Estimation de la taille d'un graphe par marches aléatoires

Matthieu Dinot, Dorian Boully

Préliminaire : spectre des matrices à coefficients positifs

Pour une matrice réelle X de taille quelconque (en particulier X peut être un vecteur de $\mathbb{R}^N$ ou une matrice (N,N)), on note $X \geqslant 0$ lorsque tous ses coefficients sont positifs ou nuls, et $X > 0$ lorsque tous ses coefficients sont strictement positifs.

Théorème 1 (Perron–Frobenius)**.** *Soit $A > 0 \in \mathcal{M}_N(\mathbb{R})$. Alors*

- *A possède une valeur propre réelle $\lambda > 0$ telle que toutes les autres valeurs propres de A sont de module strictement inférieur à λ. On dit que λ est dominante*
- *Il existe un vecteur propre $X > 0$ associé à la valeur propre λ.*
- *λ est une valeur propre simple de A, c'est à dire que sa multiplicité en tant que racine du polynôme caractéristique de A vaut 1. En particulier, le sous espace propre associé à λ est de dimension 1.*

Ce théorème nous est utile pour étudier le spectre des matrices stochastiques. Une matrice $A \in \mathcal{M}_N(\mathbb{R})$ est dite stochastique lorsque $A \geqslant 0$ et

$$\forall i \in \{1, \ldots, N\}, \quad \sum_{j=1}^{N} A(i,j) = 1.$$

Corollaire 1. *Soit A une matrice stochastique. On suppose qu'il existe un entier k tel que $A^k > 0$ (A est alors dite régulière). Alors 1 est valeur propre simple et dominante de A.*

Il est possible d'appliquer ces résultats à l'étude d'une marche aléatoire sur un graphe $G = (V,E)$ de taille N (orienté ou non). On peut définir une telle marche aléatoire à l'aide d'une matrice de transition P telle que le coefficient $P(i,j)$ est la probabilité de passer du sommet j au sommet i. On impose la contrainte $P(i,j) \neq 0$ si et seulement si $(j,i) \in E$. Ainsi P représente bien une marche aléatoire sur le graphe G. On remarque que $P^\top$ est stochastique et que si π_t représente le vecteur probabilité de présence à l'instant t, on a

$$\pi_{t+1} = P\pi_t.$$

Regardons maintenant sous quelles conditions il est possible d'appliquer le corollaire 1 à une matrice de transition P. Remarquons pour cela que $P^k(i,j)$ est non nul si et seulement si il existe un chemin de longueur k de j vers i dans le graphe G. En effet, on a

$$P^k(i,j) = \sum_{l_1,\ldots,l_{k-1}} P(i,l_1)P(l_1,l_2)\cdots P(l_{k-1},j).$$

Chaque terme de cette somme est non nul lorsque $j \to l_{k-1} \to \cdots \to l_1 \to i$ est un chemin dans le graphe G. Comme $P \geqslant 0$, $P^k(i,j)$ est non nul si et seulement si au moins un terme de la somme est non nul, ce qui permet de conclure. Pour que P soit régulière, il faut donc que G soit connexe, mais cela ne suffit pas. Par exemple, si l'on prend un graphe biparti, les chemins de longueur paire ont un départ et une arrivée dans la même « partie » du graphe, donc P^k n'aura jamais tous ses coefficients non nuls. D'ailleurs, comme indiqué dans l'énoncé (**S1**), si P représente une marche aléatoire sur un graphe biparti, où le choix du sommet au temps $t+1$ se fait uniformément parmi les voisins du sommet au temps t, alors -1 est dans le spectre de P. Par contre, une matrice de transition P sur un graphe connexe tel que tout sommet est relié à lui même est régulière. En effet, comme le graphe est connexe, il existe un entier k tel que pour tout $(i,j) \in V^2$, il existe un chemin de j vers i de longueur au plus k. En ajoutant autant de fois que nécessaire l'arête (j,j) au début du chemin, on trouve un chemin de longueur exactement k. Justement, le fait de considérer des marches aléatoires paresseuses revient à ajouter toutes les arêtes (j,j), $j \in V$ au graphe étudié, ce qui ne change bien entendu pas sa taille et permet de garantir que 1 est valeur propre dominante et simple.

1 Partie théorique

Les questions suivantes restent valables dans le cas plus général d'une marche aléatoire sur un graphe G non orienté connexe quelconque (pas forcément régulier) représentée par une matrice de transition (symétrique) P dont 1 est valeur propre simple et dominante.

T1. Soient $s \in \{0, \dots, \tau - 1\}$ et $i \in V$. On a :

$$\begin{aligned}
\mathbb{P}(X^t_{s+1} = i) &= \sum_{j \in V} \mathbb{P}((X^t_{s+1} = i) \cap (X^t_s = j)) \\
&= \sum_{j \in V} \frac{1}{d} \mathbb{1}_E(i,j) \mathbb{P}(X^t_s = j) \\
&= \sum_{j \in V} P_{ij} \mathbb{P}(X^t_s = j).
\end{aligned}$$

Ainsi $\big(\mathbb{P}(X^t_{s+1} = i)\big)_{i \in V} = P \cdot (X^t_s = i)_{i \in V}$. Une récurrence immédiate montre alors que

$$\pi = P^\tau \cdot (\delta_{i,i_0})_{i \in V} \tag{1}$$

où δ désigne le symbole de Krönecker.

T2. Le théorème spectral assure l'existence d'une matrice orthogonale O telle que

$$P = {}^tO \, \mathrm{Diag}(1, \lambda_2, \dots, \lambda_N) \, O.$$

Vu les valeurs propres de P, on voit que, pour une norme quelconque sur les matrices (N,N) (en particulier la norme subordonnée à $\|\cdot\|_2$) :

$$P^\tau = {}^tO \Delta^\tau O \xrightarrow[\tau \to +\infty]{} P^\infty := {}^tO \, \mathrm{Diag}(1, 0, \dots, 0) \, O$$

car la convergence a lieu coefficient par coefficient. La convergence de $(P^\tau)_{\tau \geqslant 1}$ pour la norme subordonnée à $\|\cdot\|_2$ entraine la convergence simple de la suite d'applications linéaires associées au sens de $\|\cdot\|_2$ vers la même limite. On peut conclure de deux manières :

— Via un calcul explicite :

$$\pi = P^\tau (\delta_{i,i_0})_{i\in V} \xrightarrow[\tau\to+\infty]{} P^\infty(\delta_{i,i_0})_{i\in V} = O_{1i_0}\,{}^tO\,{}^t(1,0,\dots,0) = O_{1i_0} O^{-1}\,{}^t(1,0,\dots,0).$$

Or $O^{-1}\,{}^t(1,0,\dots,0)$ est le vecteur propre normalisé de P de valeur propre 1 (il y en a un seul car la valeur propre 1 était simple). On vérifie facilement qu'il s'agit de $N^{-1/2}\,{}^t(1,\dots,1)$. Enfin, on a $O_{1i_0} = N^{-1/2}$ car c'est un coefficient de la première ligne de O, donc de la première colonne de ${}^tO = O^{-1}$, qui n'est autre que la matrice d'une base orthonormée de diagonalisation de P dans la base canonique. On conclut comme attendu que

$$\lim_{\tau\to+\infty} \pi = N^{-1}\,{}^t(1,\dots,1). \tag{2}$$

— En remarquant que $\pi^\infty := \lim_{\tau\to+\infty} \pi$ est le vecteur propre de P associé à la valeur propre 1 et dont la somme des coefficients vaut 1. En effet, pour tout τ, la somme des coefficients de π vaut 1, ce qui reste vrai à la limite (en particulier $\pi^\infty \neq 0$), et on a :

$$P\pi^\infty = P \lim_{\tau\to+\infty} P^\tau(\delta_{i,i_0})_{i\in V} = \lim_{\tau\to+\infty} P^{\tau+1}(\delta_{i,i_0})_{i\in V} = \pi^\infty.$$

Encore une fois, on trouve la loi uniforme sur V.

T3. Notons

$$\mathcal{A}_m = \{(y_1,\dots,y_m)\in V^m \mid \operatorname{Card}\{y_1,\dots,y_{m-1}\} = \operatorname{Card}\{y_1,\dots,y_m\} = m-(\ell-1)\},$$

de sorte que

$$(C_{\ell-1} = m) = \bigsqcup_{\mathbf{y}\in\mathcal{A}_m} (\mathbf{Y} = \mathbf{y}).$$

Notons aussi $\mathcal{B}_n(U)$ l'ensemble des n-uplets injectifs à valeurs dans $V\setminus U$. On a :

$$\begin{aligned}
\mathbb{P}((C_\ell - C_{\ell-1} > n)\cap(C_{\ell-1}=m)) &= \sum_{(y_1,\dots,y_n)\in\mathcal{A}_m} \mathbb{P}\left((C_\ell - C_{\ell-1}>n)\cap \bigcap_{1\leqslant t\leqslant m}(Y_t = y_t)\right) \\
&= \sum_{\substack{(y_1,\dots,y_m)\in\mathcal{A}_m\\(y_{m+1},\dots,y_{m+n})\in\mathcal{B}_n(\{y_1,\dots,y_m\})}} \mathbb{P}\left(\bigcap_{1\leqslant t\leqslant m+n}(Y_t=y_t)\right) \\
&= \sum_{\substack{(y_1,\dots,y_m)\in\mathcal{A}_m\\(y_{m+1},\dots,y_{m+n})\in\mathcal{B}_n(\{y_1,\dots,y_m\})}} \prod_{1\leqslant t\leqslant m+n} \mathbb{P}(Y_t=y_t) \qquad (3)\\
&= \sum_{\substack{(y_1,\dots,y_m)\in\mathcal{A}_m\\(y_{m+1},\dots,y_{m+n})\in\mathcal{B}_n(\{y_1,\dots,y_m\})}} \frac{1}{N^{m+n}} \qquad (4)\\
&= \frac{1}{N^{m+n}} \sum_{(y_1,\dots,y_m)\in\mathcal{A}_m} \operatorname{Card}\mathcal{B}_n(\{y_1,\dots,y_m\}) \\
&= \frac{1}{N^{m+n}} \sum_{(y_1,\dots,y_m)\in\mathcal{A}_m} n!\binom{N-(m-(\ell-1))}{n} \\
&= \frac{(N-m+\ell-1)(N-m+\ell-2)\cdots(N-m+\ell-n)}{N^n}\frac{\operatorname{Card}\mathcal{A}_m}{N^m} \\
&= \frac{(N-m+\ell-1)(N-m+\ell-2)\cdots(N-m+\ell-n)}{N^n}\mathbb{P}(C_{\ell-1}=m).
\end{aligned}$$

On trouve comme attendu :

$$\mathbb{P}((C_\ell - C_{\ell-1} > n) \mid (C_{\ell-1} = m)) = \frac{(N-m+\ell-1)(N-m+\ell-2)\cdots(N-m+\ell-n)}{N^n}. \tag{5}$$

On aurait pu démontrer ce fait de manière moins formelle, les idées essentielles étant que

- les Y_t sont i.i.d. et suivent une loi uniforme sur V ;
- le cardinal de $\mathcal{B}_n(U)$ ne dépend que de n et du cardinal de U, mais pas des valeurs de ses éléments.

Si l'on ne fait pas l'approximation de remplacer les variables Y_t par des variables uniformément distribuées sur V, il n'y a pas égalité entre les lignes (3) et (4). Cependant, la question **T2** montre que la ligne (4) est la limite de la ligne (3) lorsque τ tend vers l'infini. Ainsi, pour être tout à fait précis, on a montré que

$$\lim_{\tau\to+\infty} \mathbb{P}((C_\ell - C_{\ell-1} > n) \mid (C_{\ell-1} = m)) = \frac{(N-m+\ell-1)(N-m+\ell-2)\cdots(N-m+\ell-n)}{N^n}.$$

L'approximation est donc raisonnable jusqu'ici.

T4. Nous allons légèrement améliorer le résultat de la première limite pour pouvoir en déduire la seconde. Soient a, b des réels strictement positifs et $(a_N)_{N\geqslant 1}, (b_N)_{N\geqslant 1}$ des suites d'entiers telles que

$$a_N \underset{+\infty}{\sim} aN^{1/2} \quad \text{et} \quad b_N \underset{+\infty}{\sim} bN^{1/2}.$$

D'après la question précédente, on a :

$$\begin{aligned}\mathbb{P}((C_\ell - C_{\ell-1} > b_N) \mid (C_{\ell-1} = a_N)) &= \frac{(N-a_N+\ell-1)(N-a_N+\ell-2)\cdots(N-a_N+\ell-b_N)}{N^{b_N}} \\ &= \frac{(N-(a_N-(\ell-1)))!}{N^{b_N}(N-b_N-(a_N-(\ell-1)))!}\end{aligned}$$

Posons, pour $N \geqslant 1$

$$\begin{aligned} u_N &= N - (a_N - (\ell-1)) \\ v_N &= N - b_N - (a_N - (\ell-1)). \end{aligned}$$

D'après la formule de Stirling, on a :

$$\begin{aligned}\mathbb{P}((C_\ell - C_{\ell-1} > b_N) \mid (C_{\ell-1} = a_N)) &\underset{+\infty}{\sim} \frac{\sqrt{2\pi u_N}}{\sqrt{2\pi v_N}} \frac{\exp\left[u_N \log u_N - u_N\right]}{\exp\left[b_N \log N + v_N \log v_N - v_N\right]} \\ &\sim \exp\left[u_N \log u_N + v_N - u_N - b_N \log N - v_N \log v_N\right].\end{aligned}$$

On cherche un développement asymptotique en $o(1)$ de l'argument de l'exponentielle. On procède par étapes :

$$\begin{aligned}\log u_N &= \log N - \frac{a_N - (\ell-1)}{N} - \frac{(a_N-(\ell-1))^2}{2N^2} + o(1/N) \\ &= \log N - \frac{a_N - (\ell-1)}{N} - \frac{(aN^{1/2} + o(N^{1/2}))^2}{2N^2} + o(1/N) \\ &= \log N - \frac{a_N - (\ell-1)}{N} - \frac{a^2}{2N} + o(1/N).\end{aligned}$$

De même,

$$\begin{aligned}
\log v_N &= \log N - \frac{b_N + a_N - (\ell - 1)}{N} - \frac{(b_N + a_N - (\ell-1))^2}{2N^2} + o(1/N) \\
&= \log N - \frac{b_N + a_N - (\ell - 1)}{N} - \frac{(a+b)^2}{2N} + o(1/N).
\end{aligned}$$

Puis

$$\log u_N - \log v_N = \frac{b_N}{N} + \frac{b(2a+b)}{2N} + o(1/N).$$

D'où, en utilisant le fait que $a_N = aN^{1/2} + o(N^{1/2})$ et $b_N = bN^{1/2} + o(N^{1/2})$

$$\begin{aligned}
v_N(\log u_N - \log v_N) &= b_N + \frac{b(2a+b)}{2} - \frac{b_N(b_N + a_N)}{N} + o(1) \\
&= b_N + \frac{b(2a+b)}{2} - b(a+b) + o(1) \\
&= b_N - \frac{b^2}{2} + o(1). \qquad (6)
\end{aligned}$$

De plus, pour la même raison,

$$b_N \log u_N = b_N \log N - ab + o(1). \qquad (7)$$

Enfin, en utilisant les développements asymptotiques précédents (6 et 7)

$$\begin{aligned}
u_N \log u_N + v_N - u_N - b_N \log N - v_N \log v_N &= v_N(\log u_N - \log v_N) + b_N \log u_N - b_N - b_N \log N \\
&= b_N - \frac{b^2}{2} + b_N \log N - ab - b_N - b_N \log N + o(1) \\
&= -ab - \frac{b^2}{2} + o(1).
\end{aligned}$$

Cela montre que

$$\mathbb{P}((C_\ell - C_{\ell-1} > b_N) \mid (C_{\ell-1} = a_N)) \xrightarrow[N \to +\infty]{} e^{-ab-b^2/2}. \qquad (8)$$

Étudions maintenant la suite

$$\mathbb{P}\left(\left(\frac{C_\ell^2 - C_{\ell-1}^2}{2N} > y\right) \middle| \left(\frac{C_{\ell-1}^2}{2N} = \frac{\lfloor (2Nx)^{1/2} \rfloor^2}{2N}\right)\right)$$

où $x, y > 0$. On écrit pour cela

$$\begin{aligned}
&\mathbb{P}\left(\left(\frac{C_\ell^2 - C_{\ell-1}^2}{2N} > y\right) \middle| \left(\frac{C_{\ell-1}^2}{2N} = \frac{\lfloor (2Nx)^{1/2} \rfloor^2}{2N}\right)\right) \\
&= \mathbb{P}\left(C_\ell > \left(\lfloor (2Nx)^{1/2} \rfloor^2 + (2Ny)\right)^{1/2} \middle| C_{\ell-1} = \lfloor (2Nx) \rfloor^{1/2}\right) \\
&= \mathbb{P}\left(C_\ell - C_{\ell-1} > \left(\lfloor (2Nx)^{1/2} \rfloor^2 + (2Ny)\right)^{1/2} - \lfloor (2Nx)^{1/2} \rfloor \middle| C_{\ell-1} = \lfloor (2Nx)^{1/2} \rfloor\right).
\end{aligned}$$

Or,

$$\lfloor (2Nx)^{1/2} \rfloor \underset{+\infty}{\sim} (2Nx)^{1/2}$$

et

$$\left(\lfloor (2Nx)^{1/2}\rfloor^2 + (2Ny)\right)^{1/2} - \lfloor (2Nx)^{1/2}\rfloor = (2N(x+y)+o(N))^{1/2} - (2Nx)^{1/2} + o(N^{1/2})$$
$$= (2N)^{1/2}\left[(x+y)^{1/2} - x^{1/2}\right] + o(N^{1/2}).$$

On peut donc appliquer (8) avec $a = (2x)^{1/2}$ et $b = 2^{1/2}\left[(x+y)^{1/2} - x^{1/2}\right]$. On a

$$ab + b^2/2 = 2\left[x(x+y)\right]^{1/2} - 2x + (x+y) - 2\left[x(x+y)\right]^{1/2} + x = y,$$

ce qui donne

$$\mathbb{P}\left(\left(\frac{C_\ell^2 - C_{\ell-1}^2}{2N} > y\right)\middle|\left(\frac{C_{\ell-1}^2}{2N} = \frac{\lfloor (2Nx)^{1/2}\rfloor^2}{2N}\right)\right) \xrightarrow[N\to+\infty]{} e^{-y}. \tag{9}$$