ECE 2 - Année 2017-2018
Lycée français de Vienne
Mathématiques - F. Gaunard
http://frederic.gaunard.com

T.P. n°5

(Introduction aux) Chaînes de Markov

1 Introduction

Les chaînes de Markov sont un exemple de suites (X_n) de variables aléatoires **non indépendantes**. Plus précisément, elles sont un peu l'analogue aléatoire des suites récurrentes d'ordre 1. Le futur (instant $n+1$) ne dépend que du présent (instant n), indépendamment du passé (instant $0, 1, ..., n-1$).

Elles ont des propriétés très intéressantes, notamment de convergence, dans certains cas, vers un *état d'équilibre* (ou *état stable*), et ceci ne dépend pas des conditions initiales.

Les chaînes de Markov sont un objet essentiel des probabilités; elles sont utilisées en théorie des jeux, en physique, en informatique, en finance, sciences sociales, etc...

On retrouvera également dans ce chapitre des notions d'algèbre linéaire.

2 Chaînes de Markov

2.1 Une définition par l'exemple

On modélise l'évolution d'une maladie au sein d'une population au cours du temps en classant les individus en trois groupes:

- Le groupe S des individus sains et non immunisés;
- Le groupe M des individus malades;
- Le groupe I des individus immunisés.

On appelle $E = \{S, M, I\}$ l'ensemble des **états** possibles. On donne la **loi d'évolution** suivante:

- À l'instant 0, tous les individus sont dans l'état S.
- La moitié des individus dans l'état S à l'instant n restent dans le même état à l'instant $n+1$ et l'autre moitié tombe malade; un dixième des individus immunisés à l'instant n passent dans le groupe S à l'instant $n+1$, les autres restent immunisés. Un malade sur cinq le reste, les autres guérissent et deviennent immunisés.

(1) Représenter le modèle épidémiologique précédemment décrit par un graphe (appelé **diagramme de transition**);
(2) On s'intéresse à la suite (X_n) de variables aléatoires telle que X_n désigne l'état du système l'instant n ($X_n(\Omega) = \{S, M, I\}$).

(a) Recopier et compléter le programme ci-dessous pour qu'il simule la suite des états X_0, X_1, ..., X_n.

```
function X=Exemple(n)
X=zeros(1, n+1);
X(1)=1;
for k=1:n
    if X(k)==1 then
        if rand() < 1/2 then
            X(k+1)=................
        else
            X(k+1)=................
        end
    else
        if X(k)==2 then
            if ........................ then
                X(k+1)=................
            else
                X(k+1)=................
            end
        else
            if ........................ then
                X(k+1)=.................
            else
                X(k+1)=.................
           end
       end
   end
end
endfunction
```

(b) Recopier, exécuter et interpréter les commandes suivantes

```
x=0:1000;
y=Exemple(1000);
plot2d(x,y, -1)
```

(c) Même question avec les commandes ci-dessous

```
U=zeros(1,100);
for k=1:100
    U(k)=Exemple(1000)(1000);
end
frequ=zeros(1,3);
for j=1:3
    for k=1:length(U)
        if U(k)==j then
            frequ(j)=frequ(j)+1;
        end
     end
end
disp(frequ/100)
```

(3) Pour $n \in \mathbb{N}$, on note S_n (resp. M_n, I_n) l'évènement "l'individu est sain non immunisé à l'instant n" (resp. "malade", "immunisé") et s_n, m_n, i_n les probabilités correspondantes. Enfin, on note

$$U_n = \begin{pmatrix} s_n \\ m_n \\ i_n \end{pmatrix}.$$

(a) À l'aide de la formule des probabilités totales, déterminer une matrice A, appelée **matrice de transition**, telle que

$$U_{n+1} = AU_n.$$

(b) Montrer que, pour tout $n \in \mathbb{N}$, on a $U_n = A^n U_0$. Préciser U_0, appelée **loi initiale**.

(c) Définir, dans la console, la matrice de transition `A` et le vecteur de la loi initiale `u0`, et calculer $A^{1000}U_0$. Comparer la cohérence des résultats avec ceux de la Question 2c.

(4) L'instruction

```
grand(n, 'markov', A', u0)
```

donne un vecteur ligne représentant la simulation des variables $(X_1, X_2, ..., X_n)$ d'une chaîne de Markov (appelée **trajectoire**) de matrice de transition `A` et de loi initiale `u0`.

⚠ On doit rentrer en argument la transposée **`A'`** de `A` car la définition d'une matrice de transition sous `SciLab` correspond à la transposée de la notre.

⚠ La trajectoire ne contient pas l'état initial.

Utiliser la commande précédente pour simuler à nouveau et représenter graphiquement la trajectoire de la chaîne de Markov (on fera attention à rajouter l'état initial dans la trajectoire).

2.2 Blabla théorique

- On considère une suite de variables aléatoires (X_n). Lorsque la loi de X_{n+1} (le futur) ne dépend que de l'état X_n (présent), on dit que (X_n) est une **chaîne de Markov**.

- Notant $X_n(\Omega) = \{\omega_1, ..., \omega_p\}$, on appelle **matrice de transition** la matrice formée des probabilités conditionnelles

$$A = \left(P_{[X_n=\omega_j]}([X_{n+1} = \omega_i])\right)_{1 \leq i,j \leq p}.$$

☞ La matrice de transition ne dépend pas de n: la loi d'évolution d'un instant au suivant reste la même au cours du temps.

☞ Les coefficients d'une matrice de transition sont positifs et la somme des coefficients de chaque colonne vaut 1. (On dit que c'est une *matrice stochastique*.)

- La matrice colonne $U_n = (P([X_n = \omega_i])_{1 \leq i \leq p}$ donne la loi de X_n. La formule des probabilités totales appliquée au s.c.e constitué de l'ensemble des états permet de voir que

$$U_{n+1} = AU_n$$

☞ Une récurrence classique, que l'on sait tous faire *fingers in the nose*, permet de voir que

$$U_n = A^n U_0.$$

- **Supposons** (et c'est une supposition qui en est vraiment une) que pour un moment donné n_0, l'état de la chaîne vérifie

$$AU_{n_0} = U_{n_0}.$$

(C'est à dire que U_{n_0} est **vecteur propre** associé à la valeur propre 1 pour A.) Une récurrence immédiate donnerait, pour tout $n \geq n_0$ (tout instant à partir de cet instant n_0)

$$U_n = U_{n_0}.$$

Les suites $(P([X_n = \omega_i]))_n$ sont donc constantes à partir d'un certain moment et la distribution correspondante de probabilités (le vecteur U_{n_0}) s'appelle **état stable**.

☞ On admet qu'un état stable, s'il existe, est unique.

- Réciproquement, si chacune des suites $(P([X_n = \omega_i]))_n$ converge vers une limite ℓ_i, alors le vecteurs $\Pi = (\ell_i)_{1\leq i\leq p}$ vérifie, par passage à la limite dans la relation $U_{n+1} = AU_n$,

$$\Pi = A\Pi.$$

☞ La loi limite Π est donc forcément l'état stable (*i.e.* un vecteur propre stochastique associé à la valeur propre 1 pour A).

☞ Toujours dans ce cas, peu importe la position initiale, la loi se stabilise vers la même loi limite.

⚠ Même si l'état stable existe, la chaîne de Markov **peut ne pas converger**.

Exemple. On revient à l'exemple de la section précédente.

(1) Taper (et rappeler - ou expliquer - comment fonctionne) l'instruction

```
[P, D]=spec(A)
```

⚠ Attention parfois le vecteur propre associé à 1 n'est pas stochastique. Si tous ces coefficients sont positifs, on se ramène à un vecteur stochastique en divisant chaque coefficient par la somme de ceux-ci.

(2) Exécuter et interpréter

```
i0=floor(find(D==1)/length(D))
statio=P(:,i0)/sum(P(:, i0))
```

3 Exemple: PageRank de Google

Nous allons illustrer la méthode utilisée par Google pour classer les pages web par "indice de popularité". L'idée est d'étudier l'évolution des déplacements d'un individu sur le net, et de regarder la probabilité pi d'être sur un site i au bout d'une très grand nombre de déplacements (limite de la chaîne de Markov). pi est alors appelée "indice de popularité de la page i". Nous commencerons par regarder un cas très simple (et très édulcoré !) où il n'existerait sur le Web que 3 pages internet (!) puis on étudiera le cas général.

3.1 Un exemple avec 3 sites

À l'issue de sa requête, un internaute est susceptible d'aller sur trois sites A, B, et C. On suppose que l'internaute se déplace forcément vers un lien de la page Web où il se trouve.

- le site A contient un lien vers lui-même, un lien vers B, et un lien vers C ;
- le site B contient 5 liens vers lui-même, un vers A et un vers C ;
- le site C comporte un lien vers A, 7 liens vers B et 4 vers C.

À l'instant 0, l'internaute choisit au hasard l'un des trois sites. Par la suite, la probabilité de passer d'un site (à l'instant n) vers un autre (à l'instant n + 1) est proportionnelle au nombre de liens du premier site vers le deuxième.

Pour $n \in \mathbb{N}$, on désigne par a_n, b_n et c_n les probabilités des évènements A_n (resp. B_n, C_n) définis respectivement par le fait de se trouver sur le site A (resp. B, C) à l'instant n.

Enfin, on note X_n la variable aléatoire définie par

$$X_n(\omega) = \begin{cases} 1, & \text{si } \omega \in A_n \\ 2, & \text{si } \omega \in B_n \\ 3, & \text{si } \omega \in C_n \end{cases}$$

(1) En justifiant rigoureusement au moins l'une des égalités, exprimer a_{n+1}, (resp. b_{n+1}, c_{n+1}) en fonction des trois réels a_n, b_n et c_n.
(2) Écrire la matrice de transition S.
(3) Mémoriser la matrice S dans `SciLab` puis compléter et exécuter le programme suivant qui permet d'afficher la trajectoire $(X_0, X_1, ..., X_n)$ de la chaîne de Markov (X_n).

```
function X=navigation(n)
x0=grand(1,1, ........ , .......... , ..........)
X=grand(n,'markov',S',x0);
X= [x0 X];
endfunction
plot2d([0:100], navigation(100), -1)
```

(4) Est-ce qu'un site à l'air plus visité que les autres? Moins visité que les autres? Comment l'expliquer?
(5) Calculer avec `SciLab` les vecteurs de lois $U_0, U_1, ..., U_{10}$. (On recopiera et complètera le programme suivant pour définir une matrice E de taille 3×11 dont la $i-$ème colonne contiendra le vecteur U_{i-1}.)

```
E=zeros(3, 11);
E(:, 1)=(1/3)*ones(3,1);
for k=2:11
    E(:, k)=......................
end
```

(6) Tracer sur un même graphique les courbes des suites $(a_n), (b_n)$ et (c_n). Que remarque-t-on ?
(7) On appelle "indice de notoriété" d'une page A la valeur $a = \lim a_n$. Donner une valeur approchée
(8) Regarder la valeur exacte des indices de notoriété en calculant la valeur exacte de l'état stable (on utilisera la commande `spec()` comme indiqué précédemment.

3.2 Le cas général

On modélise l'ensemble des pages du Web par un ensemble fini d'états $E = \{1, 2, ..., N\}$ (le nombre N est bien très grand, de l'ordre de 10^{13}, mais reste fini).

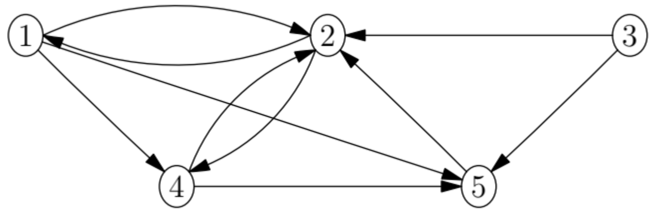
On modélise les déplacements de l'internaute par une chaînes de Markov (X_n) où la variable X_n est le numéro de la page où se situe l'internaute après n déplacements. Après avoir fait des études statistiques des comportements sur le Web, Google part du principe que:

- À l'instant 0, l'internaute choisit une page internet au hasard parmi les N pages;
- Lorsqu'un internaute est sur une page j:
  - dans 85% des cas, il se déplacera au hasard vers l'une des ℓ_j pages pointées par j (ici, une page a au plus un lien vers une autre).
  - dans 15% des cas, il se déplacera au hasard vers n'importe laquelle des pages du Web.

La matrice de transition $M = (P(i,j))$ est alors définie par

$$P(i,j) = \begin{cases} 0,85 \times \frac{1}{\ell_j} + 0,15 \times \frac{1}{N}, & \text{si } j \text{ pointe vers } i \\ 0,15 \times \frac{1}{N}, & \text{sinon} \end{cases}$$

Nous allons illustrer l'algorithme de PageRank sur un groupe autonome de $N = 5$ pages différentes, ayant des liens pointant les pages les unes vers les autres, selon le diagramme suivant:



(1) Taper et expliquer la commande

```
A=0.15*(1/5)*ones(5,5)+0.85*[0, 1/2, 0, 0, 0; 1/3, 0, 1/2, 1/2, 0; 0, 0, 0, 0, 0;
 1/3, 1/2, 0, 0, 0; 1/3, 0, 1/2, 1/2, 1]
```

(2) Déterminer la valeur exacte de la loi stationnaire (à l'aide de la commande `spec()`).

(3) Recopier et exécuter les commandes suivantes. La chaîne converge-t-elle vers l'état stationnaire?

```
E=zeros(5, 11);
E(:, 1)=(1/5)*ones(5,1);
for k=2:11
    E(:, k)=A*E(:, k-1)
end

plot2d([0:10], [E(1,:)',E(2,:)',E(3,:)',E(4,:)',E(5,:)'])
```

(4) Classer les pages par indice de popularité. Les résultats vous semblent-ils cohérents? Commenter.

4 Une autre chaîne de Markov

On veut modéliser la mobilité sociale d'une population. Pour cela, on s'intéresse à la profession des individus d'une génération, comparée à celle de leurs parents. L'INSEE fournit par exemple le tableau suivant, à partir d'un sondage effectué auprès de 18373 individus de sexe masculin

	Agriculteurs	Indépendants	Cadres	Intermédiaires	Employés	Ouvriers
Agriculteurs	621	24	10	9	7	43
Indépendants	237	550	127	138	210	621
Cadres	220	479	749	552	490	700
Intermédiaires	405	439	325	568	643	1520
Employés	214	183	100	182	393	855
Ouvriers	1126	499	116	352	770	3896
Total	2823	2174	1427	1801	2513	7635

L'effectif placé ligne i et colonne j représente le nombre de sondés qui exercent la profession i sachant que leur père exerçait la profession j. On modélise alors la situation par une chaîne de Markov (X_n) à 6 états.

(1) Taper les instructions suivantes et expliquer ce que représente `A`.

```
B=[621, 24, 10, 9, 7, 43; 237, 550, 127, 138, 210, 621; 220, 479, 749, 552, 490,
700; 405, 439, 325, 568, 643, 1520; 214, 183, 100, 182, 393, 855; 1126, 499, 116,
 352, 770, 3896]
A=zeros(6,6);
for i=1:6
    A(:, i) = B(:, i)/sum(B(:, i));
end
```

(2) Écrire une fonction `y=socio(x0,n,r)` qui renvoie un échantillon de taille `r` de la variable $X_{\mathtt{n}}$ en partant de l'état `x0`. Exécuter cette fonction en choisissant les paramètres pour estimer la probabilité (à partir d'un échantillon de taille 100) que l'arrière petit-fils d'un ouvrier soit cadre.

(3) Modifier la fonction précédente pour construire le diagramme à bâtons de la loi (empirique) de X_n (partant de `x0`).

(4) Taper maintenant

```
for i=1:6
    scf(i)
    socio(i, 100, 1000)
end
```

(a) Que représentent les différents diagrammes obtenus?

(b) Les résultats sont-ils cohérents avec la théorie des chaînes de Markov? Le vérifier en calculant la valeur exacte de l'état stable. Quelles conclusions pourrait tirer un sociologue de ces résultats? Quelles critiques pouvez-vous faire?