MA 411 : Modélisation et analyse des processus stochastiques

Chaînes de Markov à temps discret (CMTD)
Séance de TD du 05 mai 2020

Vous trouverez ci-après l'énoncé et le corrigé de l'exercice 18 de la séance de TD consacrée aux chaînes de Markov à temps discret. La correction a été rédigée dans le but de vous aider si vous êtes bloqué ou pour vérifier votre propre travail. Il se peut qu'elle contienne elle-même des erreurs. Si tel est le cas, elles seront corrigées au fur et à mesure qu'elles sont détectées. La version en ligne sur *https://chamilo.grenoble-inp.fr/courses/MA332* sera mise à jour de manière à intégrer ces corrections. Dans de nombreux exercices, il existe plusieurs méthodes pour aboutir au résultat. Si vous avez des doutes sur la méthode que vous avez vous-même employée, n'hésitez pas à m'en faire part (*laurent.lefevre@lcis.grenoble-inp.fr*).

Exercice 18 (Pagerank)
On examine dans cet exercice le fonctionnement de l'algorithme *Pagerank* de Google. On note r_i le nombre de liens existant sur la page i.
On modélise l'attitude d'un surfer de la façon suivante : une fois visitée la page i, soit il choisit une adresse Web au hasard (avec la probabilité p), soit il décide de visiter une des pages référencées sur la page i (avec la probabilité $1-p$). Dans le cas où il choisit une adresse au hasard, on considère que chaque page à la même probabilité $\frac{1}{N}$ d'être choisie. Dans le cas où le surfer choisit un des liens de la page i, il le fait également avec une distribution uniforme. Dans ce cas, chaque page a donc une probabilité $\frac{1}{r_i}$.
On considère dans cet exercice un Web miniature[1] constitué de 5 pages et représenté à la figure 1.

1. Modéliser la promenade d'un surfer sur le Web de la figure 1 par une chaîne de Markov dont on donnera le graphe et la matrice de transition.

[1] Dans le cas du Web mondial, les principes de modélisation et de définition de la mesure *pagerank* restent les mêmes. Les méthodes de modélisation, d'identification et de calcul, par contre, sont adaptées à des CMTD de très grande taille. On pourra pour une introduction rapide au sujet se reporter par exemple à la page Wikipedia en anglais sur le Pagerank.

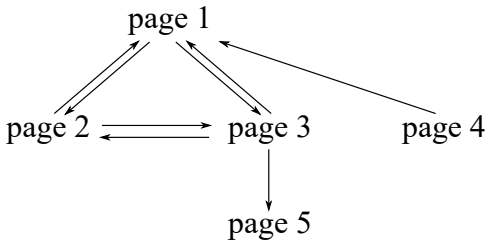

Figure 1: Le Web simplifié considére à l'exercice 18

2. Pourquoi existe-t-il nécessairement une loi limite $\boldsymbol{\pi}^{(\infty)}$? Comment peut-on la calculer?

3. Soit $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_N)$ la distribution stationnaire de probabilité. La mesure *Pagerank* de la page i est alors simplement définie par

$$\text{Pagerank}(i) := \boldsymbol{\pi}_i$$

Calculer le *Pagerank* pour toutes les pages du Web de la figure 1, successivement pour $p = 0.15$, $p = 10^{-2}$ et $p = 0.99$.

4. Imaginer une méthode de calcul itérative de la mesure *Pagerank* qui soit adaptée à un Web de très grande taille mais de structure creuse (la matrice d'incidence du graphe associé est creuse ou, de manière équivalente, une page ne pointe que sur quelques autres pages)

Correction de l'Exercice 18

- Soit $X_n \in E := \{1, 2, 3, 4, 5\}$, la $n^{\text{ième}}$ page où se trouver le surfer au cours de sa visite sur le Web de la figure 1. Comme il choisra la page suivante avec une probabilité que ne dépend que de la page où il se trouve et de la page suivante, on peut modéliser cette visite par une chaîne de Markov dont le graphe est représenté à la figure 2. La matrice de transition associée à cette CMTD s'écrit :

$$\mathbf{P} = (1-p) \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} + \frac{p}{5} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

- La CMTD de la figure 2 est nécessairement apériodique et irréductible, grâce à l'ajout de la possibilité de transition vers n'importe quelle

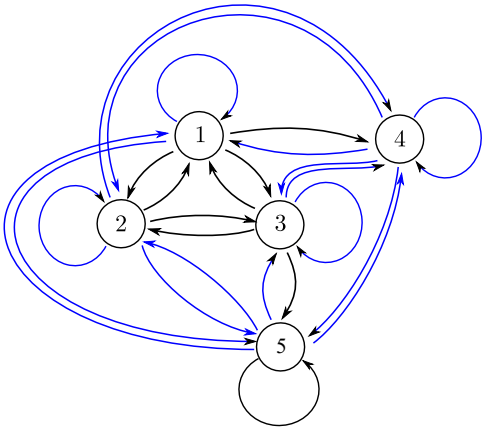

Figure 2: La chaîne de Markov associée au Web de la figure 1. Les arcs en bleu ont été ajoutés pour représenter les transitions possibles en dehors des liens hypertexte présents dans le Web initial. L'ensemble des probabilités de transition (y compris celles qui correspondent aux arcs en noir) sont modifiées par la possibilité d'aller vers n'importe quel lien du Web, au hasard, en dehors des liens hypertexte.

page du Web. Dès lors, comme il s'agit d'une chaîne finie, tous les états sont récurrents non nuls. Dans ce cas, la distribution limite de probabilités $\boldsymbol{\pi}^{(\infty)}$ existe toujours et est nécessairement égale à la distribution stationnaire $\boldsymbol{\pi}$ solution de

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$$

avec la condition de normalisation

$$\sum_{i=1}^{N} \boldsymbol{\pi}_i = 1$$

Il est intéressant de noter que sans la possibilité de transition aléatoire vers une page quelconque du Web, l'état 5 serait absorbant. Dans ce cas la distribution limite serait

$$\boldsymbol{\pi}_i^{(\infty)} = \begin{cases} 0 & \text{si } i \neq 5 \\ 1 & \text{si } i = 5 \end{cases}$$

Il suffirait alors de créer une page sans lien vers d'autres pages pour augmenter arbitrairement son *Pagerank*. Avec une probabilité p de s'échapper d'un tel piège, on retrouvera pour chaque état une distribution limite proche du nombre de liens qui pointent vers la page correspondante.

- En résolvant le système ci-dessus, pour $p = 0.15$[2], on obtient :

$$\boldsymbol{\pi} = \begin{pmatrix} 0.1623.. & 0.1444.. & 0.1604.. & 0.0300.. & 0.5029.. \end{pmatrix}$$

Cette valeur de p ne compense qu'imparfaitement la sur-représentation de la page 5. Cet effet est du à la taille très faible du Web considéré. Une augmentation de la valeur de p augmentera la probabilité de sauter vers les autres pages sans suivre les liens hypertexte. A la limite, on se rapprochera d'une distribution uniforme. Par exemple, pour pour $p = 0.99$, on obtient :

$$\boldsymbol{\pi} = \begin{pmatrix} 0.2016.... & 0.1997.. & 0.2000.. & 0.1980.. & 0.2007.. \end{pmatrix}$$

A contrario, pour p proche de zéro, nous ne compensons plus l'effet "cul de sac" de la page 5. Par exemple, pour $p = 0.01$, on obtient :

$$\boldsymbol{\pi} = \begin{pmatrix} 0.0211.... & 0.0198.. & 0.0223.. & 0.0020.. & 0.9348.. \end{pmatrix}$$

Si nous modifions le Web de la figure 1, en ajoutant un lien de la page 5 à la page 4, on obtient par contre (toujours avec la valeur $p = 0.01$) :

$$\boldsymbol{\pi} = \begin{pmatrix} 0.3028.... & 0.2413.. & 0.2711.. & 0.0931.. & 0.0917.. \end{pmatrix}$$

qui correspond à l'intuition que l'on peut avoir de l'importance relative des pages de ce "nouveau" Web (en gros, le nombre de liens qui pointent vers chaque page)

- L'application de la méthode précédente au calcul du *Pagerank* pour un Web qui comprend plusieurs milliards de pages est bien sûr exclue. On pourra utiliser pour de très grands réseaux la récurrence

$$\boldsymbol{\pi}^{(n+1)} = \mathbf{P}\boldsymbol{\pi}^{(n+1)}$$

qui concerge vers la distribution stationnaire de probabilités[3]. En effet, d'une part la matrice d'incidence des très grands réseaux est en général creuse et, d'autre part, les renouvellements de ces réseaux sont généralement progressifs et ne concernent que peu de noeuds (et d'arrêtes). Or la récurrence ci-dessus se calcule aisément pour des matrices $\mathbf{P}$ creuses (on n'effectue pas les multiplications avec les éléments nuls dans les rangées de $\mathbf{P}$). De plus, si on démarre cette récurrence avec une distribution initiale proche de la distribution stationnaire avant les derniers changements (e.g. l'ajout de quelques pages), il faudra peu d'itérations pour converger vers le nouveau point fixe qui est proche de l'ancien.

[2] C'est la valeur généralement utilisée en pratique, qui correspond à un changement aléatoire toutes les six pages environs

[3] C'est le point fixe de cette récurrence dont on peut montrer qu'elle est toujours contractante