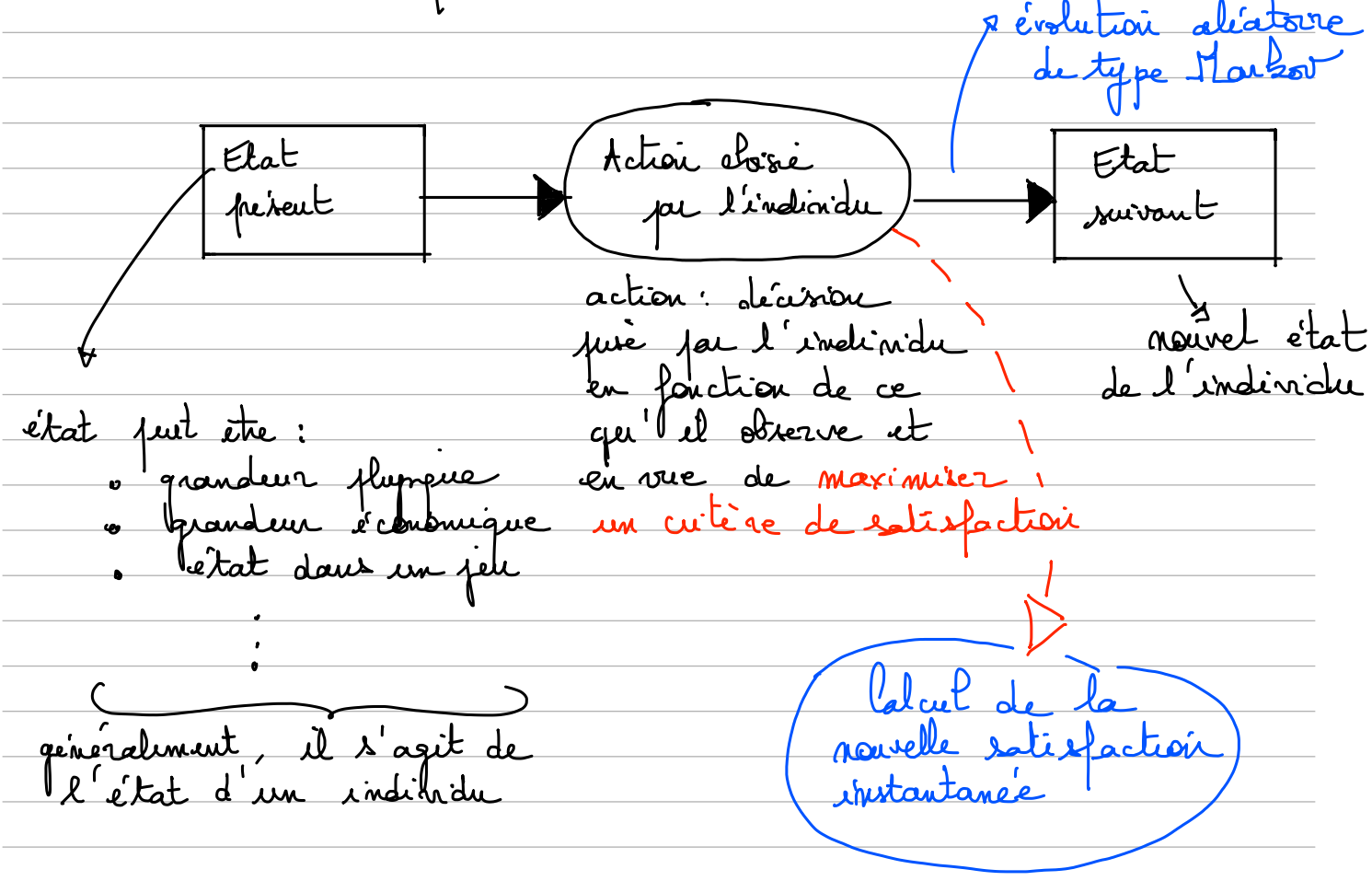
Cours MDP et applications.

Chapitre 1. Chaînes de Markov contrôlées

Philosophie générale.

1. Modélisation

Chaînes de Markov contrôlées.



Très nombreuses applications : finance, économie mais aussi biologie (ex : épidémiologie)...

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

* forme de transitions aléatoires

(les réalisations ne sont pas connues, mais la

distribution statistique et connue - en fonction des états occupés et des actions suivies -)

* forme des récompenses (ou satisfactions) en fonction des états occupés et des actions suivies.

Naturellement, les DECISIONS ne sont pas connues à l'avance.

OBJECTIF : CARACTERISER mathématiquement les meilleures décisions.

2. Apprentissage

En pratique : on ne connaît pas les quantités mathématiques régissant le modèle. Les seules choses observées sont les états occupés, les actions décidées et les récompenses reçues.

QUESTION : COMMENT APPRENDRE LA MEILLEURE DECISION A "JOUER" EN FONCTION DES OBSERVATIONS ?

→ problématique typique de l'apprentissage statistique. On parle d'APPRENTISSAGE par RENFORCEMENT (une des branches de l'apprentissage avec l'apprentissage supervisé et l'apprentissage non supervisé)

intuitivement, la "pertinence" de la décision prise doit évoluer de façon croissante avec le nombre d'observations

Idée générale : pour répondre à 2, il faut bien comprendre 1.

I Quelques rappels sur les chaînes de Markov

⚠ Ici pas de contrôle. Le cadre correspond à

celui étudié en cours de processus stochastiques.

1) Principe général.

Une chaîne de Markov décrit l'évolution (au cours du temps) d'un système aléatoire, sans mémoire. Il s'agit de la version probabiliste d'une suite récurrente de la forme

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

- x_{n+1} : état de l'individu à l'instant $n+1$
- x_n : état de l'individu à l'instant n.
- f : fonction "DÉTERMINISTE" expliquant le passage de l'état courant à l'état suivant.

Remarques : (a) La notion d'état est ici la même que la notion d'état dans le paragraphe introductif

(b) Ici, il n'y a pas possibilité pour l'individu de prendre une décision

(c) Aucun ALÉA à ce stade. Le passage de l'état x_n à l'état x_{n+1} est déterministe

Définition : On appelle espace d'états l'ensemble S des valeurs pour les états de l'individu, à n'importe quel instant. Autrement dit, on suppose, dans $(*)$,

a) $x_0 \in S$

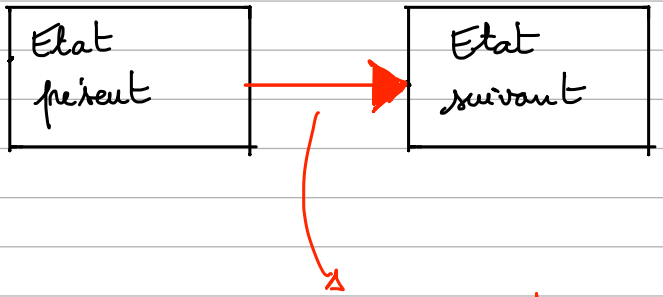
b) $\forall n \in \mathbb{N}$, $f(n, \cdot) : S \to S$, i.e

$$f : \mathbb{N} \times S \longrightarrow S$$

Dans la suite, on supposera S au + dénombrable et souvent fini.

⚠ A ce stade, pas d'aléa

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



la transition est maintenant aléatoire

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

(**) $\left(f(n,i)\right)_{i\in S} \rightsquigarrow \left(P(n,i,j)\right)_{i,j\in S}$

Définition: Une collection de réels $\left(q(i,j)\right)_{i,j\in S}$ forme une matrice de transition (sur S) si

$\forall i \in S$, $q(i,\cdot)$ forme une collection de poids de probabilité sur S, i.e

a) $\forall j \in S$, $q(i,j) \geq 0$

b) $\sum_{j\in S} q(i,j) = 1$.

Application: On suppose donc que, dans (**),

$\forall n \in \mathbb{N}$, $P(n,\cdot,\cdot)$ est une matrice de transition.

2) définitions d'une chaîne de Markov $(\Omega, \mathcal{F}, \mathbb{P})$ espace de probabilité

Définition: Etant une suite $(P(n,\cdot,\cdot))_{n\in\mathbb{N}}$ de matrices de transition sur S, une collection (X_n) de variable aléatoires à valeurs dans S (formant ainsi un processus indexée par le temps discret) constitue une chaîne de Markov de matrices de transition $(P(n,\cdot,\cdot))_{n\in\mathbb{N}}$ si l'une des définitions suivantes est vérifiée:

a) $\forall n \in \mathbb{N}$, $\forall i_1, \dots, i_n, j \in S$,

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

b) $\forall n \in \mathbb{N}$, $\forall i_1, \dots, i_n, j \in S$ tels que

$$\mathbb{P}\{X_0 = i_0, \dots, X_n = i_n\} > 0$$

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

On peut généraliser cette définition en utilisant le concept de filtration

Définition: Etant donnée $\mathcal{G}$ une sous-algèbre de $\mathcal{F}$, et $X : \Omega \to E$ une variable aléatoire à valeurs dans E dénombrable, on appelle la conditionnelle de X sachant $\mathcal{G}$ la collection

de variables aléatoires.

$$\underbrace{\left(\mathbb{P}(\{X=i\}\mid\mathcal{G})\right)}_{\mathbb{E}\left(\mathbb{1}_{\{X=i\}}\mid\mathcal{G}\right)}{}_{i\in E}$$

Si $\mathcal{G}=\sigma(Y)$ où $Y:\Omega\longrightarrow F$ dénombrable alors sur $\{Y=j\}$, $j\in F$ tel que $\mathbb{P}\{Y=j\}>0$, on a (TD)

$$\mathbb{E}\left(\mathbb{1}_{\{X=i\}}\mid\mathcal{G}\right)=\mathbb{P}\{X=i\mid Y=j\}$$

Autrement dit

$$\mathbb{E}\left(\mathbb{1}_{\{X=i\}}\mid\mathcal{G}\right)=\sum_{j\in F}\mathbb{P}\{X=i\mid Y=j\}\,\mathbb{1}_{\{Y=j\}}.$$

$\mathbb{E}\left(\mathbb{1}_{\{X=i\}}\mid\mathcal{G}\right)$ = $\sigma(Y)$

$\mathbb{P}\{X=i\mid Y=j\}$ la définition est arbitraire si $\mathbb{P}\{Y=j\}=0$.

Définition: On appelle filtration sur $(\Omega,\mathcal{S},\mathbb{P})$ toute collection $(\mathcal{S}_n)_{n\in\mathbb{N}}$ croissante de sous-tribus de $\mathcal{S}$.
(Autrement dit, $\mathcal{S}_n\subset\mathcal{S}_{n+1}$, $n\in\mathbb{N}$).

Application:

Définition: Etant une suite $(P(n,\cdot,\cdot))_{n\in\mathbb{N}}$ de matrices de transition sur S et une filtration $(\mathcal{S}_n)_{n\in\mathbb{N}}$ sur $(\Omega,\mathcal{S},\mathbb{P})$ une collection $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires à valeurs dans S constitue une $(\mathcal{S}_n)_{n\in\mathbb{N}}$-chaîne de Markov de matrices de transition $(P(n,\cdot,\cdot))_{n\in\mathbb{N}}$ si

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

$$\mathbb{P}\left(\{X_{n+1} = i\} \mid \mathcal{F}_n\right) = P(n, X_n, i) \quad (\mathbb{P}\text{ p.s.})$$

Remarque : toute chaîne de Markov au sens de la première définition forme une $(\mathcal{F}_n)_{n\in\mathbb{N}}$-chaîne de Markov pour

$$\mathcal{F}_n := \sigma(X_0, \dots, X_n).$$

en effet

sur $\{X_n = j\}$ (avec $\mathbb{P}\{X_n = j\} > 0$)

$$\mathbb{P}\left(\{X_{n+1} = i\} \mid \mathcal{F}_n\right) \overset{*}{=} \mathbb{P}\left(\{X_{n+1} = i\} \mid \{X_n = j\}\right)$$

$$\overset{*}{=} \text{déf loi conditionnelle} = \mathbb{P}(n, j, i) = P(n, X_n, j).$$

$\overset{*}{=}$ déf chaîne de Markov

4) Construction canonique.

On donne ici la construction canonique d'une chaîne de Markov généralisant la définition d'une chaîne déterministe sous mémoire donnée dans §1.

Proposition : Soit $f : \mathbb{N} \times S \times [0,1] \longrightarrow S$ une fonction $\mathcal{B}$-mesurable.

Soient $X_0 : \Omega \longrightarrow S$ et $(U_n)_{n\in\mathbb{N}^*}$ une v.a et une suite de v.a telles que

i) $U_n \sim \mathrm{Unif}[0,1]$ pour chaque $n \in \mathbb{N}$

ii) $(U_n)_{n\in\mathbb{N}\setminus\{0\}}$ suite de v.a. $\perp\!\!\!\perp$

iii) $X_0 \perp\!\!\!\perp (U_n)_{n\in\mathbb{N}\setminus\{0\}}$.

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

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

est une Chaîne de Markov de matrices de transition

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

Remarque : (a) ii) et iii) sont $\not\Leftrightarrow$ à, pour tout $n\in\mathbb{N}$, $X_0, U_1, \dots, U_n$ sont $\perp\!\!\!\perp$.

(b) démonstration faite en TD

En réalité, il existe une réciproque (preuve également faite en TD)

Proposition : Toute chaîne de Markov (de matrices de transition $(P(n,i,j))_{n,i,j}$ données) admet une forme canonique

Indice : - $f(n,i,\cdot)$ fonction quantile de la loi de $(P(n,i,j))_{j\in S}$.

II Chaînes contrôlées

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

1) Cas déterministe

Commençons par examiner ce que serait la version déterministe

Dans l'écriture

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

on dit AJOUTER la décision prise par l'individu à l'instant n.

Définition : On appelle espace d'actions un ensemble A contenant toutes les actions (ou tous les choix) possibles pour l'individu à n'importe quel instant. Dans un souci de simplification, on suppose de fait que A est indépendant du temps (mais cela pourrait être généralisé au cas où les actions possibles varient avec le temps).

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

Nous sommes donc amenés à considérer

$$f : \mathbb{N} \times S \times A \longrightarrow S$$

mesurable (i.e. $\forall (n,i) \in \mathbb{N} \times S$, $f(n,i,\cdot) : A \to S$ est mesurable au sens classique

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

(démonstration à faire en TD)

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

$$\forall n \in \mathbb{N} \qquad x_{n+1} = f(n, x_n, a_n).$$

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 représentée par la chaîne.

2) matrices de transition contrôlées.

On veut maintenant généraliser la définition donnée dans le §1 précédent au cas stochastique en nous appuyant sur le concept de matrices de transition vu dans la section I.

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

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

telle que pour tout $(n,a) \in \mathbb{N} \times A$ la collection $\big(P(n,i,a,j)\big)_{i,j}$ forme une matrice de transition

c'est à dire pour tout $(n,ya) \in \mathbb{N} \times S \times A$ $P(n,i,a,\cdot)$ est une famille de poids de probabilité sur S.

Remarque : On dit que la collection est homogène
- si

$$P(n,i,a,j) = P(i,a,j)$$

pour $P : S \times A \times S \to [0,1]$ mesurable de sorte que $P(\cdot,a,\cdot)$ forme une matrice de transition pour tout $a \in A$.

Interprétation : Si l'individu est dans l'état i à l'instant n, alors en choisissant l'action $a \in A$, il a probabilité $P(n, i, a, j)$ de se retrouver dans l'état j à l'instant suivant $n+1$.

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

0 → le système est sain non infecté

1 → le système est attaqué

A chaque instant $n \in \mathbb{N}$

En état 0 : il y a deux choix

- NE RIEN FAIRE. La protection du système est dégradée, mais cela ne coûte rien au manager du système.
- METTRE A JOUR la protection. La protection du système 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 termes de réparation, mais cela a un coût sur les activités autour.
- REPARER le système. Cela a un coût pour le service informatique mais préserve les activités périphériques.

On a donc $A = \{0, 1\}$ (0 : On agit ; On ne fait rien)

et on pose (action)

$$P(0,0,1) = p \quad \text{où } p \in [0,1].$$

(0 : état ; 1 : état)

$$P(0,0,0) = 1-p$$

$$P(0,1,1) = q$$

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

Clairement $p > q$.

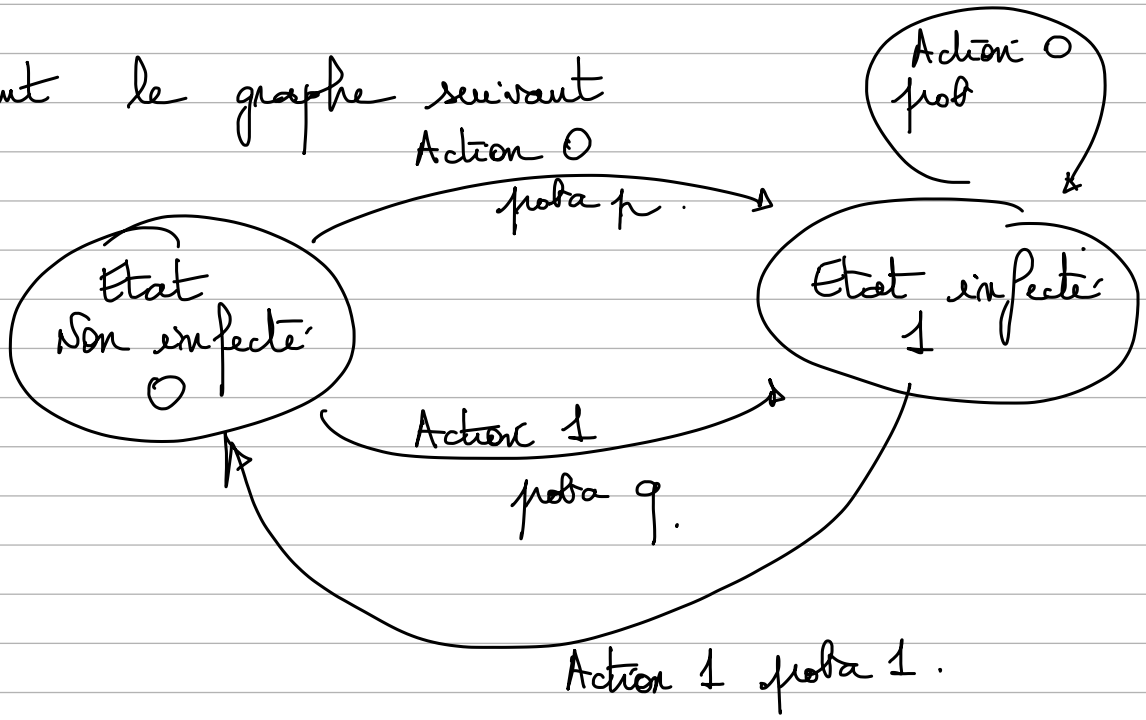
$$P(1,0,1) = 1$$

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

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

$$P(1,1,1) = 0.$$

On obtient le graphe suivant



Remarque : le modèle est simple. On oublie la protection passée dans la probabilité d'être affecté par une attaque.

On pourrait ajouter un état : "protégé mais pas mis à jour".

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 informations disponible à l'individu (ou à l'observateur) pour prendre la décision à un instant donné ? 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 : Etant donnée une filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ et une famille de matrices de transition contrôlées

$$P : \mathbb{N}\times S\times A \ni (n,i,a) \longmapsto P(n,i,a) \in [0,1]$$

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

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

$$\mathbb{P}\left(\{X_{n+1}=i\} \mid \mathcal{F}_n\right) = P(n, X_n, A_n, i)$$

Illustrons la dernière propriété dans le cas où A est au + dénombrable.

Proposition : Si A est au + dénombrable.

$\forall n\in\mathbb{N}$ $\forall (i_0, a_0, i_1, a_1, \dots, i_n, a_n) \in \underbrace{S\times A\times S\times A\times \dots}_{n \text{ fois}}$

$\forall\, j \in S$

$$\mathbb{P}\{X_0 = i_0, A_0 = a_0, \dots, X_n = i_n, A_n = i_n, X_{n+1} = j\}$$
$$= \Gamma(n, i_n, a_n, j)\ \mathbb{P}\{X_0 = i_0, A_0 = a_0, \dots, X_n = i_n, A_n = a_n\}.$$

preuve : On a

$$\{X_0 = i_0, A_0 = i_0, \dots, X_n = i_n, A_n = a_n\} \in \mathcal{F}_n$$

$$\mathbb{E}\left\{ \mathbb{1}_{\{X_{n+1} = j\}}\ \mathbb{1}_{\{X_0 = i_0, A_0 = a_0, \dots, X_n = i_n, A_n = a_n\}} \right\}$$

$$= \mathbb{E}\left\{ \mathbb{E}\left(\mathbb{1}_{\{X_{n+1} = j\}} \mid \mathcal{F}_n \right)\ \mathbb{1}_{\{X_0 = i_0, A_0 = a_0, \dots, \}} \right\}.$$

$$\Gamma(n, X_n, A_n, j).$$

$$\Gamma(n, i_n, a_n, j)$$

$$= \Gamma(n, i_n, a_n, j)\ \mathbb{P}\{X_0 = i_0, A_0 = a_0, \dots, \}.$$

Dans le cas où A est non-dénombrable, on ne peut plus procéder comme ci-dessus : les événements peuvent être de probabilité nulle (en particulier si $A_0, \dots, A_n$ sont à densité).

Proposition : Dans le cas général, si $B_0, \dots, B_n$ sont n boréliens de A, et $(i_0, \dots, i_n)$ sont dans S, alors

$$\mathbb{P}\left(\{X_0 = i_0, A_0 \in B_0, X_1 = i_1, A_1 \in B_1, \dots, X_n = i_n, A_n \in B_n, X_{n+1} = j\}\right)$$
$$= \int_{S^{n+1} \times (\mathbb{R}^d)^{n+1}} \Gamma(n, i_n, a_n)\ \mathbb{1}_{B_0 \times \dots \times B_n}(a_0, \dots, a_n)\ \mathbb{1}_{\{(i_0, \dots, i_n)\}}(j_0, \dots, j_n)\ d\mathbb{P}_{(X_0, A_0, \dots)}(j_0, a_0, \dots, j_n, a_n)$$

preuve. Comme précédemment

$$\{X_0 = i_0, A_0 \in B_0, \dots, X_n = i_n, A_n \in B_n\} \in \mathcal{F}_n$$

$$\mathbb{P}\left(\{X_0 = i_0, A_0 \in B_0, X_1 = i_1, A_1 \in B_1, \dots, X_n = i_n, A_n \in B_n, X_{n+1} = j\}\right)$$
$$= \mathbb{E}\{ \underbrace{I(n, X_n, A_n, j)}_{I(n, i_n, A_n, j)} \mathbb{1}_{\{X_0 = i_0, A_0 \in B_0, \dots, X_n = i_n, A_n \in B_n\}} \}$$

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

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

Alors

$$\mathbb{P}\{X_0 = i_0, A_0 = a_0, \dots, X_n = i_n, A_n = a_n, X_{n+1} = j\}$$
$$= \int_{S^{n+1}} I(n, i_n, a_n) \, \mathbb{P}_{(A_0, \dots, A_n) \mid X_0 = i_0, \dots, X_n = i_n}(da_0, \dots, da_n).$$

4) Mesurabilité des actions.

On distingue 3 types de propriété de mesurabilité pour $(A_n)_{n \in \mathbb{N}}$.

Définition 1. On dit que les actions $(A_n)_{n \in \mathbb{N}}$ s'écrivent sous forme fermée dépendant du passé si

$$\forall n \in \mathbb{N}, \quad A_n \text{ est } \sigma(X_0, \dots, X_n) \text{ mesurable}$$

c'est à dire $A_n = f_n(X_0, \dots, X_n)$

On choisit l'action à l'instant n en fonction des observations des états passés de 0 à n.

Définition 2 : On dit que les actions $(A_n)_{n\in\mathbb{N}}$ s'écrivent sous forme fermée markovienne si

$$\forall n, \quad A_n \text{ est } \sigma(X_n)\text{-mesurable}$$

c'est à dire $A_n = f_n(X_n)$

On choisit l'action à l'instant n en fonction de la seule observation de l'état à l'instant n.

Définition 3 : 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 v.a. $\perp\!\!\!\perp$ de loi uniforme sur $[0,1]$ telle que

- $\forall n \in \mathbb{N}$, V_n est $\mathcal{F}_n$-mesurable
- $\forall n \in \mathbb{N}$, $V_n \perp\!\!\!\perp \sigma(X_n) \vee \mathcal{F}_{n-1}$.
- $\forall n \in \mathbb{N}$, A_n est $\sigma(X_n, V_n)$ mesurable.

interprétation : à l'instant n, on ajoute, en plus de l'observation de l'état X_n à l'instant n un tirage $\perp\!\!\!\perp$ pour décider de l'action à suivre.

Exercice (TD) : On suppose A fini.

A chaque instant n, on se donne

$$f_n : S \longrightarrow \underbrace{\mathcal{P}(A)}_{\substack{\text{espace de probabilité sur } A \\ (= \text{l'ensemble de probabilité sur } A)}}$$

Montrer que f_n induit une suite d'actions

markoviennes mixtes.

Indice : On appelle $(V_n)_{n\in\mathbb{N}}$ suite de v.a. $\perp\!\!\!\perp$ de loi uniforme sur $[0,1]$ et $X_0 \perp\!\!\!\perp (V_n)_{n\in\mathbb{N}}$.

Pour $\mu \in \mathcal{P}(A)$, on appelle $Q(\mu,\cdot)$ la fonction de $[0,1]$ dans A la fonction suivante :

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

$u \in [0,1]$ où $A = \{a_1,\dots,a_{|A|}\}$

Alors on construit, pour X_0 donné,

$$A_0 = Q(f_0(X_0), V_0)$$
$$X_1 \sim P(0, X_0, A_0, \cdot)$$
$$A_1 = Q(f_1(X_1), V_1)$$
$$X_2 \sim P(1, X_1, A_1, \cdot)$$
$$\vdots$$

et $\mathcal{F}_n = \sigma(X_0, A_0, \dots, X_n, A_n)$.

Exercice (TD) On se donne

$$\left(f_n : S \longrightarrow \mathcal{P}(A)\right)_{n\in\mathbb{N}}$$

$$F : \mathbb{N}\times S\times A\times[0,1] \longrightarrow S$$

$(X_0, (V_n)_{n\in\mathbb{N}}, (U_n)_{n\in\mathbb{N}}, \log)$ collection de v.a. $\perp\!\!\!\perp$ avec $V_n \sim \mathrm{Unif}[0,1]$ et $U_n \sim \mathrm{Unif}[0,1]$. On construit

$$A_0 = Q(f_0(X_0), V_0)$$
$$X_1 = F(0, S_0, A_0, U_1)$$

$$A_1 = g(f_1(X_1), V_1)$$
$$X_2 = F(1, S_1, A_1, U_1)$$
$$\vdots$$

montrer que l'on obtient une chaîne de Markov controlée avec $\mathcal{F}_n = \sigma(X_0, V_0, U_1, \dots, U_n, V_n)$.

5) récompenses.

Etant donnée une chaîne de Markov controlée $(X_n, A_n)_{n\in\mathbb{N}}$ comme précédemment, on associe

R_n : $\mathcal{F}_n$-mesurable représentant la récompense à l'instant n

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

Objectif : Maximiser les récompenses moyennes accumulées. Ce sera le chapitre 2.

III Exemples (TD)

1) Approvisionnement de stock.

Un centre de livraison s'approvisionne de la façon suivante

- S_0 : stock initial
- A l'instant $n \in \mathbb{N}$

  S_n : état du stock (entier dans $\{0, \dots, N\}$)

  A_n : approvisionnement (entier dans $\{0, \dots, N\}$)

- Entre n et $n+1$

D_{n+1} : demande aléatoire (indépendante des observations passées) de loi homogène (sur $\{0, \dots, N_D\}$).

la demande est satisfaite si $S_n + A_n \geq D_{n+1}$, sinon elle est satisfaite dans la mesure du possible.

Ecrire la chaîne de Markov contrôlée associée

Indice : $S_{n+1} = \max(S_n + A_n - D_{n+1}, 0)$.

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 machines ne sont pas jouées.

A chaque partie n, on consigne les états des K machines. Un seul état parmi les K est de fait actualisé.

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

$$P(i, s_1, s_2).$$

Ecrire une chaîne de Markov contrôlée avec

$S^1 \times \dots \times S^K$ comme espace d'états

$\{1, \dots, K\}$ comme espace d'actions.

3) Selection d'un film sur Netflix

Un abonné Netflix sélectionne un film avec la règle suivante : il visualise les propositions de façon séquentielle et ne peut choisir que le film dont il est en train de consulter la notice (pas de retour en arrière).

$S = \{0, 1\}$

1 si le film consulté est le meilleur choix (au regard des goûts de l'abonné) parmi tous ceux consultés jusqu'ici
0 sinon

$A = \{0, 1\}$

1 si on choisit le film dont on consulte la notice, 0 sinon

On obtient donc une chaîne de Markov contrôlée. Montrer que les transitions sont

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

$\cdot\ i \in \{0, i\}$
$a \in \{0, i\}$