Processus de Décisions Markoviens (MDP) et application

François Delarue

Rémi Catellier

Table des matières

Chapitre 1

Chaînes de Markov contrôlées

Philosophie Générale

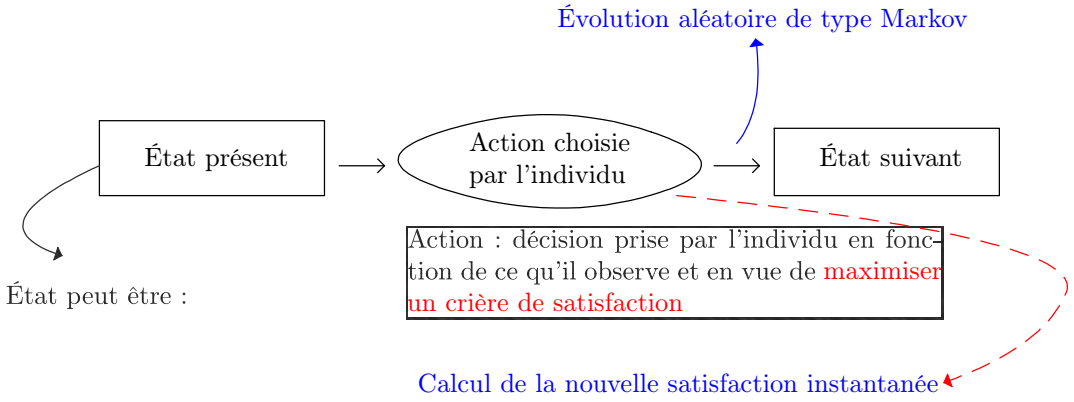

FIGURE 1. Chaînes de Markov contrôlées.

Modélisation. Il y a de nombreuses applications à ce modèle : finance, économie mais aussi biologie (ex : épidémiologie) ...

Paradigme : Dans cette phase de modélisation, on suppose CONNUES toutes les quantités régissant l'évolution du système :

- La forme des transitions aléatoires (les réalisations ne sont pas connues, mais la distribution est connuyes en fonction des états occupés et des actions suives
- La forme des récompenses (ou satisfactions) en focntion des états occupés et des actions suives.

Naturellement les décisions ne sont pas connues à l'avance.

OBJECTIF : Caractériser mathématiquement les meilleurs décisions

Appentissage. En pratique, on ne connait pas les quantités mathématiques régissant le modèle. Les seules choses observées sont les états occupés, les actions décidés et les récompenses réçues.

QUESTION : Comment apprendre la meilleurs décision à « jouer » en fonction des observations ?

C'est une préoblématique typique des l'apprentissage satistique. On parle d'*apprentissage par renforcement* (une des branches de l'apprentissage, avec l'apprentissge supervisé et l'apprentissage non supervisé).

Intuitivement, la « pertinence » de la décision apprise doit évoluer de façon croissante avec le nombre d'observations. Idée générale est que pour comprendre l'apprentissage, il faut bien comprendre la modélisation.

1. Quelques rappels sur les chaînes de Markov

ATTENTION : Ici pas de contrôle (d'action). Le cadre correspond à celui étudié en cours de processus stochastiques.

1.1. Principe général. Une chaîne de Markov décrit l'évolution (en cours du temps) d'un système *aléatoire*, sans mémoire. Il 'agit d'un version probabiliste d'une suite récursive de la forme

$$x_{n+1} = f(n, x_n) \tag{1}$$

Ici, x_{n+1} désigne l'état de l'individu à l'instant $n+1$, la fonction *déterminste* $f(n, \cdot)$ explique le passage de l'état courant à l'état suivant et x_n est l'état de l'individu à l'instant n.

REMARQUE 1.1. *(1) La notion d'état est ici la même que précédement*
(2) Il n'y a pas de possibilité pour l'individu de prendre une décision
(3) Aucun aléa à ce stade. Le passage de l'état x_n à l'état x_{n+1} est déterministe.

DÉFINITION 1.2. On appelle espace d'états l'ensemble $\mathcal{S}$ des valeurs pour les états de l'individu, à n'importe quel instant. Autrement dit, on suppose dans (1)

(1) $x_0 \in \mathcal{S}$
(2) $\forall n \in \mathbb{N}$, $f(n, \cdot) : \mathcal{S} \mapsto \mathcal{S}$, ie $f \in \mathbb{N} \times \mathcal{S} \mapsto \mathcal{S}$

Dans la suite, on suppose $\mathcal{S}$ au plus dénombrable et souvent fini. ATTENTION : à ce stade, pas d'aléa.

En présence d'aléa, il y a une incertitude sur ce que vaut x_{n+1}, même en connaissant x_n. On peut donc reprendre le grapique introductif.

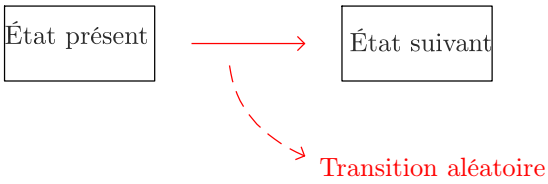

FIGURE 2. Chaîne de Markov

1.2. Matrice de transition. La fonction $f(n, \cdot)$ est remplacée par une matrice décrivant les probabilités de transiter d'un état à un autre.

$$(f(n, i))_{i \in \mathcal{S}} \longrightarrow (P(n, i, j))_{i,j \in \mathcal{S}} \tag{2}$$

DÉFINITION 1.3. Une collection de réels $(Q(i, j))_{i,j \in \mathcal{S}}$ forme une *matrice de transition* (sur $\mathcal{S}$) si pour tout $i \in \mathcal{S}$, $Q(i, \cdot)$ forme une collection de poids de probilité sur $\mathcal{S}$, c'est à dire

(1) $\forall j \in \mathcal{S}$, $Q(i, j) \geq 0$.
(2) $\sum_{j \in \mathcal{S}} Q(i, j) = 1$.

On suppose desormais que dans (2) pour tout $n \in \mathbb{N}$, $P(n, \cdot, \cdot)$ est une matrice de transition.

1.3. Chaîne de Markov. On note $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité.

DÉFINITION 1.4. Soit $(P(n, \cdot, \cdot))_{n \in \mathbb{N}}$ de matrices de transitions sur $\mathcal{S}$. Une collection $(X_n)_{n \in \mathbb{N}}$ de variables aléatoires à valeurs dans $\mathcal{S}$ (formant ainsi un processus indéxé par le temps discret) constitue une chaîne de Marov de matrices de transiçtion $(P(n, \cdot, \cdot))_{n \in \mathbb{N}}$ si l'une des définitions suivantes est vérifiée :

(1) Pour tout $n \in \mathbb{N}$n pour tout $i_0, \cdots, i_n, j \in \mathcal{S}$

$$\mathbb{P}(X_0 = i_0, \cdots, X_n = i_n, X_{n+1} = j) = \mathbb{P}(n, i_n, j)\mathbb{P}(X_0 = i_0, \cdots, X_n = i_n)$$

(2) Pour tout $n \in \mathbb{N}$n pour tout $i_0, \cdots, i_n, j \in \mathcal{S}$ tels que $\mathbb{P}(X_0 = i_0, \cdots, X_n = i_n) > 0$,

$$\mathbb{P}(X_{n+1} = j | X_0 = i_0, \cdots, X_n = i_n) = \mathbb{P}(X_{n+1} = j | X_n = i_n) = P(n, i_n, j).$$

On peut généraliser cette définition en utilisant le concept de filtration :

Définition 1.5. Soit $\mathcal{G}$ une sous tribu de $\mathcal{F}$, soit $X : \Omega \to E$ une variable aléatoire à valers dans E dénombrable, on appelle loi conditionnelle de X sachant $\mathcal{G}$ la collection de variables aléatoires

$$(\mathbb{P}(X = i|\mathcal{G}))_{i \in E} := \mathbb{E}[\mathbb{1}_{\{X=i\}}|\mathcal{G}]$$

Remarque 1.6. *Si $\mathcal{G} = \sigma(Y)$ où $Y : \Omega \to F$ est une variable aléatoire dans un espace dénombrable F, alors, sur $(Y = j)$ pour $j \in F$ tel que $\mathbb{P}(Y = j) > 0$, on a*

$$\mathbb{E}[\mathbb{1}_{\{X=i\}}|\mathcal{G}] = \sum_{\substack{j \in F \\ \mathbb{P}(Y=j)>0}} \mathbb{P}(X = i|Y = j)\mathbb{1}_{Y=j}.$$

Exercice 1. Prouver la remarque précédente.

Solution 1. On rappel que

$$\sigma(Y) = \{Y^{-1}(B) : B \in \mathcal{P}(F)\},$$

où $\mathcal{P}(F)$ désigne l'ensemble des parties de F. De plus, comme F est dénombrable, on a

$$Y^{-1}(B) = \coprod_{j \in B} Y^{-1}(\{j\}),$$

où l'union est disjointe et au plus dénombrable. Ainsi, grâce au propriétés de l'espérance conditionnelle,

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[\mathbb{1}_{\{X=i\}}|\mathcal{G}]\mathbb{1}_{Y^{-1}(B)}] =& \mathbb{E}[\mathbb{1}_{\{X=i\}}\mathbb{1}_{Y^{-1}(B)}] \\
=& \sum_{j \in B} \mathbb{E}[\mathbb{1}_{X=i}\mathbb{1}_{Y^{-1}(\{j\})}] \\
=& \sum_{j \in B} \mathbb{P}(X = i, Y = j) \\
=& \sum_{\substack{j \in B \\ \mathbb{P}(Y=j)>0}} \mathbb{P}(X = i|Y = j)\mathbb{P}(Y = j) \\
=& \mathbb{E}\left[\sum_{\substack{j \in F \\ \mathbb{P}(Y=j)>0}} \mathbb{P}(X = i|Y = j)\mathbb{1}_{Y=j}\mathbb{1}_{Y^{-1}(B)} \right].
\end{aligned}$$

où l'avant dernière égalité provient de la formule des probabilités totales. Notons également que

$$\sum_{\substack{j \in F \\ \mathbb{P}(Y=j)>0}} \mathbb{P}(X = i|Y = j)\mathbb{1}_{Y=j} \in \mathcal{G} = \sigma(Y).$$

Ainsi, grâce à la définition de l'espérance conditionnelle, on vient de montrer que

$$\mathbb{E}[\mathbb{1}_{\{X=i\}}|\mathcal{G}] = \sum_{\substack{j \in F \\ \mathbb{P}(Y=j)>0}} \mathbb{P}(X = i|Y = j)\mathbb{1}_{Y=j} p.s.$$

Définition 1.7. On appelle filtration sur $(\Omega, \mathcal{F}, \mathbb{P})$ toute collection $(\mathcal{F}_n)_n$ croissante de sous-tribu de $\mathcal{F}$. Autrement dit pour tout $n \in \mathbb{N}$, $\mathcal{F}_n$ est une tribu et $\mathcal{F}_n \subset \mathcal{F}_{n+1}$.

Définition 1.8. Étant donnée une suite $(P(n,\cdot,\cdot))_{n\in\mathbb{N}}$ de matrices de transition sur $\mathcal{S}$ et une filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ sur $(\Omega,\mathcal{F},\mathbb{P})$, une collection $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires à valeur dans $\mathcal{S}$ est une $(\mathcal{F}_n)_{n\in\mathbb{N}}$ chaîne de Markov de matrices de transition $(P(n,\cdot,\cdot))_{n\in\mathbb{N}}$ si

(1) $\forall n\in\mathbb{N}$, X_n est $\mathcal{F}_n$-mesurable
(2) $\forall n\in\mathbb{N}$, $\forall i\in\mathcal{S}$,

$$\mathbb{P}(X_{n+1}=i|\mathcal{F}_n)=\mathbb{P}(n,X_n,i)\quad \mathbb{P}-p.s.$$

Remarque 1.9. *Toute chaîne de Markov au sens de la première définition est une $(\sigma(X_0,\cdots,X_n))_{n\in\mathbb{N}}$ chaîne de Markov. En effet, sur $(X_n=j)$, avec $\mathbb{P}(X_n=j)>0$,*

$$\mathbb{P}(X_{n+1}=i|\mathcal{F}_n)=\mathbb{P}(X_{n+1}=i|X_n=j)=P(n,j,i)=P(n,X_n,i).$$

La première égalité vient de la définition d'une chaîne de Markov, la deuxième vient de la définition de la matrice de transition, et la troisième du fait que nous travaillons sur $(X_n=j)$.

1.4. Construction canonique. On donne ici la construction canonique d'une chaîne de Markov généralisant lma définition d'une chaîne déterminiqte sans mémoire donnée dans la Partie 1.1.

Proposition 1.10. *Soit $f:\mathbb{N}\times\mathcal{S}\times[0,1]\to\mathcal{S}$ une fonction mesurable.*
Soient $X_0:\Omega\to\mathcal{S}$ et $(U_n)_{n\in\mathbb{N}\setminus\{0\}}$ une variable aléatoire et une suite de variables aléatoires telles que

(1) Pour tout $n\in\mathbb{N}$, $U_n\sim\mathcal{U}([0,1])$.
(2) $(U_n)_{n\in\mathbb{N}\setminus\{0\}}$ est une suite de variables aléatoires indépendantes.
(3) X_0 et $(U_n)_{n\in\mathbb{N}\setminus\{0\}}$ sous indépendantes.

Alors la suite $(X_n)_{n\in\mathbb{N}}$ définie pour tout $n\in\mathbb{N}$ par

$$X_{n+1}=f(n,X_n,U_{n+1})$$

est une chaîne de Markov de matrices de transition

$$((P(n,i,j))_{i,j\in\mathcal{S}})_{n\in\mathbb{N}}=((\mathbb{P}(f(n,i,U_1)=j))_{i,j\in\mathcal{S}})_{n\in\mathbb{N}}.$$

Remarque 1.11. *Les points 2) et 3) sont équivalents à dire que pour tout $n\in\mathbb{N}$, $X_0,U_1,\cdots,U_n$ soint indépendants.*

Exercice 2. Prouver la proposition précédente.

Solution 2. On remarque par une récurence imémdiate que pour tout $n\in\mathbb{N}$, X_n est $\sigma(X_0,U_1,\cdots,U_n)$ mesurable. En particulier pour tout $n\in\mathbb{N}$ et tout $k\leq n$, X_k est indépendant de U_{n+1}.
Soit $n\in\mathbb{N}$, $i_0,\cdots,i_n,j\in\mathcal{S}$. On a

$$\begin{aligned}\mathbb{P}(X_{n+1}=j,X_0=i_0,\cdots,X_n=i_n)=&\mathbb{P}(f(n,X_n,U_{n+1})=j,X_0=i_0,\cdots,X_n=i_n)\\=&\mathbb{P}(f(n,i_n,U_{n+1})=j,X_0=i_0,\cdots,X_n=i_n)\\=&\mathbb{P}(f(n,i_n,U_{n+1})=j)\mathbb{P}(X_0=i_0,\cdots,X_n=i_n).\end{aligned}$$

La premièer égalité provient de la définition de $(X_n)_{n\in\mathbb{N}}$. La seconde égalité vient du fait que l'on a spécifié $X_n=i_n$ et la troisième égalité vient de l'indépendance de $(X_0,\cdots,X_n)$ avec U_{n+1}.

Proposition 1.12. *Toute chaine de Markov d'espace d'état $\mathcal{S}$ et de matrices de transition $((P(n,i,j))_{i,j\in\mathcal{S}})_{n\in\mathbb{N}}$ admet une forme canonique.*

Exercice 3. Prouver la proposition précédente. *Indication : soit $n\in\mathbb{N}$ and $i\in\mathcal{S}$, soit $f(n,i,\cdot)$ la fonction quantile de la loi de poids $(P(n,i,j))_{j\in\mathcal{S}}$.*

Solution 3. Comme $\mathcal{S}$ est au plus dénombrable, il existe une suite $(j_k)_{k\in\mathbb{N}}$ telle $\mathcal{S} = \coprod_{k\geq 0}\{j_k\}$. Pour $x \in [0,1]$, $n \in \mathbb{N}$ et $i \in \mathcal{S}$, définissons

$$f(n,i,x) = j_0 \mathbb{1}_{[0,P(n,i,j_0))}(x) + \sum_{k\geq 0} j_{k+1} \mathbb{1}_{\left[\sum_{\ell=0}^{k} P(n,i,j_\ell), \sum_{\ell=0}^{k+1} P(n,i,j_\ell)\right)}(x).$$

Soit $U \sim [0,1]$ et $k \in \mathbb{N}$. Alors

$$\begin{aligned}\mathbb{P}(f(n,i,U) = j_k) =& \mathbb{P}\left(U \in \left[\sum_{\ell=0}^{k-1} P(n,i,j_\ell), \sum_{\ell=0}^{k} P(n,i,j_\ell)\right)\right) \\ =& \sum_{\ell=0}^{k} P(n,i,j_\ell) - \sum_{\ell=0}^{k-1} P(n,i,j_\ell) = P(n,i,j_k).\end{aligned}$$

Ici nous avons utilisé la convention $\sum_{\ell=0}^{-1} P(n,i,j_\ell) = 0$. Notons également que f est une fonction mesurable. Sont $(U_n)_{n\in\mathbb{N}\setminus\{0\}}$ une suite de variable aléatoire iid de loi $\mathcal{U}([0,1])$ indépendente de X_0. Soit pour $n \in \mathbb{N}$, $\tilde{X}_{n+1} = f(n, \tilde{X}_n, U_{n+1})$ avec $\tilde{X}_0 = X_0$. Par la Proposition 1.10, $(\tilde{X}_n)_{n\in\mathbb{N}}$ est une chaîne de Markov de matrices de transition définie pour $n \in \mathbb{N}$ et $i, j \in \mathcal{S}$ par

$$\tilde{P}(n,i,j) = \mathbb{P}(f(n,i,U_1) = j) = P(n,i,j).$$

Ainsi, par récurrence, pour tout $n \in \mathbb{N}$ et $i_0, \cdots, i_n \in \mathcal{S}$,

$$\mathbb{P}(X_0 = i_0, \cdots, X_n = i_n) = \mathbb{P}(X_0 = i_0) \prod_{k=0}^{n-1} P(k, i_k, i_{,k+1}) = \mathbb{P}(\tilde{X}_0 = i_0, \cdots, \tilde{X}_n = i_n),$$

et $(\tilde{X}_n)_{n\in\mathbb{N}}$ et $(X_n)_{n\in\mathbb{N}}$ ont la même loi, ce qui est l'ennoncée de la proposition.

Remarque 1.13. *Nous avons prouvé au passage que la données de la suite des matrices de transition $(P(n,\cdot,\cdot))_{n\in\mathbb{N}}$ ainsi que de la condition initiale X_0 caractérisent la loi de la chaîne de Markov $(X_n)_{n\in\mathbb{N}}$.*

2. Chaînes contrôlées

Nous avons maintenant tous les éléments pour définir un système contrôlé, analogue des chaînes de Markov.

2.1. Cas déterministe. Commençons par examiné ce que serait la version déterministe. Dans l'écriture

$$x_{n+1} = f(n, x_n)$$

donnée à l'Équation (1). On doit ajouter la décision prise pas l'individu à l'insant n.

Définition 2.1. On appel *espace d'actions* un ensembla $\mathcal{A}$ contenant toutes les actions (ou tous les choix) possibles pour l'individu à n'importe quel instant. Dans un soucis de simplication, on suppose de fait que $\mathcal{A}$ est indépendant du temps (mais ceci pourrait être généralisé au cas où les actions possibles varient avec le temps).

Dans la suite, $\mathcal{A}$ est souvent au plus dénombrable, mais on peut également avoir des cas où $\mathcal{A}$ est un Borélien de $\mathbb{R}^d$, pour un certain $d \geq 1$. On fait donc cette hypothèse ici.

Nous sommes donc amené à considérer

$$f : \mathbb{N} \times \mathcal{S} \times \mathcal{A} \to \mathcal{S}$$

mesurable, c'est à dire pour tout $(n,i) \in \mathbb{N} \times \mathcal{S}$, $f(n,i,\cdot) : \mathcal{A} \to \mathcal{S}$ est mesurable au sens classique, c'est à dire que pour tout $j \in \mathcal{S}$,

$$\{a \in \mathcal{A} : f(n,i,a) = j\} \in \mathcal{B}(\mathbb{R}^d).$$

Exercice 4. Prouver que la définition précédente est compatible avec la définition standard de la mesurabilité.

Solution 4. Remarquons que $f : \mathbb{N} \times \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ est mesurable si pour tout $B \in \mathcal{P}(\mathcal{S})$, $f^{-1}(B)$ est mesurable pour la tribu sur $\mathbb{N}\times\mathcal{S}\times\mathcal{A}$. Notons que comme $\mathcal{S}$ est au plus dénombrable, la tribu à l'arrivée est naturellement l'ensemble des parties de $\mathcal{S}$, notée $\mathcal{P}(\mathcal{S})$. Si $B \in \mathcal{P}(\mathcal{S})$, il existe une suite $(j_k)_{k\in\mathbb{N}} \in \mathcal{S}^{\mathbb{N}}$ telle que $B = \bigcup_{k\in\mathbb{N}}\{j_k\}$ et $f^{-1}(B) = \bigcup_{k\in\mathbb{N}} f^{-1}(\{j_k\})$

Notons également que pour $j \in \mathcal{S}$,

$$f^{-1}(\{j\}) = \coprod_{\substack{n\in\mathbb{N}\\ i\in\mathcal{S}}} \Big(f^{-1}(\{j_k\})\bigcap(\{n\}\times\{i\}\times\mathcal{A})\Big)$$

et que

$$\begin{aligned} f^{-1}(\{j\})\bigcap(\{n\}\times\{i\}\times\mathcal{A}) =& \{(m,l,a)\in\mathbb{N}\times\mathcal{S}\times\mathcal{A} : f(m,l,a)=j\}\bigcap(\{n\}\times\{i\}\times\mathcal{A}) \\ =& \{(n,i,a) : a\in\mathcal{A},\, f(n,i,a)=j\} \end{aligned}$$

Comme f est mesurable,

$$f^{-1}(\{j\})\bigcap(\{n\}\times\{i\}\times\mathcal{A}) = \{(n,i,a) : a\in\mathcal{A},\, f(n,i,a)=j\} \in \mathcal{P}(\mathbb{N})\times\mathcal{P}(\mathcal{S})\times(\mathcal{A}\cap\mathcal{B}(\mathbb{R}^d)),$$

ce qui est équivalent à

$$\{a\in\mathcal{A} : f(n,i,a)=j\} \in \mathcal{A}\cap\mathcal{B}(\mathbb{R}^d),$$

et qui prouve le résultat.

Définition 2.2. Une suite $(x_n, a_n)_{n\in\mathbb{N}}$ à valeurs dans $\mathcal{S}\times\mathcal{A}$ forme une chaîne sans mémoire contrôlée par $(a_n)_{n\in\mathbb{N}}$ et dirigée par f si pour tout $n \in \mathbb{N}$,

$$x_{n+1} = f(n, x_n, a_n).$$

Attention : il n'y a pas de dynamique pour $(a_n)_{n\in\mathbb{N}}$, chaque a_n est choisi librement par l'individu dont l'évolution est représenté par la chaîne.

2.2. Matrices de transition contrôlées. On veut maintenant généraliser la définition données dnas la précédente partie au cas stocastique en nous appuyant sur le concept de matrices de transition vu dans la Section 1.3.

Définition 2.3. On appelle matrices de transition contrôlées toute application mesurable

$$P : \mathbb{N}\times\mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0,1]$$

telle que pour tout $n, a \in \mathbb{N}\times\mathcal{A}$, la collection $(P(n,i,a,j))_{i,j\in\mathcal{S}}$ forme une matrice de transition, c'est à dire que pour tout $(n,i,a) \in \mathbb{N}\times\mathcal{S}\times\mathcal{A}$, $P(n,i,a,\cdot)$ est une famille de poids de probbilité sur $\mathcal{S}$.

Remarque 2.4. *On dit que la collection est homogène si $P(n,i,a,j) = P(i,a,j)$ pour tout $(n,i,a,j) \in \mathbb{N}\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, et pour un certain $P : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0,1]$ mesurable, de sorte que pour tout $a \in \mathcal{A}$, $P(\cdot,a,\cdot)$ est une matrice de transition.*

L'interprétation de la Définition 2.3 est que si l'individu est dans l'état i à l'instant n, alors, en choisissant l'action $a \in \mathcal{A}$, il a une probabilité $P(n,i,a,j)$ de se retrouver dans l'état j à l'instant suivant $n+1$.

Exemple 2.5. On considère un service informatique soumis régulièrement à des attaques de hackers. Le service a deux états :

- L'état 0 : le système est sain, non infecté
- Létat 1 : le système est attaqué

A chaque instant $n \in \mathbb{N}$, si le système est dans l'état 0, il y a deux choix :

- Ne rien faire. La protection du système est dégradéen mais cela ne coûte rien au manager du système
- Mettre à jour la protection. La protection du sytème est renforcée, mais il y a des frais de mise à jour.

En état 1, il y a deux choix :

- Ne rien faire : le système fonctionne en mode dégradé. Cela ne coûte rien en terme de réparations, mais cela a un coût sur les activités autours
- Réparer le système. Cela a un coût pour le service informatique mais préserve les activités périphériques.

On a donc $\mathcal{A} = \{0, 1\}$, où 0 représente : « ne rien faire » et 1 représente « agir ». On pose alors

$$P(\underbrace{0}_{\text{état}}, \overbrace{0}^{\text{action}}, \underbrace{1}_{\text{état}}) = p \in [0,1], \quad P(0,0,0) = 1-p$$

En mode sain, sans rien faire, il y a une probabilité p de passer en mode dégradé et une probabilité $1-p$ de rester en mode sain.

$$P(0,1,1) = q, \quad P(0,1,0) = 1-q,$$

En mode sain, si on agit, il y a une probbilité q de passer en mode dégradé et une probabilité $1-q$ de rester en mode sain.

$$P(1,0,1) = 1, \quad P(1,0,0) = 0,$$

En mode dégradé, si on ne fait rien on reste en mode dégradé.

$$P(1,1,0) = 1, \quad P(1,1,1) = 0,$$

En mode dégradé, si on agit on passe en mode sain. On obtien le graphe suivant :

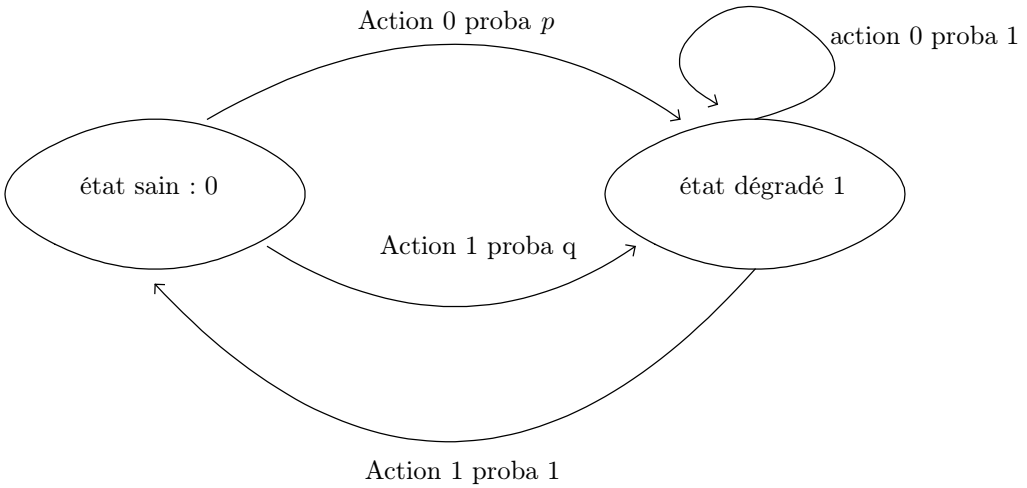

FIGURE 3.

REMARQUE 2.6. *Le modèle est simple, on a oublié la protection passée dans la probabilité d'être affacté par une attaque. On pourrait ajouter un état « protégé mais pas mis à jour » par exemple. Par ailleurs nour n'avons pas encore inclu le coût de chaque action dans la modélisation.*

2.3. Chaînes de Markov contrôlées. On veut maintenant associer une suite d'états et une suite d'actions aux matrices de transition contrôlées. Il y a néanmoins une subtilité : quelles sont les onformations disponibles à l'individu (ou à l'observateur) pour prendre une décision à un instnat donnée ? Il s'agit de fait d'une question de mesurabilité.

On commence par la définition suivante, qui généralise celle introduite pour les chaînes de Markov.

DÉFINITION 2.7. Etant donnée une filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ et une familele de matrices de transitions contrôlées

$$P : \mathbb{N} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1],$$

une suite $(X_n, A_n)_{n\in\mathbb{N}}$ de variables aléatoires à valeurs dans $\mathcal{S}\times\mathcal{A}$ est une $(\mathcal{F}_n)_{n\in\mathbb{N}}$ chaîne de Markov contrôlée associée à P si

(1) Pour tout $n\in\mathbb{N}$, (X_n, A_n) est $\mathcal{F}_n$ mesurable.
(2) Pour tout $n\in\mathbb{N}$ et pour tout $j\in\mathcal{S}$,

$$\mathbb{P}(X_{n+1}=j|\mathcal{F}_n)=P(n,X_n,A_n,j).$$

Ilustrons cette dernière propriété dans le cas où $\mathcal{A}$ est au plus dénombrable. Soit $n\in\mathbb{N}$, $i_0,\cdots,i_n,j\in\mathcal{S}$ et $a_0,\cdots,a_n\in\mathbb{N}$.

$$\begin{aligned}\mathbb{P}(X_0=i_0,A_0=a_0,\cdots,X_n=i_n,A_n=a_n,X_{n+1}=j)&\\ =P(n,i_n,a_n,j)\mathbb{P}(X_0=i_0,A_0=a_0,\cdots,X_n=i_n,A_n=a_n).&\quad(3)\end{aligned}$$

Démonstration. On a $(X_0=i_0,A_0=a_0,\cdots,X_n=i_n,A_n=a_n)\in\mathcal{F}_n$ par définition de la mesurabilité. Ainsi

$$\begin{aligned}\mathbb{P}(X_0=i_0,A_0=a_0,\cdots,X_n=i_n,A_n=a_n,X_{n+1}=j)&=\mathbb{E}[\mathbb{1}_{X_{n+1}=j}\mathbb{1}_{X_0=i_0,A_0=a_0,\cdots,X_n=i_n,A_n=a_n}]\\ &=\mathbb{E}[\mathbb{E}[\mathbb{1}_{X_{n+1}=j}|\mathcal{F}_n]\mathbb{1}_{X_0=i_0,A_0=a_0,\cdots,X_n=i_n,A_n=a_n}]\\ &=\mathbb{E}[P(n,X_n,A_n,j)\mathbb{1}_{X_0=i_0,A_0=a_0,\cdots,X_n=i_n,A_n=a_n}]\\ &=P(n,i_n,a_n,j)\mathbb{P}(X_0=i_0,A_0=a_0,\cdots,X_n=i_n,A_n=a_n)\quad(4)\end{aligned}$$

□

Dans le cas où $\mathcal{A}$ n'est pas dénombrable, on ne peut plus procéder comme ci-dessus : les évènements peuvent être de probabilité nulle (en particulier si les $A_0,\cdots,A_n$ sont à densité.

Proposition 2.8. *Dans le cas général, si $B_0,\cdots,B_n\in\mathcal{B}(\mathcal{A})$, et $i_0,\cdots,i_n,j\in\mathcal{S}$,*

$$\begin{aligned}\mathbb{P}(X_0=i_0,A_0\in A_0,\cdots,X_n=i_n,A_n\in B_n,X_{n+1}=j)&\\ =\int_{\mathcal{S}^{n+1}\times(\mathbb{R}^d)^{n+1}}P(n,i_n,a_n,j)\mathbb{1}_{B_0\times\cdots\times B_n}(a_0,\cdots,a_n)\mathbb{1}_{\{i_0,\cdots,i_n\}}(j_0,\cdots,j_n)&\\ \mathrm{d}\mathbb{P}_{(X_0,A_0,X_1,A_1,\cdots,X_n,A_n)}(j_0,a_0,\cdots,j_n,a_n).&\quad(5)\end{aligned}$$

Démonstration. Comme précédemment, $(X_0=i_0,A_0\in A_0,\cdots,X_n=i_n,A_n\in B_n)\in\mathcal{F}_n$, ainsi

$$\begin{aligned}\mathbb{P}(X_0=i_0,A_0\in A_0,\cdots,X_n=i_n,A_n\in B_n,X_{n+1}=j)&\\ =\mathbb{E}[P(n,X_n,A_n,j)\mathbb{1}_{(X_0=i_0,A_0\in A_0,\cdots,X_n=i_n,A_n\in B_n)}]&\\ =\mathbb{E}[P(n,i_n,A_n,j)\mathbb{1}_{(X_0=i_0,A_0\in A_0,\cdots,X_n=i_n,A_n\in B_n)}],&\quad(6)\end{aligned}$$

ce qui prouve la proposition quand on écrit les espérance à l'aide des lois des variables aléatoires. □

Application : Si $\mathbb{P}(X_0=i_0,\cdots,X_n=i_n)>0$, on appelle loi conditionnelle de $(A_0,\cdots,A_n)$ sachant $(X_0=i_0,\cdots,X_n=i_n)$ la mesure de probabilité

$$\mathbb{P}_{(A_0,\cdots,A_n)|(X_0=i_0,\cdots,X_n=i_n)}(B_0\times\cdots\times B_n)=\mathbb{P}(A_0\in B_0,\cdots,A_n\in B_n|X_0=i_0,\cdots,X_n=i_n).$$

Alors

$$\begin{aligned}\mathbb{P}(X_0=i_0,A_0\in A_0,\cdots,X_n=i_n,A_n\in B_n,X_{n+1}=j)&\\ =\int_{\mathcal{A}^{n+1}}P(n,i_n,a_n,j)\mathbb{P}_{(A_0,\cdots,A_n)|(X_0=i_0,\cdots,X_n=i_n)}(\mathrm{d}a_0,\cdots,\mathrm{d}a_n)&\quad(7)\end{aligned}$$

2.4. Mesurabilité des actions. On distingue 3 types de propriétés de mesurabilité pour $(A_n)_{n\in\mathbb{N}}$.

Définition 2.9. On dit que les actions $(A_n)_{n\in\mathbb{N}}$ s'écrivent sous forme fermé dépendant du passé si pour tout $n \in \mathbb{N}$, A_n est $\sigma(X_0, \cdots, X_n)$ mesurable, c'est à dire si $A_n = f_n(X_0, \cdots, X_n)$.
On choisit l'action à l'instant n en fonction des observations des états passés de 0 à n.

Définition 2.10. On dit que les actions $(A_n)_{n\in\mathbb{N}}$ s'écrivent sous une forme fermée markovienne si pour tout n, A_n est $\sigma(X_n)$ mesurable, c'est à dire qu'il existe une fonction mesurable f_n telle que $A_n = f_n(X_n)$.
On choisit l'action à l'instant n en fonctino de la seule observation de l'état à l'instant n.

Définition 2.11. On dit que les actions $(A_n)_{n\in\mathbb{N}}$ s'écrivent sous forme mixte markovienne s'il existe une suite $(V_n)_{n\in\mathbb{N}}$ de variables aléatoires uniformes sur $[0, 1]$ telle que

(1) pour tout $n \in \mathbb{N}$, V_n est $\mathcal{F}_n$-mesurable.
(2) pour tout $n \in \mathbb{N}$, V_n est indépedante de $\sigma(X_n) \vee \mathcal{F}_{n-1}$
(3) pour tout $n \in \mathbb{N}$, A_n est $\sigma(X_n, V_n)$ mesurable.

A l'instant n, on ajoute, en plus de l'observation de l'état X_n, à l'instant n, un tirage indépendant pour décider de l'action à suivre.

Exercice 5 (Forme canonique des chaînes de Markov contrôlées 1). On suppose $\mathcal{A}$ fini. Soit $P : \mathbb{N} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ une collection de matrices de transition contrôlées et soit $(f_n : \mathcal{S} \to \mathcal{P}(\mathcal{A}))_{n\in\mathbb{N}}$ une suite de fonction de $\mathcal{S}$ à valeur de $\mathcal{P}(\mathcal{A})$, l'ensemble des mesures de probabilités sur $\mathcal{A}$.

Montrer qu'il existe une fonction $G : \mathcal{P}(\mathcal{A}) \times [0, 1] \to \mathcal{A}$ telle que si $(V_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoire iid de loi $\mathcal{U}([0, 1])$ indépendante de X_0, et que si $(X_n, A_n)_{n\in\mathbb{N}}$ est définit par la relation de récurrence

$$A_n = G(f_n(X_n), V_n), \quad X_{n+1} \sim P(n, X_n, A_n, \cdot),$$

alors $(X_n, A_n)_{n\in\mathbb{N}}$ est une chaîne de Markov contrôlée de matrices de transitions contrôlées P avec une suite de contrôles $(A_n)_{n\in\mathbb{N}}$ de forme mixte markovienne.

Solution 5. On remarque que si $\mathcal{A} = \{a_1, \cdots, a_{|\mathcal{A}|}\}$, l'ensemble des mesures de probabilité sur $\mathcal{A}$ peut être décrit sous la forme

$$\mathcal{P}(\mathcal{A}) = \left\{ \mu = (p_1, \cdots, p_{|\mathcal{A}|}) \in [0, 1]^{|\mathcal{A}|} : \sum_{i=1}^{|\mathcal{A}|} p_i = 1 \right\}.$$

via l'identification $\mu = (p_1, \cdots, p_{|\mathcal{A}|}) \in \mathcal{P}(A)$, $\mu(\{a_k\}) = p_k$. De la même façon que pour la forme canonique des chaîne de Markov (non contrôlées), on écrit

$$G(\mu, u) = a_1 \mathbb{1}_{[0,\mu(a_1))}(u) + \sum_{i=2}^{|A|} a_i \mathbb{1}_{[\mu(a_1)+\cdots+\mu(a_{i-1}),\mu(a_1)+\cdots+\mu(a_i))}(u).$$

Soit $(V_n)_{n\in\mathbb{N}}$ une suite de variable iid de loi $\mathcal{U}([0, 1])$

$$A_0 = G(f_0(X_0), U_0).$$

Remarquons aussi que pour $n \in \mathbb{N}$, $i \in \mathcal{S}$ et $a \in \mathcal{A}$, $(P(n, i, a, j))_{j\in\mathcal{S}}$ définit une mesure de probabilité sur $\mathcal{S}$, de la même façon que pour les mesures sur $\mathcal{A}$. Soit $X_1 \sim P(0, X_0, A_0, \cdot)$. On définit alors $A_1 = G(f_1(X_1), V_1)$, de telle sorte que A_1 est bien un contrôle mixte (en fonction de X_1 et V_1). Par récurrence maintenant, on suppose que $X_0, A_0, \cdots, X_n, A_n$ sont construits. On note alors $X_{n+1} \sim P(n, X_n, A_n, \cdot)$, de telle sorte que

$$
\begin{aligned}
&(X_0 = i_0, A_0 = a_0, \cdots, X_n = i_n, A_n = a_n, X_{n+1} = j) \\
&\quad = \mathbb{E}[\mathbb{E}[\mathbb{1}_{X_{n+1}=j} | X_0 = i_0, A_0 = a_0, \cdots, X_n = i_n, A_n = a_n] \mathbb{1}_{X_0=i_0, A_0=a_0, \cdots, X_n=i_n, A_n=a_n}] \\
&\qquad = \mathbb{E}[P(n, X_n, A_n, j) \mathbb{1}_{X_0=i_0, A_0=a_0, \cdots, X_n=i_n, A_n=a_n}] \\
&\qquad = P(n, i_n, a_n, j) \mathbb{P}(X_0 = i_0, A_0 = a_0, \cdots, X_n = i_n, A_n = a_n). \quad (8)
\end{aligned}
$$

On définit alors $X_{n+1} = G(f_{n+1}(X_{n+1}), U_{n+1})$. On a bien définit une chaîne de Markov contrôlée $(X_n, A_n)_{n\in\mathbb{N}}$, de matrices de transition contrôlées P et telle que $(A_n)_{n\in\mathbb{N}}$ est sous forme Markovienne mixte.

Exercice 6 (Forme canonique des chaînes de Markov contrôlées 2). On suppose $\mathcal{A}$ fini. Pour tout $n \in \mathbb{N}$ on se donne $f_n : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, où ici $\mathcal{P}(\mathcal{A})$ désigne l'ensemble des probabilités sur $\mathcal{A}$. $F : \mathbb{N} \times \mathcal{S} \times \mathcal{A} \times [0,1] \to \mathcal{S}$ mesurable. Soit $(U_n)_{n\in\mathbb{N}}$ et $(V_n)_{n\in\mathbb{N}}$ deux suites de variables aléatoires iid de loi $\mathcal{U}([0,1])$ et indépendantes de X_0. Soit $G : \mathcal{P}(A) \times [0,1] \to \mathcal{A}$ la fonction définie à l'exercice précédent.

Montrer que la suite de variables aléatoires $(X_n, A_n)_{n\in\mathbb{N}}$ définie pour tout $n \in \mathbb{N}$ par

$$
A_n = G(f_n(X_n), V_n), \quad X_{n+1} = F(n, X_n, A_n, U_n)
$$

est une chaîne de Markov contrôlée de matrices de transition contrôlées

$$
(P(n, i, a, j) = \mathbb{P}(F(n, i, a, U_1) = j))_{n\in\mathbb{N}, i,j\in\mathcal{S}, a\in\mathcal{A}}
$$

pour la filtration $(\mathcal{F}_n = \sigma(X_0, V_0, U_0, \cdots V_n, U_n))_{n\in\mathbb{N}}$ et telle que $(A_n)_{n\in\mathbb{N}}$ s'éxrite sous forme mixte markovienne.

Solution 6. On utilise l'exercice précédent, ce qui permet de conclure.

Remarque 2.12. *Nous avons montré dans les deux exercices précédent, et en utilisant les idées de la Proposition 1.12 que se donner une collection de matrices de transition contrôlées ainsi qu'une suite de fonction qui à un élément de $\mathcal{S}$ associe une probabilité sur l'espace des contrôles permet de définir une chaîne de Markov contrôlée avec une suite de contrôles markoviens mixtes. C'est un analogue des Propositions 1.10 et 1.12 pour les chaînes de Markov contrôlées, et cela aura une grande influence dans la simulation numérique de telles chaînes.*

2.5. Récompenses. Etant donnée une chaîne de Markov controlée $(X_n, A_n)_{n\in\mathbb{N}}$ comme précédemment, on associe $(R_n)_n$ une suite de variables $\mathcal{F}_n$-mesurable, représentant la récompense à l'instant n. Typiquement il existe une suite de fonction $(r_n)_{n\in\mathbb{N}}$ telles que pour tout $n \in \mathbb{N}$,

$$
R_n = r_n(X_n, A_n).
$$

L'objectif dans la suite (chapitre 2) sera de maximiser les récompenses moyennes accumulées.

3. Exemples

3.1. Approvisionnement de stock. Un centre de livrasion s'approvisionne de la façon suivate :

- S_0 : stock initial
- A l'instant $n \in \mathbb{N}$, S_n représente l'état du stock (un entier dans $\{0, \cdots, N_S\}$) et A_n représente l'approvisionnement (un entier das $\{0, \cdots, N_A\}$).
- Entre n et $n+1$, D_{n+1} représente la demande aléatoire (indépendante des observations passées), de loi homogène (sur $\{0, \cdots, N_D\}$).
- La demande est satisfiate si $S_n + A_n \geq D_{n+1}$, sinon elle est satisfaite dans la mesure du possible.

Exercice 7. Ecrire la chaîne de Markov contrôlée associé. *Indication :* $S_{n+1} = \max(S_n + A_n - D_{n+1}, 0)$.

SOLUTION 7. Pour $k \in \{0, \cdots, N_D\}$ on note $p_k = \mathbb{P}(D_1 = k)$. Remarquons que pour $n \in \mathbb{N}$, $p_k = \mathbb{P}(D_{n+1} = k)$.

Nous allons désormais construire une matrice de transition contrôlée homogène telle que

$$\begin{aligned}\mathbb{P}(S_{n+1} = j, X_n = i, A_n = a) \\ = \mathbb{P}((X_n + A_n - D_{n+1})_+ = j, X_n = i, A_n = a) = P(i,a,j)\mathbb{P}(X_n = i, A_n = a).\end{aligned}$$

Soit $i \in \{1, \cdots, N_S\}$ et $a \in \{0, \cdots, N_A\}$. Le gérant du magasin ne fera jamais une commande d'approvisionnement qui va dépasser sa capacité de stockage. Ainsi, on définit

$$P(i,a,j) = 0 \quad i + a > N_S, \quad j \in \{0, \cdots, N_S\}.$$

Supponsons maintenant que $0 \leq i + a \leq N_S$. On suppose que $j \in \{1, \cdots, N_S\}$. Si $j > i + a$, quelque soit la demande, il n'est pas possible qu'à la fin de la période de vente, le stock soit plus élevé qu'au début. Ainsi

$$P(i,a,j) = 0 \quad i + a \leq N_S, \quad j > i + a.$$

Supposons désormais que $0 \leq j \leq i + a \leq N_S$. Si $j < i + a - N_D$, il n'est pas possible que les stocks du magasin aient autant baissé, et dans ce cas

$$P(i,a,j) = 0 \quad i + a \leq N_S, \quad j < i + a - N_D.$$

On suppose désormais que $i + a - N_D \leq j \leq i + a$ et que $i + a - N_D \geq 0$. Alors, il n'existe qu'une possilité pour que le stock passe de $i + a$ à j, c'est que la demande soit exactement égale à $i + a - j$. On pose alors

$$P(i,a,j) = p_{i+a-j}.$$

De la même façon si $i + a - N_D < 0$ et $j \in \{1, \cdots, i + a\}$,

$$P(i,a,j) = p_{i+a-j}.$$

Enfin, si $i + a - N_D < 0$, pour que le stock passe à zéro, il faut que la demande ai été plus grande que $i + a$. On pose

$$P(i,a,0) = \sum_{k=i+a}^{N_D} p_k.$$

On remarque que l'on a construit P telle que pour tout $i, j \in \{0, \cdots, N_S\}$, tout $a \in \{0, \cdots, N_A\}$ et tout $n \in \mathbb{N}$,

$$\mathbb{P}(S_{n+1} = j, X_n = i, A_n = a) = P(i,a,j)\mathbb{P}(X_n = i, A_n = a).$$

3.2. Bandit multi-bras. Un joueur observe les résultats de K machines à sous différentes. A chaque étape, il en choisit une et une seule et la joue, les $K - 1$ autres ne sont pas jouées.

A chaque partie n, on consigne les états des machines à sous, ainsi, un seul de ces états parmis les $K - 1$ est actualisé.

On suppose que la machine $i \in \{1, \cdots, K\}$ peut être dans les états S^i. La machine i passe d'un état $s_1 \in S^i$ à un état $s_2 \in S^i$ avec probabilité $p(i, s_1, s_2)$.

EXERCICE 8. Ecrire une matrice de transition contrôlée avec $S^1 \times \cdots \times S^K$ comme espace d'états et $\{1, \cdots, K\}$ comme espace d'actions qui modélise le problème précédent.

SOLUTION 8. La chaîne de Markov contrôlée va décrire la dynamique suivante : à un instant $n \in \mathbb{N}$ les machines sont à l'état $s = (s^1, \cdots, s^K) \in \mathcal{S} := S^1 \times \cdots \times S^K$. Le joueur observe les machine, choisi la machine $i \in \mathcal{A} = \{1, \cdots, K\}$. L'état de cette dernière est actualisé aléatoirement. Soit $i \in \{1, \cdots, K\}$, $s_1^i, s_2^i \in S^i$ et $s^1, \cdots, s^{i-1}, s^{i+1}, \cdots, s^K \in S^1 \times \cdots \times S^{i-1} \times S^{i+1} \times \cdots \times S^K$, et soit $s_1 = (s^1, \cdots, s^{i-1}, s_1^i, s^{i+1}, \cdots, s^K)$ et $s_2 = (s^1, \cdots, s^{i-1}, s_2^i, s^{i+1}, \cdots, s^K)$. On définit

$$P(s_1, i, s_2) = p(i, s_1^i, s_2^i).$$

Si il existe deux indices $j \neq k \in \{1, \cdots, K\}$ telles que $s, \tilde{s} \in \mathcal{S}$ et $s^k \neq \tilde{s}^k$ et $s^j \neq \tilde{s}^j$, on pose

$$P(s, i, \tilde{s}) = 0$$

pour tout $i \in \{1, \cdots, K\}$.

On remarque que P est une matrice de transition contrôlée homogène.

3.3. Selection d'un film sur Netflix. Un abonné selectionne un film sur Netflix avec la règle suivante : il visualise les suggestions de façon séquentielle et ne peut choisir que le film dont il est en train de regarder la notice (pas de retour en arrière possible).

On note $\mathcal{S} = \{0, 1\}$ l'espace d'état. L'état est 1 si le film consulté est le meilleurs choix (au regard des goûts de l'abonné) parmis tout ceux consultés jusqu'alors. Létat est 0 si ce n'est pas le cas.

On note $\mathcal{A} = \{0, 1\}$ l'espace d'actions : 1 si on choisit le film dont on est en train de consulter la notice, 0 sinon.

Exercice 9. Montrer qu'on obtien une chaîne de Markov contrôlée de matrices de transition

$$P(n, i, a, 1) = \frac{1}{n+1} = 1 - P(n, i, a, 0), \quad i \in \{0, 1\}, a \in \{0, 1\}.$$

Solution 9.