TP°2 INFO : Chaine de Markov.

1. Introduction **Etude de la fréquentation de sites internet :**

Dans le cadre de la préparation de leurs élèves de ECE, des professeurs ont mis à la disposition de leurs étudiants un site internet comportant leur cours, TD et autres informations. Ils aimeraient savoir quel est le site le plus fréquenté

Situation avec deux sites : Etude avec deux professeurs et des résultats de fréquentation connus.

Dans cette première partie, on se place dans le cadre de deux sites que l'on notera 1 pour celui du prof de maths et 2 pour celui du prof d'ESH. On admet dans cette partie uniquement que l'on dispose déjà de certains résultats de fréquentation obtenu par exemple par le webmestre du lycée. On admet que le site de chaque professeur contient un lien vers celui de son collègue.

L'étude a montré que la semaine 0, l'étudiant internaute choisit de commencer par le site de Maths.
Puis si la semaine n, il était sur le site 1 alors il cliquera sur le lien vers le site 2 avec une probabilité de 0.4 et sur un lien le laissant sur le site 1 avec la probabilité 0.6.
Si la semaine n, il était sur le site 2 alors il cliquera vers le site 1 avec une probabilité égale à 0.2 et vers le même site avec une probabilité 0.8.
Pour n non nul, on note S_n la variable aléatoire égale au numéro du site où se trouve l'internaute la semaine numéro n.
On constate dans cette situation que S_{n+1} ne dépend que de S_n , on dit alors que $(S_n)_{n\in\mathbb{N}^*}$ est une chaine de Markov à deux états (le nombre de sites)

Etude théorique de la situation 1.

On complétera le fichier info « Markov2sites » en fonction de l'avancement de l'étude mathématique.

1. Représenter la situation par un graphe probabiliste.

On note $U_n = \begin{pmatrix} P(S_n=1) \\ P(S_n=2) \end{pmatrix}$

2. Donner U_0 . (Le vecteur U_0 s'appelle l'état initial) ligne 20

3. A l'aide de la formule des probabilités totales, déterminer la matrice A telle que pour tout entier naturel n non nul, on a $U_{n+1} = AU_n$. ligne 7

Définition : La matrice obtenue s'appelle la matrice de transition de la chaine de Markov $(S_n)_{n\in\mathbb{N}^*}$

On a donc $a_{i,j} = P_{(S_n=i)}(S_{n+1}=j)$ pour $(i,j)\in[\![1,N]\!]\times[\![1,N]\!]$ où N désigne le nombre de sites.(ou nombre d'états)

4. Exprimer U_n en fonction de A et U_0.

5. Déterminer le vecteur $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ tel que $AX = X$ et $x_1 + x_2 = 1$. (Vérification dans la console)

Définition : On dit que *X* est la loi invariante de la chaine de Markov $(S_n)_{n\in\mathbb{N}^*}$.

6. On souhaite justifier qu'il existe une existe une matrice P et une matrice D telles que $A = PDP^{-1}$. Ligne 17

   a) Justifier que $sp(A) = \{1; 0.4\}$. Que pouvez-vous en déduire concernant A ?
   b) Justifier qu'il existe une existe une matrice P et une matrice D telles que $A = PDP^{-1}$ et donner *D.*
   c) Donner la dimension des espaces propres et en déduire $E_1(A)$.
   d) Vérifier que $E_{0.4}(A) = Vect\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right)$
   e) Proposer un couple (P,D) tel que $A = PDP^{-1}$.
   f) Exprimer A^n à l'aide de P, D, P^{-1} *et* n.
   g) En déduire $\lim\limits_{n \to +\infty} U_n$ sous forme de produit matriciel. On notera U_∞ le vecteur obtenu.
   h) Déterminer à l'aide de scilab des matrices P et D avec la commande $[P, D] = spec(A)$ et justifier que ces matrices sont en accord avec les résultats théoriques obtenus en 6.c et 6.d

Cette commande sera à retenir dans la perspective des concours. Elle donne pour une matrice A diagonalisable, deux matrices P et D telles que $A = PDP^{-1}$.
On obtient donc grâce à cette dernière le spectre de *A* et des vecteurs propres associés à chacune des valeurs propres présentes sur *D*. On fera attention à bien respecter l'ordre entre valeur propre et vecteur propre.

   i) En déduire une approximation des coordonnées de U_∞ dans la base canonique de $M_{2,1}(\mathbb{R})$ (Console)

Simulation informatique.

La commande R=grand(n, 'markov',A', E_0) permet de simuler la trajectoire d'un internaute en fonction du nombre n de semaines. Le résultat étant stocké dans le vecteur R.

- n désigne le nombre d'états successifs que l'on souhaite observés (différentes pages en fonction des semaines).
- A' désigne la transposée de la matrice de transition de la chaine de Markov.
- E_0 désigne l'état initial de notre internaute

Si on prend pour E_0 un vecteur de taille p, alors la commande simulera p trajectoires sur n semaines avec les états initiaux contenus dans E_0.

7. Ecrire la commande permettant de simuler la trajectoire de l'internaute sur n semaines avec l'état initial proposé dans la situation de départ. Ligne 29
   Faire un test dans la console avec n=10 et expliquer les résultats obtenus.
8. Observez le graphique permettant de visualiser la trajectoire (figure 1) ligne 34
9. Définir une matrice $M \in M_{2,n}(\mathbb{R})$ telle que la colonne i de M contienne le vecteur U_i . lignes 37 & 38
   Commentez les colonnes 1 et 3 de M.

Modèle théorique :

10. Observez sur la figure 2, l'évolution de la probabilité de $P(S_n = i)$ pour $i \in [\![1; n]\!]$
    On a restreint le graphique à un rectangle fixé à l'aide de l'option rect=[0 0 1 n] dans plot2d
    Commentez la commande de la ligne 46
11. Comment identifier graphiquement les coordonnées du vecteur U_∞ . Que constatez-vous ?
12. Quelle est la page la plus fréquentée ?
13. Le constat fait en Q.12. dépend-il de l'état initial ?

Modèle empirique

Dans ce modèle, les trajectoires sont simulées par l'ordinateur à l'aide de la commande R=grand(n, 'markov',A', E_0) où E_0 représente cette fois un vecteur donnant les positions initiales **des p simulations effectuées**.
Pour chaque simulation, une trajectoire est ainsi proposée.
On se propose ici d'observer l'évolution des fréquences d'apparition des numéros des sites en fonctions du nombre de semaine. On parle ici de **fréquence empirique**.

14. Que représente le vecteur site1 en ligne 52
15. Expliquer la commande de la ligne 54
16. Expliquer les commandes des lignes 59 & 60
17. Commentez la figure 3. Q en terme de convergence pour la fréquence empirique.

Situation avec 2 sites : Etude avec deux professeurs et aucun résultat de fréquentation.

Dans cette seconde situation, on considère que nos professeurs ne disposent d'aucune donnée. Ils décident (au moins l'un deux) d'utiliser l'algorithme de Page Rank crée en 1998 (peu avant votre naissance….) par les fondateurs de Google, Brin et Page. Cet algorithme est basé sur l'idée que l'internaute saute de manière aléatoire d'un lien à un autre en fonction de l'architecture de chaque site (liens ou non vers d'autres sites)
On considère ici que chaque site possède un lien vers lui-même et vers l'autre et on suppose donc ici que le clic se fait de manière aléatoire.

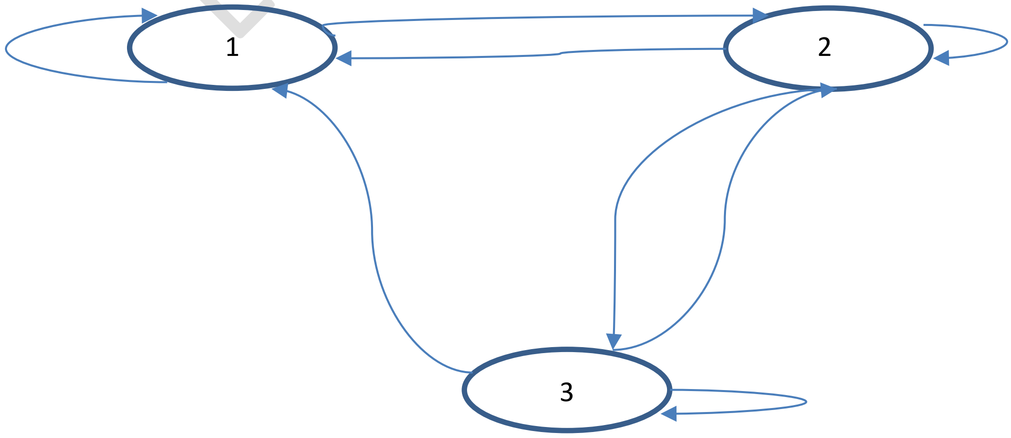
1. Proposer le graphe probabiliste associé.
2. Reprendre toutes les questions de 2 à 17.

   Pour la question 6, on admet que $A = PDP^{-1}$ où $P = \frac{1}{2}\begin{pmatrix} \sqrt{2} & -\sqrt{2} \\ \sqrt{2} & \sqrt{2} \end{pmatrix}$ et $D = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$

Situation A avec 3 sites : On complique la situation avec un troisième site noté 3. La chaine de Markov associé est donc à trois états.

On pose alors $U_n = \begin{pmatrix} P(S_n = 1) \\ P(S_n = 2) \\ \mathrm{P}(\mathrm{S}_n = 3) \end{pmatrix}$

On considère des choix aléatoires et le graphe suivant (certains liens n'existent pas)



1. Compléter le graphe probabiliste
2. Donner la matrice de transition.
3. Justifier que $P(S_n = 1) = P(S_n = 2)$
4. Déterminer X la loi invariante de cette chaine de Markov
5. Déterminer à l'aide de scilab P, D, P^{-1} puis en déduire U_∞
6. Comparer U_∞ et X.
7. Visualiser U_∞ à l'aide de la figure 2.
8. Conclure en termes de fréquentation des sites.
9. Vérifier en modifiant l'état initial que ce comportement est indépendant de l'état initial.

Situation B avec 3 sites et fréquentation connue.

Dans cette partie, on suppose que la matrice de transition est donnée par $A = \begin{pmatrix} \frac{2}{5} & \frac{1}{4} & \frac{1}{10} \\ \frac{2}{5} & \frac{1}{2} & \frac{2}{5} \\ \frac{1}{5} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}$

1. Définir le graphe probabiliste associé à A.
2. La commande V=kernel(A-eye(3,3)) fournit une base de l'ensemble des solutions de AX=X.
   Vérifier que $V \approx \begin{pmatrix} 0.39 \\ 0.74 \\ 0.53 \end{pmatrix}$ puis expliquer en utilisant la console comme calculatrice que $X \approx \begin{pmatrix} 0.24 \\ 0.44 \\ 0.32 \end{pmatrix}$
3. Vérifier avec scilab que $U_\infty \approx \begin{pmatrix} 0.24 \\ 0.44 \\ 0.32 \end{pmatrix}$.
4. Commentez les résultats observés sur les figures 2 et 3
5. Vérifier en modifiant l'état initial que la conclusion de l'exercice précédent semble toujours valable.
6. Quel semble être le site le moins fréquenté ? (Horreur….)

Situation C avec 3 sites et la jalousie du professeur dont le site est le moins fréquenté.

Dans la situation B, le professeur dont le site est le moins fréquenté (lequel) décide d'enlever les liens pointant de son site vers ceux de son collègue mais laisse un lien vers son propre site.

1. Donner le graphe probabiliste associé à cette situation.
2. Observez les résultats associés aux figures 2 et 3 en modifiant l'état initial.
3. Conclure ?
4. Si on extrapole à plus de sites cette attitude, à quoi conduirait-elle ?

```
//TP markov 2 sites
clear, clf

//Matrice de transition

//Situation 2 sites avec fréquentation connue
A=[0.6 0.2;0.4 0.8]

//Situation 2 sites avec architecture.
//A=[0.5 0.5;0.5 0.5]
disp(A,'A=')

//Diagonalisation de A
[P,D]=spec(A)
disp(P,'P=')
disp(D,'D=')
disp(P*D*inv(P),'PDP-1=')

//Définition de l'état initial
//U=[1;0]     //initial sur site 1
U=[0;1]   //initial sur site 2
disp(U,'U0=')

//Simulation de la trajectoire

//n=input('donner le nombre n de semaines')
n=100
//vecteur trajectoire
R=grand(n,'markov',A',1)
disp(R,'Trajectoire ')

//visualisation de la trajectoire.
scf(1)
plot2d(1:n,R,-1,rect=[0 0 n+1 3])

for k=1:n
    U=A*U
    M(:,k)=U
end

//Construction de la matrice M
disp(M,'M=')
```

```

//Evolution de la probabilité théorique P(Sn=i)
scf(2)
plot2d(1:n,M(1,:),1,rect=[0 0 100 1])
plot2d(1:n,M(2,:),2,rect=[0 0 100 1],leg="P(Sn=2)")

//Evolution de la fréquence empirique sur p simulations de n semaines et E0=1
//p=input("entrer le nombre de simulations")
p=10000
site1=ones(1,p)
//site2=2*ones(1,p)
Rp=grand(n,'markov',A',site1)    //p répétition avec initial site 1
//Rp=grand(n,'markov',A',site2)    //p répétition avec initial site 1
F1=zeros(1,n)
F2=zeros(1,n)
for k=1:n
    F1(k)=length(find(Rp(:,k)==1))/p          //fréquence du 1 en semaine k
    F2(k)=length(find(Rp(:,k)==2))/p          //fréquence du 2 en semaine k
end

scf(3)
plot2d([1:n],F1,1)
plot2d([1:n],F2,2,leg="site 2")
```