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.

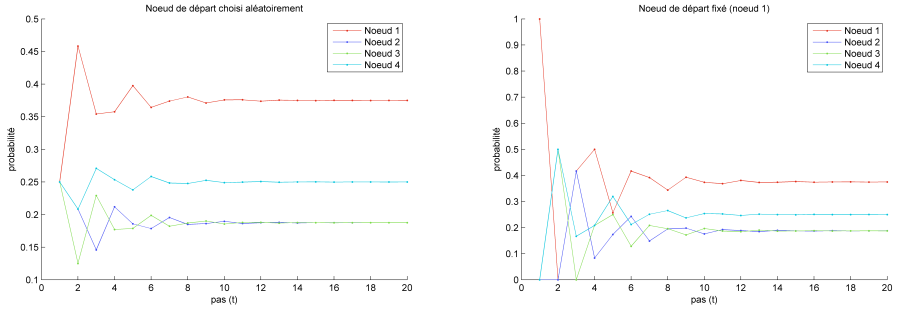


FIGURE 1.1 – Évolution de la distribution de probabilité

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 :

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

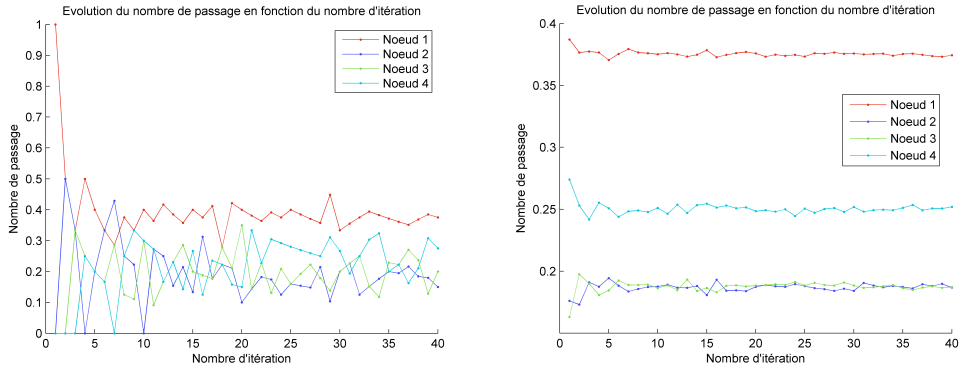
On constate que les lignes de la matrice sont très proches

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 :

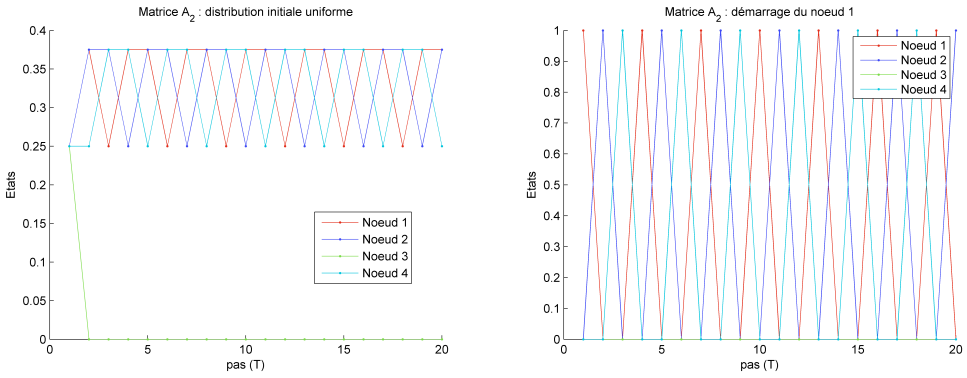
$$Q_2 = \begin{pmatrix} 0 & 1.0000 & 0 & 0 \\ 0 & 0 & 0 & 1.0000 \\ 0.5000 & 0.5000 & 0 & 0 \\ 1.0000 & 0 & 0 & 0 \end{pmatrix} \quad Q_3 = \begin{pmatrix} 0 & 1.0000 & 0 & 0 & 0 \\ 1.0000 & 0 & 0 & 0 & 0 \\ 0 & 0.3333 & 0 & 0.3333 & 0.3333 \\ 0 & 0 & 0 & 0 & 1.0000 \\ 0 & 0 & 0 & 1.0000 & 0 \end{pmatrix}$$

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*.

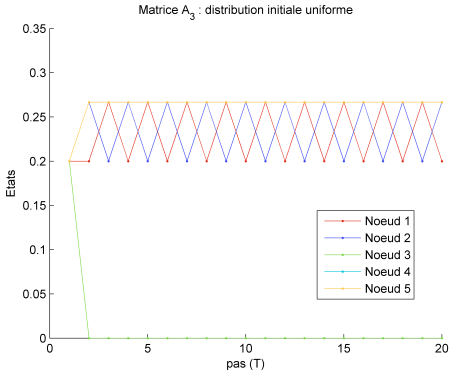
Ces deux phénomènes sont précisément des **limitations** du modèle du surfeur aléatoire simpliste présenté dans cette première partie. D'une part, les oscillations empêchent d'atteindre une distribution stationnaire. D'autre part, la probabilité nulle d'atteindre un nœud existant rend l'éventuelle distribution 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.

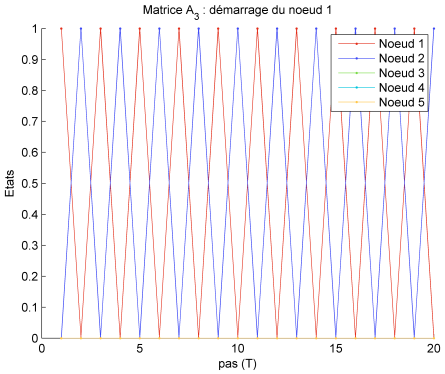


(a) Distribution initiale uniforme

(b) Depuis le noeud 1



(c) Distribution initiale uniforme



(d) Depuis le noeud 1

FIGURE 1.3 – Évolution des distributions de probabilités

2) Tout d'abord, affichons les résultats obtenus à partir de la fonction `findStationnaryPi` (*cfr.* question 1.1.4 - méthode des puissances) appliquée aux différentes matrices de transition :

	Nœud 1	Nœud 2	Nœud 3	Nœud 4	Nœud 5
Q_1	0.3750	0.1875	0.1875	0.2500	—
Q_2	0.3750	0.3750	0	0.2500	—
Q_3	0.2667	0.2000	0	0.2667	0.2667

Notons que les distributions pour Q_2 et Q_3 ont été calculés à partir d'une distribution uniforme.

Dans cette section, deux méthodes de calcul ont été utilisées :

– **Résolution du système linéaire** : On a $\pi_\infty Q = \pi_\infty$ et $\sum_i \pi_{\infty,i} = 1$, que l'on peut transformer en un système :

$$(Q' - \mathbb{1})\pi_\infty^T = b$$

où Q' est formé de la matrice Q^T à laquelle une ligne de '1' a été ajoutée (matrice de taille $(k+1) \times k$, k étant le nombre de nœuds) et b est un vecteur colonne (de taille $k+1$) composé de zéros sauf le dernier élément qui vaut '1'.

– **Résolution par vecteurs/valeurs propres** : On a le système suivant :

$$\pi_\infty Q = \pi_\infty$$
$$\Leftrightarrow Q^T \pi_\infty^T = \pi_\infty^T$$

Cette équation est une équation de type $Ax = \lambda x$, où A est de taille $n \times n$, x est $n \times 1$ et λ est un scalaire. Ce type d'équation est en fait un calcul de recherche de vecteurs et valeurs propres. Dans notre cas, la valeur propre λ vaut 1, il s'agit alors de calculer le vecteur propre correspondant et de le normaliser afin de respecter l'équation $\sum_i \pi_{\infty,i} = 1$, ce qui a été implémenté dans la fonction `getStationnaryPiBySystem`.

Ces deux méthodes donnent un résultat exactement identique pour chaque matrice de transition, à l'exception du troisième graphe pour lequel une distribution supplémentaire a été trouvée. En effet, la fonction `linsolve` émet un warning "*Rank deficient*" nous informant que le système est indéterminé, il n'y a pas suffisamment d'équations, mais ne renvoie qu'une seule solution. On peut en déduire qu'il existe une infinité de solutions.

On peut obtenir ce même résultat à partir des deux distributions obtenues par la méthode des vecteurs/valeurs propres : supposons les distributions stationnaires π_A et π_B, par définition :

$$\pi_A Q_3 = \pi_A$$

et

$$\pi_B Q_3 = \pi_B$$

On peut alors trouver une autre distribution (stationnaire) qui est combinaison des deux précédentes :

$$(\alpha\pi_A + \beta\pi_B)Q_3 = (\alpha\pi_A + \beta\pi_B)$$

On peut en déduire qu'il existe une infinité de distributions stationnaires pour la matrice Q_3.

Contrairement à la matrice A_1, les matrices A_2 et A_3 ne donnent pas le même résultats qu'avec la fonction `findStationnaryPi`. En effet, en exécutant celle-ci, le message "*Number of maximum iterations reached*" apparaît, signifiant que la sortie a une plus forte probabilité d'être un état non-stationnaire, c'est-à-dire que la fonction a été arrêtée avant d'atteindre un état stationnaire, s'il y en a un. Dans notre

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) SoitX_n la variable aléatoire qui prend la valeur p_i à l'état n si le surfeur est sur la page i. On veut calculer :

$$P(X_{10} = p_2 | X_{20} = p_3, X_1 = p_1)$$

En appliquant la formule de Bayes et en se rappelant de la propriété d'une chaîne de Markov, on obtient :

$$\frac{P(X_{10} = p_2 | X_1 = p_1) \cdot P(X_{20} = p_3 | X_{10} = p_2,)}{P(X_{20} = p_3 | X_1 = p_1)}$$

Lorsque l'on passe à la notation matricielle, on a :

$$\frac{Q^{10}(p_3, p_2) \cdot Q^{9}(p_2, p_1)}{Q^{19}(p_3, p_1)}$$

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 $\alpha = 1$

2) Nous avons analysé l'évolution du PageRank de certaines pages lorsque alpha évolue. En regardant la Figure 1.4(a) , on observe que plus alpha est grand, plus le PageRank des pages tend à s'uniformiser. En effet, un PageRank élevé au départ va diminuer avec l'augmentation de α, tandis qu'un PageRank assez bas va augmenter avec α vu que plus alpha est élevé, plus la probabilité d'arriver sur une telle page s'uniformise.

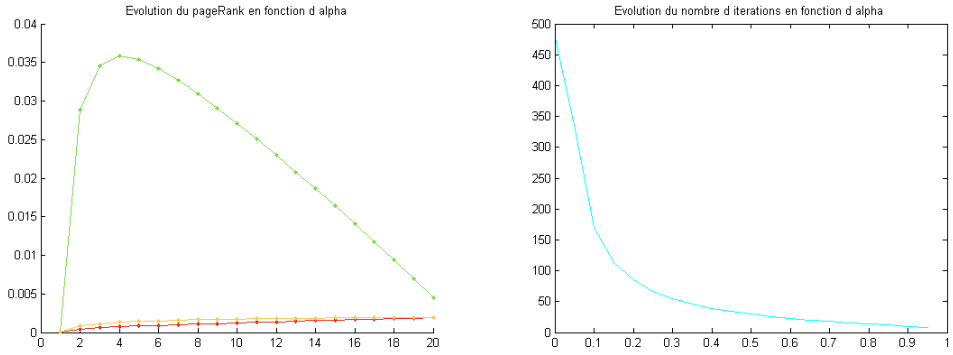


FIGURE 1.4 – Évolution du PageRank et du nombre d'itérations avec α

3) Nous avons déterminé l'évolution du nombre d'itérations nécessaires pour trouver une distribution stationnaire lorsque α varie. Les résultats sont présentés sur la Figure 1.4(b) :

Les résultats observés sont attendus car il est logique de trouver une distribution stationnaire très rapidement lorsque α tend vers 1, étant donné que on ne tient compte que de la téléportation : la probabilité d'arriver sur une page ou une autre partant d'une page donnée est uniforme. Dans le cas contraire, lorsque alpha tend vers zéro, on repose plus sur le graphe de départ ce qui est plus compliqué à calculer qu'une répartition uniforme.

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) :

$$\left(\frac{\beta_{in}}{\beta_{ij}} + 1\right)\theta_{ij} + \sum_{k=1,k\neq j}^{n-1} \theta_{ik} = 1$$

$$\Leftrightarrow \frac{\beta_{in}}{\beta_{ij}}\theta_{ij} + \theta_{ij} = 1 - \sum_{k=1,k\neq j}^{n-1} \theta_{ik}$$

$$\Leftrightarrow \frac{\beta_{in}}{\beta_{ij}}\theta_{ij} + \theta_{ij} = \theta_{ij} + \theta_{in}$$

$$\Leftrightarrow \theta_{ij} = \frac{\theta_{in}}{\beta_{in}}\beta_{ij} \quad (2)$$

$$\Leftrightarrow \sum_{i=0}^{n-1} \theta_{ij} = \frac{\theta_{in}}{\beta_{in}} \sum_{i=0}^{n-1} \beta_{ij}$$

$$\Leftrightarrow 1 - \theta_{in} = \frac{\theta_{in}}{\beta_{in}} (\sigma_i - \beta_{in})$$

$$\Leftrightarrow \theta_{in} = \frac{\beta_{in}}{\sigma_i} \quad (3)$$

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

Plusieurs techniques de lissage existent dont la méthode de Laplace (*Laplace smoothing*) qui consiste à ajouter 1 à tous les éléments comptés.

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 sur base des chaînes générées
- 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 données la matrice Q exacte, l'erreur quadratique moyenne sur les 50 générations et Q_{est}, une des matrices estimées :

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

$$E = \begin{pmatrix} 0.0006 & 0.0008 & 0.0015 & 0.0005 & 0.0003 \\ 0.0005 & 0.0003 & 0.0005 & 0.0008 & 0.0003 \\ 0.0012 & 0.0009 & 0.0004 & 0.0005 & 0.0005 \\ 0.0004 & 0.0002 & 0.0009 & 0.0004 & 0.0012 \\ 0.0017 & 0.0006 & 0.0004 & 0.0016 & 0.0004 \end{pmatrix}$$

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

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.

On peut conclure que la taille des traces relativement courte permettent d'obtenir une erreur quadratique tout à fait acceptable.

Afin de pouvoir mettre le problème à l'échelle, il est plus intéressant d'étudier le paramètre $\frac{m}{k}$. En effet, plus ce paramètre est petit, moins la chaîne de Markov aura tendance à être représentative de la matrice de transition et inversément. Dans notre cas, nous obtenons un rapport $\frac{m}{k} = 20$ qui donne déjà une bonne précision.

m	**Moyenne de l'erreur**	$\frac{m}{k}$
100	0.0075	20
1000	6.9755×10^{-4}	200
5000	1.5653×10^{-4}	1000
10000	7.2594×10^{-5}	2000

TABLE 2.1 – Erreur moyenne en fonction du paramètre m

2.1.2 Utilisation des modèles estimés

Une méthode pour déterminer l'origine des traces d'un ensemble de traces consiste à calculer la probabilité $P(Q_i|tr_k)$ que tr_k provienne d'une certaine matrice de transition Q_i (la matrice de transition définissant le comportement d'un surfeur). Il suffira ensuite de prendre le surfeur dont la matrice de transition mène à la plus grande probabilité. Autrement dit :

$$arg\max_i P(Q_i|tr_k)$$

A l'aide de la formule de Bayes, on peut développer :

$$P(Q_i|tr_k) = \frac{P(tr_k|Q_i)P(Q_i)}{\sum_{j=1}^{n} P(tr_k|Q_j)P(Q_j)}$$

où le terme $P(tr_k|Q_i)$ est la vraisemblance de la trace connaissant la matrice Q_i et $P(Q_i)$ est la probabilité qu'une trace provienne du surfeur i.

Dans notre cas, la vraisemblance est un nombre tellement petit qu'il n'est pas représentable sur un ordinateur (à moins d'utiliser des bibliothèques spéciales). Nous avons donc décidé d'utiliser la **log-vraisemblance** de sorte que les nombres soient manipulables.

Étant donné que les traces tr_k ne peuvent appartenir qu'à deux surfeurs, nous pouvons aussi modifier notre règle de décision. Pour ce faire, nous définissons :

$$\begin{aligned} r &= \log \frac{P(tr_k|Q_1)}{P(tr_k|Q_2)} \\ &= \log \frac{P(tr_k|Q_1)P(Q_1)}{P(tr_k|Q_2)P(Q_2)} \\ &= \log P(tr_k|Q_1) - \log P(tr_k|Q_2) + \log \frac{P(Q_1)}{P(Q_2)} \end{aligned}$$

Si $r > 0$ alors la trace d'origine provient du **surfer 1** sinon elle provient du **surfer 2**. Notons que dans notre cas :

$$P(Q_i) = \frac{1}{2}, \forall i \in [1, 2]$$

En effet, nous savons que parmi les 20 traces à évaluer, 10 appartiennent au premier (soit la moitié) et 10 au deuxième (soit l'autre moitié).

Calcul de la vraisemblance

Pour une trace X, une matrice N contenant les éléments β_{ij} (nombre de transition de l'état i à j dans X) et une matrice de transition Q, la vraisemblance est donnée par :

$$P(X|Q) = \prod_{i=1}^{n} \prod_{j=1}^{n} \theta_{ij}^{\beta_{ij}}$$

Comme expliqué plus haut, nous utiliserons la log-vraisemblance qui se calcule comme suit :

$$logP(X|Q) = \sum_{i=1}^{n} \sum_{i=1}^{n} \beta_{ij} \log \theta_{ij}$$

Test de la méthode

Analyse des 20 traces

En appliquant la méthode expliquée dans la Section 2.1.2, nous avons obtenu les résultats donnés dans la Table 2.2.

Trace	$\log P(Q_1\|tr_k)$	$\log P(Q_2\|tr_k)$	r	Provient de
1	−3890.9	−3890.5	−0.4360	**2**
2	−3865.5	−3923.2	57737	**1**
3	−4247.5	−4318.1	70562	**1**
4	−3849.6	−3917.7	68022	**1**
5	−4253.0	−4099.6	−153.35	**2**
6	−3763.3	−3890.7	127.45	**1**
7	−4276.2	−4228.4	−47847	**2**
8	−3866.2	−3914.1	47912	**1**
9	−4144.5	−4144.5	−13472	**2**
10	−4161.5	−4161.5	−10209	**2**
11	−4304.2	−4304.2	−183.48	**2**
12	−4290.9	−4290.9	−155.67	**2**
13	−3887.7	−3887.7	52513	**1**
14	−3874.6	−3874.6	10682	**1**
15	−3930.0	−3930.0	19723	**1**
16	−4270.7	−4270.7	−104.7	**2**
17	−3932.0	−3897.0	−34937	**2**
18	−4099.3	−4111.8	12536	**1**
19	−4212.2	−4158.6	−53645	**2**
20	−3863.4	−3990.3	126.89	**1**

TABLE 2.2 – Recherche des surfeur ayant engendrés les traces

2.1.3 5)

Cette démarche a de nombreuses applications possibles dans des domaines très variés.

Par exemple, nous avons implémenté ce système pour définir l'appartenance d'un texte à un ensemble : on peut définir si un texte a été écrit par une personne ou l'autre, si ce texte est en français, anglais, si un code est en langage C ou Java, et bien d'autres sur base d'un échantillon de texte.

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.