\leq n} X_iX_j.$$

Or $X_i^2 = X_i$ (car X_i est une variable de Bernoulli) et $\displaystyle\sum_{1\leq j<i\leq n} X_iX_j = \sum_{1\leq i<j\leq n} X_iX_j$. Donc :

$$Z_n^2 = \sum_{k=1}^{n} X_k + 2\sum_{1\leq i<j\leq n} X_iX_j.$$

(c) On a pour $i < j$, $(X_i = 1) \cap (X_j = 1)$ est réalisé si et seulement si les événements $(T_{k,i} = 0)$ pour $k < i$, $(T_{i,k} = 0)$ pour $k > i$, $(T_{k,j} = 0)$ pour $k < j$ et $k \neq i$, $(T_{j,k} = 0)$ pour $k > j$ sont réalisés.

Or ces événements sont au nombre de $(n-1) + (n-2) = 2n-3$, indépendants et de probabilité p. D'où $P((X_i = 1) \cap (X_j = 1)) = (1-p)^{2n-3}$.

Par linéarité de l'espérance avec la question 4.(b) :

$$E(Z_n^2) = n(1-p)^{n-1} + 2 \sum_{1 \leq i < j \leq n} E(X_i X_j).$$

Or $E(X_i X_j) = P(X_i X_j = 1) = P((X_i = 1) \cap (X_j = 1)) = (1-p)^{2n-3}$.

Comme il y a $\dfrac{n(n-1)}{2}$ couples (i, j) tels que $1 \leq i < j \leq n$, alors on a bien :

$$E(Z_n^2) = n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3}.$$

5. (a) Voici la fonction demandée :

```python
def Z(lst):
    c = 0
    for k in range(len(lst)):
        if len(lst[k]) == 0:
            c = c+1
    return c
```

(b) Ce scripte utilise la loi faible des grands nombres qui permet d'affirmer que la fréquence empirique de réalisation d'un événement (ici $(Z_n = 0)$), lors d'une répétition d'un grand nombre d'expériences identiques liées à la réalisation de cet événement et indépendantes, est proche de sa probabilité. On peut conjecturer que, lorsque n est grand, $P(Z_n = 0)$ est proche de 1 pour $c > 1$ et de 0 pour $c < 1$.

6. (a) On remarque que $(1-p_n)^{n-1} \sim (1-p_n)^n$ car $1 - p_n \underset{n\to+\infty}{\longrightarrow} 1$. De plus,

$$(1-p_n)^n = \exp\left(n \ln\left(1 - c\frac{\ln(n)}{n}\right)\right).$$

Comme $\dfrac{\ln(n)}{n} \underset{n\to+\infty}{\longrightarrow} 0$ par croissance comparée,

$$n \ln\left(1 - c\frac{\ln(n)}{n}\right) = n\left(-c\frac{\ln(n)}{n} - c^2\frac{\ln(n)^2}{n^2} + o\left(\frac{\ln(n)^2}{n^2}\right)\right) = -c\ln(n) + o(1).$$

Donc :

$$(1-p_n)^n = \exp(-c\ln(n) + o(1)) = n^{-c} \times \underbrace{e^{o(1)}}_{\underset{n\to+\infty}{\longrightarrow} 1} \sim n^{-c}.$$

(b) Si $c > 1$, $E(Z_n) \sim n^{1-c} \underset{n\to+\infty}{\longrightarrow} 0$.

D'après Markov, $P(Z_n \geq 1) \leq E(Z_n)$ donc par encadrement $P(Z_n \geq 1) \underset{n\to+\infty}{\longrightarrow} 0$.

Donc $P(Z_n = 0) = 1 - P(Z_n \geq 1) \underset{n\to+\infty}{\longrightarrow} 1$.

(c) Appliquons l'inégalité de Bienaymé-Tchébichev à Z_n pour $\varepsilon = E(Z_n)$:

$$P(|Z_n - E(Z_n)| \geq E(Z_n)) \leq \frac{V(Z_n)}{E(Z_n)^2}.$$

Or si $(Z_n = 0)$ est réalisé, alors $(|Z_n - E(Z_n)| \geq E(Z_n))$ aussi. Par croissance de P, on a donc :

$$P(Z_n = 0) \leq P(|Z_n - E(Z_n)| \geq E(Z_n)) \leq \frac{V(Z_n)}{E(Z_n)^2}.$$

Avec K-H puis les questions 4.(a) et 4.(c), on a :

$$\begin{aligned}
\frac{V(Z_n)}{E(Z_n)^2} &= \frac{E(Z_n^2) - E(Z_n)^2}{E(Z_n)^2} = \frac{E(Z_n^2)}{E(Z_n)^2} - 1 \\
&= \frac{n(1-p_n)^{n-1} + n(n-1)(1-p_n)^{2n-3}}{n^2(1-p_n)^{2n-2}} - 1 \\
&= \frac{1}{n(1-p_n)^{n-1}} + \frac{n-1}{n(1-p_n)} - 1.
\end{aligned}$$

Or $\dfrac{n-1}{n(1-p_n)} \sim \dfrac{n}{n \times 1} = 1 \underset{n\to+\infty}{\longrightarrow} 1$ et $\dfrac{1}{n(1-p_n)^{n-1}} \sim \dfrac{1}{n \times n^{-c}} = \dfrac{1}{n^{1-c}} \underset{n\to+\infty}{\longrightarrow} 0$ si $c < 1$.
Par encadrement, $P(Z_n = 0) \underset{n\to+\infty}{\longrightarrow} 0$

(d) On a confirmation de la conjecture.

Exercice 2 (Distance en variation et couplage - ESSEC II 2006)

1. Lorsque $\mathcal{K} = \{0;1\}$, on a $D(P,Q) = \dfrac{1}{2}(|p_0 - q_0| + |p_1 - q_1|)$. Comme $p_0 + p_1 = 1$, on a donc $p_0 = 1 - p_1$ et de même $q_0 = 1 - q_1$.

   Donc $D(P,Q) = \frac{1}{2}(|q_1 - p_1| + |p_1 - q_1|) = |p_1 - q_1|$ (car $|-x| = |x|$ pour tout x réel).

2. Supposons que $\mathcal{K} = \mathbb{N}$. Pour tout $k \in \mathcal{K}$, on a :

   $$0 \leq |p_k - q_k| \leq |p_k| + |q_k| = p_k + q_k \text{ (d'après l'inégalité triangulaire).}$$

   Comme les séries de terme général p_k et q_k convergent, celle de terme général $p_k + q_k$ également.

   Et par comparaison de série à termes positifs, la série de terme général $|p_k - q_k|$ converge.

3. Pour tout $A \in \mathcal{A}$, on a $0 \leq P(A) \leq 1$ et $0 \leq Q(A) \leq 1$.

   Donc $-1 \leq P(A) - Q(A) \leq 1$ et donc $|P(A) - Q(A)| \leq 1$.

   On a bien $|P(A) - Q(A)| \in [0,1]$.

4. Pour tout $A \in \mathcal{A}$, $P(A) = \sum\limits_{k\in A} p_k$ et $\sum\limits_{k\in A} p_k = 1 - \sum\limits_{k\in \bar{A}} p_k$ (et de même pour Q).

   Donc $P(A) - Q(A) = \sum\limits_{k\in A}(p_k - q_k)$ d'une part.

   Et $P(A) - Q(A) = 1 - \sum\limits_{k\in \bar{A}} p_k - 1 + \sum\limits_{k\in \bar{A}} q_k = \sum\limits_{k\in \bar{A}}(q_k - p_k) = -\sum\limits_{k\in \bar{A}}(p_k - q_k)$ d'autre part.

   Donc :

   $$2|P(A) - Q(A)| = |P(A) - Q(A)| + |P(A) - Q(A)| = \left|\sum_{k\in A}(p_k - q_k)\right| + \left|\sum_{k\in \bar{A}}(p_k - q_k)\right|.$$

5. On a : $D(P,Q) = \dfrac{1}{2}\sum\limits_{k\in\mathcal{K}} |p_k - q_k|$ et $\mathcal{K} = A \cup \bar{A}$ (union disjointe).

   Donc $D(P,Q) = \dfrac{1}{2}\left(\sum\limits_{k\in A}|p_k - q_k| + \sum\limits_{k\in \bar{A}}|p_k - q_k|\right)$.

   Avec l'inégalité triangulaire, $\sum\limits_{k\in A}|p_k - q_k| \geq \left|\sum\limits_{k\in A}(p_k - q_k)\right|$ et $\sum\limits_{k\in \bar{A}}|p_k - q_k| \geq \left|\sum\limits_{k\in \bar{A}}(p_k - q_k)\right|$.

Alors, en utilisant la question précédente,

$$2D(P,Q) = \sum_{k\in A} |p_k - q_k| + \sum_{k\in \bar{A}} |p_k - q_k| \geq \left|\sum_{k\in A}(p_k - q_k)\right| + \left|\sum_{k\in \bar{A}}(p_k - q_k)\right| = 2\,|P(A) - Q(A)|\,.$$

6. Soit $A = \{k \in \mathcal{K} \mid q_k \geq p_k\}$. Pour tout $k \in A$, $|p_k - q_k| = -(p_k - q_k)$.

   Et pour $k \in \bar{A}$, $|p_k - q_k| = p_k - q_k$ (car dans $\bar{A}$, $q_k \geq p_k$ est faux).

   Donc :

$$\begin{aligned} D(P,Q) &= \frac{1}{2}\left(\sum_{k\in A} |p_k - q_k| + \sum_{k\in \bar{A}} |p_k - q_k|\right) = \frac{1}{2}\left(-\sum_{k\in A}(p_k - q_k) + \sum_{k\in \bar{A}}(p_k - q_k)\right) \\ &= \frac{1}{2}\left(\left|\sum_{k\in A}(p_k - q_k)\right| + \left|\sum_{k\in \bar{A}}(p_k - q_k)\right|\right) = |P(A) - Q(A)| \end{aligned}$$

   car $\sum\limits_{k\in A} p_k - q_k \leq 0$ et $\sum\limits_{k\in \bar{A}} p_k - q_k \geq 0$.

7. On réutilise la question précédente :

   Avec $A = \{k \in \mathcal{K} \mid q_k \geq p_k\}$, on a : $D(P,Q) = |P(A) - Q(A)|$.

   Or pour $k \in A$, $\min(p_k, q_k) = p_k$ et pour $k \in \bar{A}$, $\min(p_k, q_k) = q_k$.

   Donc $1 - \sum\limits_{k\in \mathcal{K}} \min(p_k, q_k) = 1 - \sum\limits_{k\in A} p_k - \sum\limits_{k\in \bar{A}} q_k$

   Comme $1 - \sum\limits_{k\in A} p_k = P(\bar{A})$ et $\sum\limits_{k\in \bar{A}} q_k = Q(\bar{A})$, on obtient :

$$1 - \sum_{k\in \mathcal{K}} \min(p_k, q_k) = P(\bar{A}) - Q(\bar{A}) = (1 - P(A)) - (1 - Q(A)) = Q(A) - P(A)\,.$$

   Et comme ici $Q(A) - P(A) = \sum\limits_{k\in A} (q_k - p_k) \geq 0$, on a :

$$Q(A) - P(A) = |P(A) - Q(A)| = D(P,Q).$$

   Finalement, on a :

$$D(P,Q) = 1 - \sum_{k\in \mathcal{K}} \min(p_k, q_k).$$

8. L'événement $(X = Y)$ est plus facile à décomposer que $(X \neq Y)$, on s'intéresse à l'événement contraire : $\mathbf{P}(X \neq Y) = 1 - \mathbf{P}(X = Y)$.

   On décompose $(X = Y) = \bigcup\limits_{k\in \mathcal{K}} ((X = k) \cap (Y = k))$ (union disjointe).

   Donc $\mathbf{P}(X = Y) = \sum\limits_{k\in \mathcal{K}} \mathbf{P}((X = k) \cap (Y = k))$.

   Or :

   - $\mathbf{P}((X = k) \cap (Y = k)) \leq \mathbf{P}(X = k) = p_k$ car $((X = k) \cap (Y = k)) \subset (X = k)$.
   - $\mathbf{P}((X = k) \cap (Y = k)) \leq \mathbf{P}(Y = k) = q_k$ car $((X = k) \cap (Y = k)) \subset (Y = k)$.

   Donc $\mathbf{P}((X = k) \cap (Y = k))$ est inférieur au plus petit des deux :

$$\mathbf{P}((X = k) \cap (Y = k)) \leq \min(p_k, q_k)$$

   Finalement, en utilisant la question précédente :

$$\mathbf{P}(X \neq Y) = 1 - \mathbf{P}(X = Y) = 1 - \sum_{k\in \mathcal{K}} \mathbf{P}((X = k) \cap (Y = k)) \geq 1 - \sum_{k\in \mathcal{K}} \min(p_k, q_k) = D(P,Q)\,.$$

9. Comme $Y_1, \cdots, Y_n$ sont des variables aléatoires indépendantes et de même loi de Poisson de paramètre λ/n, on sait (c'est du cours !) que $\sum\limits_{i=1}^{n} Y_i$ suit une loi de Poisson de paramètre la somme des paramètres, c'est-à-dire λ.

10. f est définie, continue et dérivable sur $\mathbb{R}$ et $f'(x) = e^x - (1-x)e^x = xe^x$.

    Sur $[0,1]$, $f' \geq 0$ donc f est croissante. Comme $f(0) = 0$ et $f(1) = 1$, on a donc pour tout $x \in [0,1]$, $f(x) \in [0,1]$.

11. Par définition, les seules valeurs de X_i étant 0 et 1, X_i suit une loi de Bernoulli.

    Reste à déterminer le paramètre $\mathbf{P}(X_i = 1)$. On passe par $\mathbf{P}(X_i = 0)$ donné par l'énoncé :

$$\begin{aligned} \mathbf{P}(X_i = 0) &= \mathbf{P}(U_i = 0 \cap Y_i = 0) \overset{\text{indép.}}{=} \mathbf{P}(U_i = 0)\,\mathbf{P}(Y_i = 0) = (1 - f(\lambda/n))\,e^{-\lambda/n} \\ &= \left[1 - \left(1 - \left(1 - \frac{\lambda}{n}\right)e^{\lambda/n}\right)\right]e^{-\lambda/n} = 1 - \frac{\lambda}{n}. \end{aligned}$$

    Donc $\mathbf{P}(X_i = 1) = 1 - \mathbf{P}(X_i = 0) = \dfrac{\lambda}{n}$ et X_i suit une loi de Bernoulli de paramètre λ/n.

    Par indépendance des X_i, $\sum\limits_{i=1}^{n} X_i$ suit une loi de binomiale de paramètres $(n, \lambda/n)$ (toujours d'après le cours !).

12. Comme les valeurs de X_i sont 0 et 1,

$$\begin{aligned} \mathbf{P}(X_i = Y_i) &= \mathbf{P}(((X_i = 0) \cap (Y_i = 0)) \cup ((X_i = 1) \cap (Y_i = 1))) \\ &= \mathbf{P}((X_i = 0) \cap (Y_i = 0)) + \mathbf{P}((X_i = 1) \cap (Y_i = 1)) \quad \text{(par incomp.)} \\ &= \mathbf{P}(X_i = 0) + \mathbf{P}(Y_i = 1) \quad (\text{car } (X_i = 0) \subset (Y_i = 0) \text{ et } (Y_i = 1) \subset (X_i = 1)) \\ &= \mathbf{P}(X_i = 0) + \mathbf{P}(Y_i = 1) = 1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^{-\lambda/n} \end{aligned}$$

    Donc $\mathbf{P}(X_i \neq Y_i) = \dfrac{\lambda}{n} - \dfrac{\lambda}{n}e^{-\lambda/n} = \dfrac{\lambda}{n}\left(1 - e^{-\lambda/n}\right)$.

    Reste à montrer que $1 - e^{-x} \leq x$ pour tout $x \geq 0$. Soit $g(x) = x - 1 + e^{-x}$ définie, continue et dérivable sur $\mathbb{R}$ et $g'(x) = 1 - e^{-x} \geq 0$ pour $x \geq 0$. Donc g est croissante sur $\mathbb{R}^+$ et comme $g(0) = 0$, $g \geq 0$ sur $\mathbb{R}^+$.

    Donc $1 - e^{-\lambda/n} \leq \dfrac{\lambda}{n}$ et $\mathbf{P}(X_i \neq Y_i) \leq \dfrac{\lambda^2}{n^2}$.

13. Si $\sum\limits_{i=1}^{n} X_i \neq \sum\limits_{i=1}^{n} Y_i$, alors il existe au moins un $i \in [\![1, n]\!]$ tel que $X_i \neq Y_i$.

    Donc $\left(\sum\limits_{i=1}^{n} X_i \neq \sum\limits_{i=1}^{n} Y_i\right) \subset \left(\bigcup\limits_{i=1}^{n} (X_i \neq Y_i)\right)$ et $\mathbf{P}\left(\sum\limits_{i=1}^{n} X_i \neq \sum\limits_{i=1}^{n} Y_i\right) \leq \mathbf{P}\left(\bigcup\limits_{i=1}^{n} (X_i \neq Y_i)\right)$.

14. $\bigcup\limits_{i=1}^{n} (X_i \neq Y_i)$ n'est pas une réunion d'événements incompatibles.

    Mais, comme $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \leq \mathbf{P}(A) + \mathbf{P}(B)$ pour tous événements A et B et en généralisant cette formule par récurrence, on obtient :

$$\mathbf{P}\left(\bigcup_{i=1}^{n} (X_i \neq Y_i)\right) \leq \sum_{i=1}^{n} \mathbf{P}(X_i \neq Y_i)$$

    et donc

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i \neq \sum_{i=}^{n} Y_i\right) \leq \sum_{i=1}^{n} \mathbf{P}\left(X_i \neq Y_i\right) \leq \sum_{i=1}^{n} \frac{\lambda^2}{n^2} = \frac{\lambda^2}{n}.$$

Ici, la distance entre deux lois est celle entre les probablités définies par ces lois :

- P définie par $P(\{k\}) = \mathbf{P}(X = k)$ pour tout $k \in \mathbb{N}$ où $X \hookrightarrow \mathcal{B}(n, \lambda/n)$,
- Q définie par $Q(\{k\}) = \mathbf{P}(Y = k)$ pour tout $k \in \mathbb{N}$ où $Y \hookrightarrow \mathcal{P}(\lambda)$.

Soient $X = \sum\limits_{i=1}^{n} X_i \hookrightarrow \mathcal{B}(n, \lambda/n)$ et $Y = \sum\limits_{i=1}^{n} Y_i \hookrightarrow \mathcal{P}(\lambda)$

On a, avec la question 8 de la partie 1 et le calcul précédent, que $D(P, Q) \leq \mathbf{P}(X \neq Y) \leq \dfrac{\lambda^2}{n}$.

Finalement, $D(\mathcal{B}(n, \lambda/n), \mathcal{P}(\lambda)) \leq \dfrac{\lambda^2}{n}$.

15. Quand n tend vers l'infini, cette distance entre les deux lois tend vers 0.

    Cette distance étant la somme des écarts entre les probabilités données par les deux lois, l'erreur que l'on fera en employant $\mathcal{P}(\lambda)$ au lieu de $\mathcal{B}(n, \lambda/n)$ tendra vers 0.