Projet 1 - Chaînes de Markov

Eléments de processus stochastiques

Floriane Magera
Romain Mormont
Fabrice Servais
Troisième bachelier en sciences de l'ingénieur

Table des matières

Chapitre 1

Question 1

1.1 Etude du modèle de base

2) Avant de calculer la matrice de transition, il est nécessaire de caractériser la marche aléatoire. Autrement dit, il faut définir les poids/probabilités que nous appliquerons aux arêtes du graphe sur lequel le surfeur va évoluer. Nous avons déduit de l'énoncé du projet que les différentes possibilités de quitter un nœud devaient être **équiprobables** et nous utiliserons donc cette hypothèse dans la suite du rapport.

Suite à ce choix, la formation de la matrice de transition est très simple. Si l'on note A la matrice d'adjacence, alors il suffit d'appliquer la formule suivante pour calculer l'élément Q(i,j) :

$$Q(i,j) = A(i,j) \times \frac{1}{n} \sum_{j=1}^{n} A(i,j)$$

Cette formule permet de placer à 0 les éléments de Q représentant une transition impossible et de placer à une certaine probabilité les autres éléments de Q (toute probabilité non nulle d'une ligne de Q étant équiprobable comme attendu).

3) Nous avons choisi un nombre de pas $t = 20$. Le cas où le surfeur démarre aléatoirement sur le graphe est représenté par une distribution initiale π_0 uniforme et le cas où il démarre d'une page fixe est représenté par une distribution initiale π_0 où toutes les probabilités sont nulles sauf celle située l'index correspondant au nœud de départ. L'évolution des probabilités dans les deux cas est donnée sur le Figure 1.1.

La matrice $Q^{(20)}$ obtenue est la suivante :

$$Q^{(20)} = \begin{pmatrix} 0.3751 & 0.1874 & 0.1875 & 0.2500 \\ 0.3751 & 0.1877 & 0.1874 & 0.2499 \\ 0.3749 & 0.1875 & 0.1875 & 0.2500 \\ 0.3749 & 0.1875 & 0.1875 & 0.2500 \end{pmatrix}$$

Si on observe les matrices $Q^{(k)}$ pour $k > 20$, on peut constater que les éléments se stabilisent et que les lignes s'égalisent.

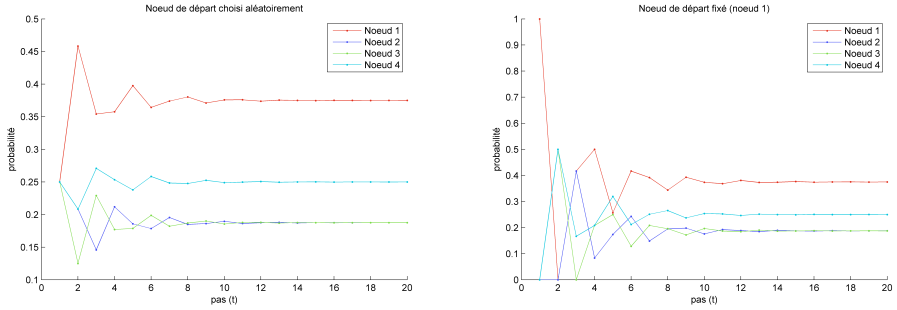


FIGURE 1.1 – Évolution de la distribution de probabilité

4) La distribution stationnaire a été calculée par la méthode des puissances. Nous avons donc multiplié les distributions $\pi^{(k)}$ successives par Q jusqu'à ce que cette distribution ce stabilise. Le critère de stabilisation choisi était le suivant :

$$\max\left(\left|\pi_j^{(k)} - \pi_j^{(k-1)}\pi\right|\right) < \varepsilon$$

où π_j est la $j^{\text{ième}}$ composante du vecteur π. La distribution stationnaire obtenue est donnée ci-dessous :

$$\pi_\infty = \begin{pmatrix} 0.3750 & 0.1875 & 0.1875 & 0.2500 \end{pmatrix}$$

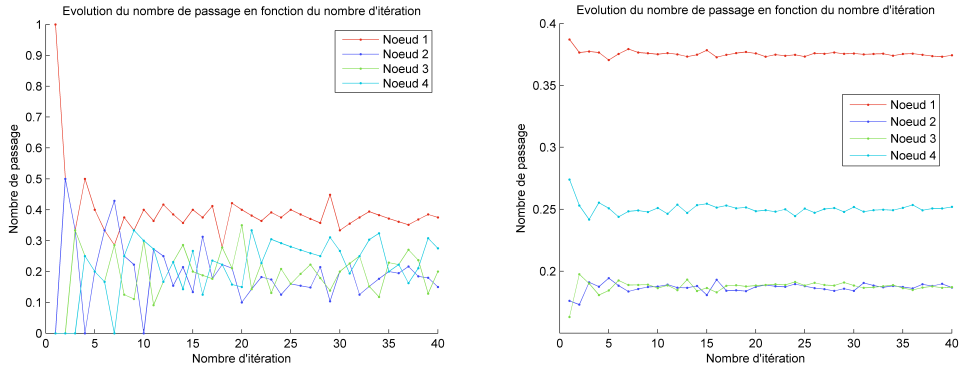
On constate que les lignes de la matrice sont très proches

5) On constate que le noeud 1 possède le meilleur PageRank, suivi des noeuds 2 et 3 à égalité et du noeud 4. On peut expliquer ce classement intuitivement :

- le **noeud 1** possède le plus d'arêtes entrantes donc ayant le plus de chance d'être visité
- le **noeud 3** possède le moins d'arêtes entrantes donc ayant le moins de chance d'être visité
- les **noeuds 2** et **4** possèdent le même nombre intermédiaire (par rapport aux deux autres) d'arêtes entrantes. Le PageRank du noeud 4 est néanmoins plus élevé que celui du noeud 2 puisque le noeud 4 possèdent une arête entrante venant du noeud 1 qui est le plus visité.
- malgré un nombre d'arête entrante plus élevé que pour le noeud 3, le **noeud 2** possède une PageRank égal. Cela est dû au fait que, d'une part, le noeud 3 peut être visité depuis le noeud le plus visité (noeud 1) ce qui améliore son PageRank et, d'autre part, que le noeud 2 ne peut être accéder depuis des noeuds moins visités (noeud 3 et 4) ce qui abaisse son PageRank.

6) Dans un premier temps, nous avons généré une chaîne pour chaque longueur. Le résultat obtenu est donné sur la Figure 1.2(a). On peut déjà observer que les différentes courbes obtenues oscillent autour de leur probabilité stationnaire correspondante. Néanmoins, étant donné la présence d'oscillations, nous avons décidé de refaire l'expérience en générant cette fois-ci 1000 chaînes pour chaque longueur. Nous avons ensuite moyenné les différentes probabilités afin d'obtenir un résultat plus précis (voir Figure 1.2(b)). Les courbes obtenues nous permettent de confirmer les premières observations.

Remarquons aussi que, quelque soit la distribution de départ, la distribution converge vers la distribution stationnaire.



(a) 1 génération (b) 1000 générations

FIGURE 1.2 – Évolution du nombre de passage par un noeud

7)

1.1.1 Analyse des matrices A_2 et A_3

1) Pour pouvoir calculer des éventuelless distribution stationnaires sur base des matrices A_2 et A_3, nous avons calculés les matrices de transition Q_2 et Q_3 :

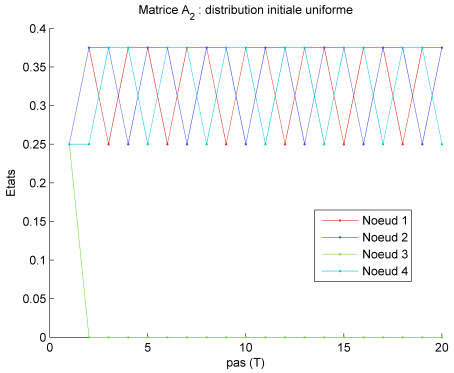
$$Q_2 = \begin{pmatrix} 0 & 1.0000 & 0 & 0 \\ 0 & 0 & 0 & 1.0000 \\ 0.5000 & 0.5000 & 0 & 0 \\ 1.0000 & 0 & 0 & 0 \end{pmatrix} \quad Q_3 = \begin{pmatrix} 0 & 1.0000 & 0 & 0 & 0 \\ 1.0000 & 0 & 0 & 0 & 0 \\ 0 & 0.3333 & 0 & 0.3333 & 0.3333 \\ 0 & 0 & 0 & 0 & 1.0000 \\ 0 & 0 & 0 & 1.0000 & 0 \end{pmatrix}$$

Nous avons ensuite appliqué la méthode des puissance pour obtenir des distributions $\pi^{(k)}$ successives. Le résultat est donné sur la Figure 1.3. On constate qu'aucune distribution stationnaire n'a pu être trouvée étant donné la présence d'**oscillations**.

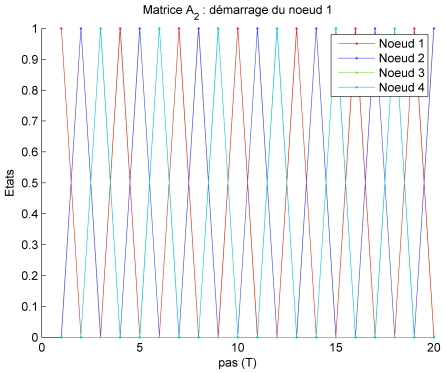
On constate aussi que la **probabilité** d'atteindre une certain nœud **tombe à 0** ou est directement nulle dès la première itération dans les deux situations. Ce phénomène est dû à la présence de *dangling nodes* dans les deux graphes. On peut directement voir la présence de type de nœud sur les matrices de transitions : elle se manifeste par une colonne composée uniquement de probabilités nulles et implique qu'il est impossible de passer d'un nœud quelconque au *dangling node*.

Ces deux phénomènes sont précisément des **limitations** du modèle du surfeur aléatoire simpliste présenté dans cette première partie. D'une part, les oscillations empêchent d'atteindre une distribution stationnaire. D'autre part, la probabilité nulle d'atteindre un nœud existant rend l'éventuelle distribution stationnaire (et donc le PageRank) peut représentative de l'organisation des nœuds puisqu'elle en omet d'en prendre certains en compte.

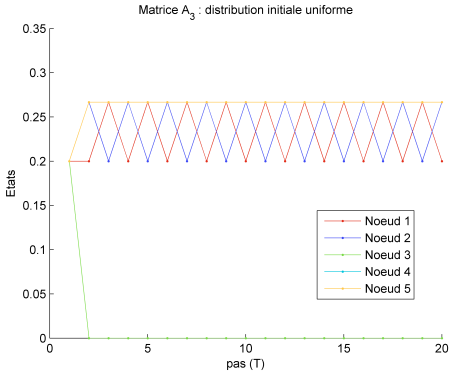
Les **causes de ces phénomènes** sont d'une part la présence d'un cycle infini dans les graphes et la présence de *dangling nodes*. Ces problèmes vont être contournés en introduisant la téléportation dans la section suivante.



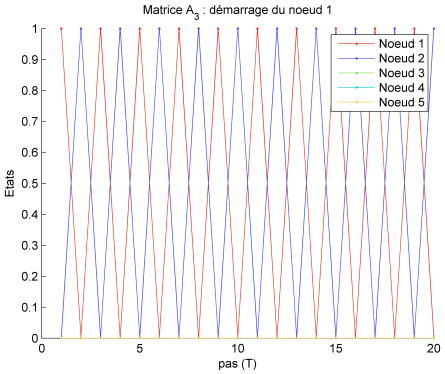
(a) Distribution initiale uniforme



(b) Depuis le noeud 1



(c) Distribution initiale uniforme



(d) Depuis le noeud 1

FIGURE 1.3 – Évolution des distributions de probabilités

2)

1.2 Téléportation

1) La formule utilisée pour calculer la matrice de transition Q_t du modèle du surfeur avec téléportation est la suivante :

$$Q_t = (1 - \alpha)Q' + \alpha\tilde{Q}$$

où Q' et $\tilde{Q}$ sont des matrices de transition et α la probabilité de téléportation.

La première est la matrice de transition du graphe initial auquel on a rajouté des arêtes partant des *dangling nodes*. Elle a été calculée en remplaçant par $\frac{1}{n}$ tous les éléments de la matrice Q situés dans des lignes ne contenant que des 0. Cette valeur $\frac{1}{n}$ a été choisie en considérant une densité de probabilité uniforme entre les différentes arêtes partant des *dangling nodes*.

La seconde est la matrice de transition du graphe complet formé des noeuds du graphe initial. Autrement dit, la matrice de transition représentant la téléportation. Une combinaison linéaire de paramètre α est ensuite appliquée aux deux matrices pour trouver la matrice Q_t.

2) Pour que la distribution stationnaire π_s soit unique, **il faut que la chaîne de Markov soit irréductible**. Autrement dit, il faut que pour tout couple de nœuds (i_1, i_2), il existe une arête les reliant (un probabilité non-nulle de passer de i_1 à i_2). Cette propriété est vérifiée avec le modèle du surfeur modifié puisque la téléportation permet, depuis tout noeud, de se diriger vers un autre noeud tant que $\alpha > 0$.

A partir du moment où $\alpha = 0$, on est plus assuré que chaque paire de nœuds est reliée par une arête et donc que π_s est bien stationnaire.

3) Le classement des sites les plus visités, obtenus à l'aide de la distribution stationnaire, est donné dans la Table 1.1.

N°	PageRank	Page
1	0.0045	http ://purl.org/rss/1.0/modules/content
2	0.0027	http ://www.ulg.ac.be
3	0.0024	http ://ogp.me/ns♯
4	0.0024	http ://www.gre-liege.be
5	0.0023	http ://blog.intelliterwal.net
6	0.0023	http ://www.jalios.com
7	0.0022	http ://www.vmfnet.be
8	0.0022	http ://www.alinoa.be
9	0.0022	http ://www.ulb.ac.be
10	0.0022	http ://www.cedia.ulg.ac.be

TABLE 1.1 – Classement des sites ayant le meilleur PageRank ($\alpha = 0.15$)

4)

1.3 Effet de α

1) Pour prouver que le score PageRank de toute page est au moins $\frac{\alpha}{n}$ (n est le nombre de pages), on peut développer une expression "*explicite*" des éléments de la matrice Q_t en utilisant la formule donnée précédemment :

$$Q_t(i,j) = q_{ij}(1-\alpha) + \frac{1}{n}\alpha$$

où q_{ij} est un élément de la matrice Q'. Connaissant la relation qui lie $\pi^{(k)}$ et $\pi^{(k-1)}$, on a :

$$\begin{aligned}
\pi_j^{(k)} &= \sum_{i=1}^{n} Q_t(i,j)\pi_i^{(k-1)} \\
&= \sum_{i=1}^{n} \left(q_{ij}(1-\alpha) + \frac{\alpha}{n}\right)\pi_i^{(k-1)} \\
&= \sum_{i=1}^{n} q_{ij}(1-\alpha)\pi_i^{(k-1)} + \sum_{i=1}^{n} \frac{\alpha}{n}\pi_i^{(k-1)} \\
&= \frac{\alpha}{n} + \underbrace{(1-\alpha)\sum_{i=1}^{n} q_{ij}\pi_i^{(k-1)}}_{>0}
\end{aligned}$$

Le deuxième terme est inférieur à 1 (et même inférieur à $(1-\frac{\alpha}{n})$ afin de respecter le deuxième axiome de Kolmogorov) et surtout, positif. De ce fait, on peut affirmer que :

$$\pi_j^{(k)} \geq \frac{\alpha}{n}$$

On peut interpréter le cas où $\alpha = 0$ comme le cas où il n'y a pas de téléportation et le cas où $\alpha = 1$ comme le cas où il n'y a que téléportation (le surfeur n'utilise plus les liens). Remarquons les valeurs prises par les distributions dans les deux cas :

$$\pi_{\alpha=0}^{(k)} = \pi^{(k-1)}Q'$$

$$\pi_{j,\alpha=1}^{(k)} = \tfrac{1}{n}, \forall j$$

Dans le deuxième cas, les PageRank de toutes les pages seront égaux et vaudront $\frac{1}{n}$. Afin de vérifier cette affirmation, nous avons calculé la distribution stationnaire pour un $\alpha = 1$ et l'avons mise en parallèle avec la distribution stationnaire pour $\alpha = 0.15$. Le résultat est édifiant (voir Table 1.2, on constate en effet un PageRank uniforme dans le cas où $\alpha = 1$.

α	0.15	1
Moyenne	0.002	0.002
Ecart-type	1.3000e-04	**4.7753e-18**
Min.	0.0019	0.002
Max.	0.0045	0.002

TABLE 1.2 – Statistiques à propos des distributions stationnaires avec $\alpha = 0.15$ et $\alpha = 1$

Chapitre 2

Question 2

2.1 Estimation d'une matrice de transition

2.1.1 Méthode d'estimation

Avant tout, voici les quelques notations que nous utiliserons :
- n, le nombre de noeuds dans le graphe
- X, la trace fournie (chaîne de Markov)
- Q_{est}, la matrice de transition recherchée

$$Q_{est} = \begin{pmatrix} \theta_{11} & \cdots & \theta_{1n} \\ \vdots & \ddots & \vdots \\ \theta_{n1} & \cdots & \theta_{nn} \end{pmatrix}$$

- N, la matrice dont l'élément β_{ij} est le nombre de transition de l'état i à j dans la trace X
- σ_i, la somme de la $i^{\text{ème}}$ ligne de la matrice N
- $\underline{\theta_i}$; le vecteur contenant les probabilités pour passer du nœud i à un autre nœud :

$$\underline{\theta_i} = [\theta_{i1} \cdots \theta_{in}]$$

- $P(X|\underline{\theta_i})$, la probabilité d'observer les départs du nœud i présents dans la trace connaissant $\underline{\theta_i}$

$$P(X|\underline{\theta_i}) = \theta_{i1}^{\beta_{i1}} \times \cdots \times \theta_{i(n-1)}^{\beta_{i(n-1)}} \times \left(1 - \sum_{k=1}^{n-1} \theta_{ik}\right)^{\beta_{in}} = \left(1 - \sum_{k=1}^{n-1} \theta_{ik}\right)^{\beta_{in}} \times \prod_{k=1}^{n-1} \theta_{ik}^{\beta_{ik}}$$

- S_i, le système de $n-1$ équations à résoudre pour trouver la ligne i de la matrice de transition par la méthode du maximum de vraisemblance. On note S_{ij} la $j^{\text{ième}}$ équation de ce système.

$$S_i \equiv \begin{cases} \dfrac{\partial P(X|\underline{\theta_i})}{\partial \theta_{i1}} = 0 \\ \vdots \\ \dfrac{\partial P(X|\underline{\theta_i})}{\partial \theta_{i(n-1)}} = 0 \end{cases}$$

La résolution de ce système ne donne que les $n-1$ probabilités de la ligne i. Il suffit de les sommer pour obtenir la $n^{\text{ième}}$.

Développons l'équation trouvée ci-dessus (pour $j \in [2, n-2]$ mais que l'on peut facilement généraliser pour les cas où $j = 1$ ou $(n-1)$) :

$$\begin{aligned} S_{ij} &= \left[\beta_{ij} \theta_{ij}^{\beta_{ij}-1} \left(1 - \sum_{k=1}^{n-1} \theta_{ik}\right)^{\beta_{in}} - \beta_{in} \left(1 - \sum_{k=1}^{n-1} \theta_{ik}\right)^{\beta_{in}-1} \theta_{ij}^{\beta_{ij}} \right] \times \prod_{k=1, k\neq j}^{n-1} \theta_{ik}^{\beta_{ik}} \\ &= \left[\beta_{ij} \left(1 - \sum_{k=1}^{n-1} \theta_{ik}\right) - \beta_{in}\theta_{ij} \right] \times \left(1 - \sum_{k=1}^{n-1} \theta_{ik}\right)^{\beta_{in}-1} \times \theta_{ij}^{\beta_{ij}-1} \times \prod_{k=1, k\neq j}^{n-1} \theta_{ik}^{\beta_{ik}} = 0 \end{aligned}$$

Il nous reste à résoudre l'équation suivante :

$$\begin{aligned} & \beta_{ij} \left(1 - \sum_{k=1}^{n-1} \theta_{ik}\right) - \beta_{in}\theta_{ij} = 0 \\ \Leftrightarrow\ & \beta_{ij} \left(1 - \sum_{k=1}^{n-1} \theta_{ik}\right) = \beta_{in}\theta_{ij} \\ \Leftrightarrow\ & \frac{\beta_{in}}{\beta_{ij}}\theta_{ij} + \sum_{k=1}^{n-1} \theta_{ik} = 1 \\ \Leftrightarrow\ & \left(\frac{\beta_{in}}{\beta_{ij}} + 1\right)\theta_{ij} + \sum_{k=1, k\neq j}^{n-1} \theta_{ik} = 1 \quad (1) \end{aligned}$$

Le système S_i peut se réécrire sous forme matricielle de la manière suivante :

$$S_i \equiv \begin{pmatrix} \left(\frac{\beta_{in} + \beta_{i1}}{\beta_{i1}}\right) & 1 & \cdots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & \left(\frac{\beta_{in} + \beta_{i(n-1)}}{\beta_{i(n-1)}}\right) \end{pmatrix} \begin{pmatrix} \theta_{i1} \\ \vdots \\ \theta_{i(n-1)} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

La résolution de ce système pour chaque ligne permet d'obtenir une estimation de la matrice de transition. Néanmoins, cette méthode est extrêmement inefficace puisqu'elle nécessite n résolutions du système (la résolution d'un système étant elle-même une opération coûteuse en terme de temps de calcul). Bien qu'à notre échelle ($n = 50$), cette complexité élevée ne soit pas gênante, une simplification de la méthode d'estimation serait tout de même bienvenue. Repartons

de l'équation (1) :

$$\begin{aligned}
&\left(\frac{\beta_{in}}{\beta_{ij}} + 1\right)\theta_{ij} + \sum_{k=1,k\neq j}^{n-1} \theta_{ik} = 1 \\
\Leftrightarrow\ &\frac{\beta_{in}}{\beta_{ij}}\theta_{ij} + \theta_{ij} = 1 - \sum_{k=1,k\neq j}^{n-1} \theta_{ik} \\
\Leftrightarrow\ &\frac{\beta_{in}}{\beta_{ij}}\theta_{ij} + \theta_{ij} = \theta_{ij} + \theta_{in} \\
\Leftrightarrow\ &\theta_{ij} = \frac{\theta_{in}}{\beta_{in}}\beta_{ij} \quad (2) \\
\Leftrightarrow\ &\sum_{i=0}^{n-1} \theta_{ij} = \frac{\theta_{in}}{\beta_{in}} \sum_{i=0}^{n-1} \beta_{ij} \\
\Leftrightarrow\ &1 - \theta_{in} = \frac{\theta_{in}}{\beta_{in}} (\sigma_i - \beta_{in}) \\
\Leftrightarrow\ &\theta_{in} = \frac{\beta_{in}}{\sigma_i} \quad (3)
\end{aligned}$$

En injectant l'équation (3) dans l'équation (2), on obtient l'équation :

$$\theta_{ij} = \frac{\beta_{ij}}{\sigma_i}$$

La formule ci-dessus nous permet de simplifier l'estimation puisque qu'il suffit maintenant de **diviser chaque élément β_{ij} de la matrice N par la somme de la ligne dans laquelle il se trouve** pour trouver la matrice de transition. Il ne reste plus qu'à régler le problème où

$$\beta_{ij} = 0, \forall j \in [1, n]$$

En effet, cette situation mènerait à une division par zéro. Il faut donc appliquer un traitement à la matrice N afin de supprimer les valeurs nulles.

Smoothing

Un des problèmes liés à l'estimation par le maximum de vraisemblance se pose lorsque le set de données permettant d'estimer Q ne contient pas d'information sur certains nœuds, *i.e.* $\beta_{ij} = 0$. En effet, lorsque par la suite, nous souhaiterions estimer la probabilité qu'une certaine séquence, contenant un nœud qui n'a pas été compté pour estimer Q, corresponde à ce modèle, nous obtiendrons une probabilité nulle.

Plusieurs techniques de lissage existent : soient la transition du nœud ω_{i-1} vers le nœud ω_i observée, $c(x)$ le nombre de fois où le nœud x a été compté, $c(x, y)$ le nombre de fois la transition du nœud x vers le nœud y a été compté et n le nombre de nœuds :

- **Méthode de Laplace (*Laplace smoothing*) :** Cette méthode consiste à ajouter 1 à tous les éléments comptés. On obtient alors une probabilité de la forme :

$$P_{Laplace}(\omega_i|\omega_{i-1}) = \frac{c(\omega_{i-1}, \omega_i) + 1}{c(\omega_{i-1}) + n}$$

– ***Add-k smoothing*** **:** Cette méthode généralise la méthode de Laplace où, cette fois, k est ajouté au numérateur et $k.n$ est ajouté au dénominateur :

$$P_{Add-k}(\omega_i|\omega_{i-1}) = \frac{c(\omega_{i-1},\omega_i)+k}{c(\omega_{i-1})+k.n}$$

– ***Unigram prior smoothing*** **:** Il s'agit d'une variante de l'*add-k smoothing* où la probabilité d'obtenir l'unigramme ω_i. Par rapport à la formule de l'*add-k smoothing*, on prend $k = m/n$ où m est un paramètre à déterminer. On obtient alors :

$$P_{Add-k-var}(\omega_i|\omega_{i-1}) = \frac{c(\omega_{i-1},\omega_i)+m.\left(\frac{1}{n}\right)}{c(\omega_{i-1})+m}$$

La "subtilité" de l'*unigram prior smoothing* consiste à remplacer la constante $\frac{1}{n}$ au numérateur par $P(\omega_i)$. On a au final :

$$\begin{aligned} P_{UnigramPrior}(\omega_i|\omega_{i-1}) &= \frac{c(\omega_{i-1},\omega_i)+m.P_{ML}(\omega_i)}{c(\omega_{i-1})+m} \\ &= \frac{c(\omega_{i-1},\omega_i)+m\frac{c(\omega_i)}{n}}{c(\omega_{i-1})+m} \end{aligned}$$

– ***Jelinek-Mercer smoothing*** **(interpolation) :** Il s'agit d'une interpolation linéaire entre le modèle de l'unigramme et du bigramme afin de pouvoir prendre en compte les probabilités de ces deux modèles :

$$\begin{aligned} P_{Jelinek-Mercer}(\omega_i|\omega_{i-1}) &= \lambda.P_{ML}(\omega_i|\omega_{i-1}) + (1-\lambda).P_{ML}(\omega_i) \\ &= \lambda\frac{c(\omega_i,\omega_{i-1})}{c(\omega_i)} + (1-\lambda)\frac{c(\omega_i)}{n} \end{aligned}$$

Test de la méthode

Pour tester cette méthode, nous avons procédé en plusieurs étapes :
– Définition d'une matrice d'adjacence A de taille $k \times k$ (i, $k = 5$)
– Calcul de la matrice de transition Q et génération de n (ici, $n = 50$) chaîne de Markov X_i de taille m (ici, $m = 1000$)
– Génération de n estimations $Q_{est,i}$ de la matrice de transition
– Calcul de l'erreur quadratique moyenne E (EQM) pour chaque θ_{ij}

En utilisant cette méthode nous avons obtenu des résultats intéressants. Ci-dessous sont donnés la matrice Q exacte, l'erreur quadratique moyenne sur les 50 générations et Q_{est}, une des matrices estimées :

$$Q = \begin{pmatrix} 0.1000 & 0.3500 & 0.3500 & 0.1000 & 0.1000 \\ 0.1000 & 0.1000 & 0.1000 & 0.6000 & 0.1000 \\ 0.3500 & 0.3500 & 0.1000 & 0.1000 & 0.1000 \\ 0.1000 & 0.1000 & 0.3500 & 0.1000 & 0.3500 \\ 0.3500 & 0.1000 & 0.1000 & 0.3500 & 0.1000 \end{pmatrix}$$

$$E = \begin{pmatrix} 0.0066 & 0.0459 & 0.0585 & 0.1327 & 0.0284 \\ 0.0456 & 0.1188 & 0.0153 & 0.0657 & 0.0335 \\ 0.0046 & 0.0018 & 0.0127 & 0.0203 & 0.0039 \\ 0.0141 & 0.0236 & 0.0510 & 0.0018 & 0.0511 \\ 0.3890 & 0.0006 & 0.0335 & 0.2911 & 0.0082 \end{pmatrix}$$

$$Q_{est} = \begin{pmatrix} 0.0883 & 0.3282 & 0.3659 & 0.0952 & 0.1224 \\ 0.1075 & 0.1061 & 0.1016 & 0.5906 & 0.0942 \\ 0.3312 & 0.3629 & 0.1060 & 0.0906 & 0.1093 \\ 0.0946 & 0.1123 & 0.3322 & 0.1064 & 0.3546 \\ 0.3529 & 0.0997 & 0.1124 & 0.3390 & 0.0960 \end{pmatrix}$$

Nous avons calculé la moyenne des erreurs quadratiques contenues dans E et avons trouvé une erreur moyenne d'environ 5% pour cette configuration. La Table 2.1 contient l'évolution de cette grandeur lorsque m varie. On constate (et c'était attendu) que l'erreur quadratique diminue au fur et à mesure que la taille des chaînes de Markov utilisées augmente.

Cette Table reprend aussi le temps d'exécution pour le calcul des n générations. On constate que ce dernier croît linéairement avec m.

On peut conclure que la taille des traces utilisées pour estimer une matrice de transition doit être raisonnablement longue et que $m = 1000$ (et 5000) semble(nt) être un (de) bon(s) compromis entre temps d'exécution et précision.

m	**Moyenne de l'erreur**	**Temps de calcul** (sec)
100	0.3265	0.0379
1000	**0.0509**	0.3109
5000	0.0096	1.5646
10000	0.0066	3.0695

TABLE 2.1 – Erreur moyenne en fonction du paramètre m

2.1.2 Utilisation des modèles estimés

2.2 Estimation de α

1) Le paramètre α est un paramètre de la matrice de transition représentant un surfeur aléatoire. Nous avons défini dans la Section 1.3, la formule définissant chaque élément de cette matrice de transiton en fonction de α pour le modèle du surfeur aléatoire :

$$Q_t(i,j) = q_{ij}(1-\alpha) + \frac{\alpha}{n}$$

Un première approche est d'utiliser cette formule afin de définir le paramètre recherché.

La seule inconnue du membre de droite est α. En effet, nous connaissons de manière exacte le terme q_{ij} puisque nous connaissons le graphe sur lequel le surfeur se déplace. En ce qui concerne le membre de gauche, nous sommes capable de l'estimer par la méthode développée au point précédent. Il est important de garder en tête que **cette matrice de transition est estimée** et que chaque probabilité de la matrice est entachée d'une erreur qui se propagera, lors de la résolution, dans la valeur de α.

Nous avons à notre disposition n^2 équations qui vont toutes nous donner des valeurs différentes pour α.

Annexe A

Comparaison de différentes méthodes de lissage

A.1 Estimation de Q

On compare ici les précisions atteignables, en utilisant la méthode explicitée au point 2.1.1 - “Test de la méthode”.

On note m_t la taille de la trace :

m_t	Laplace	Unigram prior ($m = 0.06 * 50$)	Interpolation ($\lambda = 0.3$)
100	1.3755×10^{-3}	1.5444×10^{-3}	4.2131×10^{-4}
300	1.4585×10^{-3}	1.4937×10^{-3}	2.8596×10^{-4}
500	1.1939×10^{-3}	1.2146×10^{-3}	2.5819×10^{-4}
800	9.0119×10^{-4}	9.1241×10^{-4}	2.4232×10^{-4}
1000	7.6995×10^{-4}	7.7718×10^{-4}	2.2381×10^{-4}
2000	4.3842×10^{-4}	4.3775×10^{-4}	2.2826×10^{-4}
5000	1.8841×10^{-4}	1.8972×10^{-4}	2.2237×10^{-4}
10000	9.6356×10^{-5}	9.6745×10^{-5}	2.2023×10^{-4}

Table A.1 – Précision de l’estimation de Q

On cherche maintenant à estimer le paramètre m de la méthode “uniform prior” tel que celui-ci maximise la précision. On fixe la taille de la trace à 5000.

On cherche ici à estimer le paramètre λ de la méthode par interpolation tel que celui-ci maximise la précision. On fixe la taille de la trace à 5000.

k	$m = k.50$	Unigram prior
0.01	0.5	1.9824×10^{-4}
0.1	5	1.8813×10^{-4}
0.3	15	1.5445×10^{-4}
0.5	25	1.3759×10^{-4}
0.7	35	1.2522×10^{-4}
0.9	45	1.1752×10^{-4}
1	50	1.1499×10^{-4}
2	100	1.0787×10^{-4}
3	150	1.149×10^{-4}
5	250	$1.332\,35 \times 10^{-4}$

TABLE A.2 – Estimation du paramètre m pour la méthode “uniform prior”

λ	Interpolation
0.1	2.2511×10^{-4}
0.2	2.2374×10^{-4}
0.3	2.2222×10^{-4}
0.4	2.2023×10^{-4}
0.5	2.19×10^{-4}
0.6	2.1351×10^{-4}
0.7	2.0713×10^{-4}
0.8	2.0226×10^{-4}
0.9	1.6776×10^{-4}
1	1.9933×10^{-4}

TABLE A.3 – Estimation du paramètre λ pour la méthode par interpolation