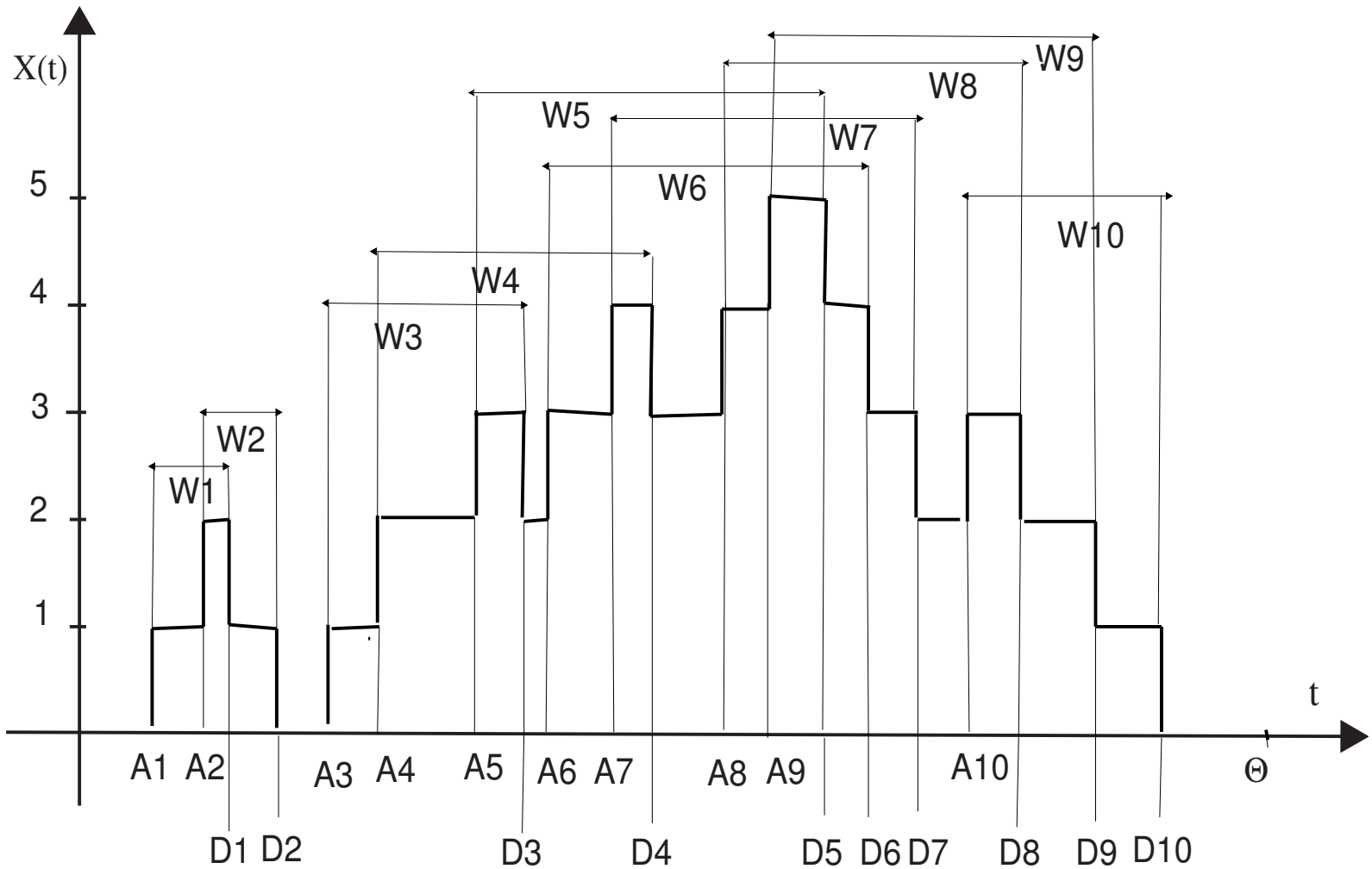
ISM-AG 2
UE MI0B903T
Processus stochastiques (partie 2)

Claudie Hassenforder-Chabriac
Agnès Lagnoux

Simulation graphique



Paramètres opérationnels en régime transitoire

- Θ : temps total de l'observation ;
- $T(n,\Theta)$: temps total durant lequel il y a n clients ;
- W_k : temps de séjour du $k^{\text{ième}}$ client dans le système : $W_k = D_k - A_k$;
- $P(n,\Theta) = \frac{T(n,\Theta)}{\Theta}$: proportion de temps pendant laquelle le système contient n clients ;
- $\alpha(\Theta)$: nombre de clients arrivant dans le système pendant la période $[0,\Theta]$;
- $\delta(\Theta)$: nombre de clients quittant le système pendant la période $[0,\Theta]$.

Paramètres de performance en régime transitoire

- Débit moyen d'entrée : $d_e(\Theta) = \frac{\alpha(\Theta)}{\Theta}$;
- Débit moyen de sortie : $d_s(\Theta) = \frac{\delta(\Theta)}{\Theta}$;
- Nombre moyen de clients :
  $L(\Theta) = \frac{1}{\Theta} \sum_{n=0}^{+\infty} nT(n,\Theta) = \sum_{n=0}^{+\infty} nP(n,\Theta)$;
- Temps moyen de séjour : $W(\Theta) = \frac{1}{\alpha(\Theta)} \sum_{k=1}^{\alpha(\Theta)} W_k$
- Taux d'utilisation : $U(\Theta) = \sum_{n=1}^{+\infty} P(n,\Theta) = 1 - P(0,\Theta)$

Paramètres de performance en régime permanent

En régime permanent, on s'intéressera à l'existence et aux valeurs des limites lorsque Θ tend vers l'infini de tous les paramètres du régime transitoire :

- $d_e = \lim_{\Theta \to +\infty} d_e(\Theta)$;
- $d_s = \lim_{\Theta \to +\infty} d_s(\Theta)$;
- $L = \lim_{\Theta \to +\infty} L(\Theta)$;
- $W = \lim_{\Theta \to +\infty} W(\Theta)$;
- $U = \lim_{\Theta \to +\infty} U(\Theta)$.

Stabilité, Ergodicité

Définition

Un système est stable si et seulement si

$$\lim_{\Theta \to +\infty} d_s(\Theta) = \lim_{\Theta \to +\infty} d_e(\Theta) = d.$$

Définition

Un système est ergodique si et seulement si, quelle que soit la réalisation particulière étudiée du processus :

$$\lim_{\Theta \to +\infty} \sum_{n=0}^{+\infty} n^k P(n,\Theta) = \lim_{t \to +\infty} \sum_{n=0}^{+\infty} n^k \pi_n(t) \text{ pour tout } k = 1,2,\cdots$$

Ds 1 syst. ergodique et en régime permanent, proportions de tps passé ds 1 état = proba. d'être ds cet état.

Liens entre distribution stationnaire et paramètres de performance

Lorsque le système est stable et ergodique, on a donc

$$\boxed{\lim_{\Theta \to +\infty} P(n,\Theta) = \lim_{t \to +\infty} \pi_n(t) = \pi_n}.$$

On a alors :

- $L = \sum_{n=1}^{+\infty} n\pi_n$;
- $L_q = \sum_{n=C+1}^{+\infty} (n-C)\pi_n$;
- $d = d_e = d_s = \sum_{n=0}^{+\infty} \lambda_n \pi_n = \sum_{n=1}^{+\infty} \mu_n \pi_n$.

Formule de Little

Théorème

Le nombre moyen de clients, le temps moyen passé dans le système et le débit moyen d'un système stable en régime permanent se relient de la façon suivante :

$$\boxed{L = W \times d}$$

En effet, $L(\Theta) = \sum_{n=0}^{+\infty} nP(n,\Theta) = \frac{1}{\Theta} \sum_{n=0}^{+\infty} nT(n,\Theta)$ et $W(\Theta) = \frac{1}{\delta(\Theta)} \sum_{k=1}^{\delta(\Theta)} W_k$. Or $\sum_{n=0}^{+\infty} nT(n,\Theta) = \sum_{k=1}^{\delta(\Theta)} W_k$ car les deux sommations de cette égalité représentent deux façons de calculer l'aire sous la courbe de X_t.

Liens entre paramètres de performance

La loi de Little peut s'appliquer de différentes façons :

- au système entier : $L = W \times d$;
- à la file d'attente seule : $L_q = W_q \times d$;
- au serveur seul : $L_S = W_S \times d = \frac{1}{\mu} \times d$ en notant $\frac{1}{\mu}$ le temps de service moyen.

Ces trois relations ne sont pas indépendantes. On peut en effet déduire l'une d'entre elles à partir des deux autres en remarquant que :

$$W = W_q + \frac{1}{\mu} \text{ et } L = L_q + L_S.$$

Dans le cas où $C = 1$ (un seul serveur), $L_S = U$.

Étude de la file M/G/1

Le processus d'arrivée des clients dans la file est toujours supposé poissonien de taux λ mais, maintenant, le temps de service Y d'un client est distribué selon une loi qui n'est plus supposée exponentielle.
Pour étudier simplement ce système, le service n'étant plus exponentiel, il ne suffit plus de savoir qu'un client est en service, pour prédire quand ce service va se terminer. Il faut en plus savoir depuis combien de temps le service a commencé. On va voir ici 2 méthodes différentes :

- méthode de la chaîne de Markov induite ;
- méthode de l'analyse de la valeur moyenne.

Méthode de la chaîne de Markov induite : généralités

On considère le processus (X_t) juste après les instants $D_1, \cdots, D_k, \cdots$ où les clients terminent leur service, et on pose $X_k = X_{D_k^+}$. On définit ainsi une chaîne de Markov $(X_k)_{k\geq 1}$ dite chaîne de Markov induite.

Cette chaîne peut être facilement étudiée et notamment, on peut déterminer $\pi_n(k) = \mathbb{P}(X_k = n)$, puis la distribution stationnaire de (X_k), solution de $\pi = \pi P$ où P est la matrice de transition de la chaîne. On a alors

$$\boxed{\pi_n = \lim_{k\to+\infty} \mathbb{P}(X_k = n) = \lim_{t\to+\infty} \mathbb{P}(X_t = n)}.$$

Mise en oeuvre de la méthode

On introduit une variable aléatoire N qui est égale au nombre de clients qui entrent pendant un service.
La loi de N est déteminée par les a_n, où

$$\begin{aligned} a_n &= \mathbb{P}(N=n) = \int \mathbb{P}(N=n|Y=t) f_Y(t)\, dt \\ &= \int_0^{+\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} f_Y(t)\, dt \end{aligned}$$

Détermination de la matrice de transition

$P = (q_{i,j})_{i,j\in\mathbb{N}}$ où $q_{i,j}$ est la probabilité de transition de l'état i vers l'état j. On a

$$\begin{cases} q_{0,j} = a_j & \text{si } j \geq 0 \\ q_{i,j} = a_{j-i+1} & \text{si } 1 \leq i \leq j+1 \\ q_{i,j} = 0 & \text{sinon.} \end{cases}.$$

En effet,

- si $X_k = 0$, $X_{k+1} = j$ correspond à l'arrivée de j clients pendant le service du $(k+1)^{\text{ième}}$ client ;
- si $X_k = i \geq 1$, on a $X_{k+1} = j$ si $X_{k+1} - X_k = j - i$, ce qui correspond à l'arrivée de $j - i + 1$ clients, car il ne faut pas oublier que le $(k+1)^{\text{ième}}$ client vient de partir.

Détermination de la distribution stationnaire

La distribution stationnaire doit vérifier $= P$ soit
$\pi_j = \sum_{i=0}^{+\infty} \pi_i q_{i,j} = a_j \pi_0 + \sum_{i=1}^{j+1} a_{j-i+1}$ pour tout j.
Pour la déterminer, il est plus simple de déterminer sa fonction génératrice en fonction de celle de N.

Théorème

Si $\Pi(z) = \sum_{n=0}^{+\infty} \pi_n z^n$, *si* $A(z) = \sum_{n=0}^{+\infty} a_n z^n$ *et si* $\rho = \frac{\lambda}{\mu}$, *alors :*

$$\boxed{\Pi(z) = \frac{(1-\rho)A(z)(z-1)}{z-A(z)}}.$$

Détermination des paramètres de performance

Si Y est la variable aléatoire représentant la durée du service et si $\rho = \frac{\lambda}{\mu}$ où $\mathbb{E}(Y) = \frac{1}{\mu}$, alors

$$\boxed{L = \rho + \frac{\rho^2 + \lambda^2 \mathrm{var}(Y)}{2(1-\rho)}}.$$

En effet, $L = \Pi'(1) = \lim_{h \to 0} \frac{\Pi(1+h) - \Pi(1)}{h}$. Le développement limité à l'ordre 1 de $\Pi(z)$ en 1 utilise celui de $A(z)$ à l'ordre 2 et,

$$\begin{aligned} A(z) &= \sum_{j=0}^{+\infty} a_j z^j = \sum_{j=0}^{+\infty} z^j \int_0^{+\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!} \, f_Y(t) \, dt \\ &= \int_0^{+\infty} e^{-\lambda t} e^{\lambda z t} \, f_Y(t) \, dt \end{aligned}$$

$$A'(1) = \lambda \int_0^{+\infty} t f_Y(t) dt = \rho \; ; \; A''(1) = \lambda^2 \int_0^{+\infty} t^2 f_Y(t) dt = \lambda^2 \mathbb{E}(Y^2)$$

Analyse de la valeur moyenne

Si on n'a pas besoin de connaître la loi , mais seulement les paramètres de performance, la méthode qui suit est plus simple. Elle utilise le temps moyen résiduel de service t_r = temps qu'il reste au serveur au moment où un client arrive, pour terminer son service.

Si $m = \frac{1}{\mu} = \mathbb{E}(Y)$, on a $\boxed{\begin{cases} L_q = \sum_{n=1}^{+\infty} (n-1)\pi_n \\ W_q = \sum_{n=1}^{+\infty} \pi_n [t_r + (n-1)m] \end{cases}}$.

En appliquant la formule de Little à la file $L_q = \lambda W_q$ et au service $1 - \pi_0 = \lambda \mathbb{E}(Y) = \rho$, on obtient

$$W_q = (1-\pi_0)t_r + mL_q = \rho t_r + m\lambda W_q = \rho t_r + \rho W_q$$

soit $\boxed{W_q = \frac{\rho}{1-\rho} t_r}$.

Détermination du temps moyen résiduel de service

Proposition

Dans une file $M/G/1$, le temps moyen résiduel de service t_r, vu par un client qui arrive dans la file est

$$\boxed{t_r = \frac{m}{2}\left(1 + \frac{\mathrm{var}(Y)}{m^2}\right)}.$$

On note Z la variable aléatoire mesurant la durée d'un service vue par l'arrivée d'un client (différente de la variable aléatoire Y mesurant la durée d'un service). Un client qui arrive dans la file a plus de chance de "tomber" sur un service long que sur un service court. La probabilité de tomber sur un service de longueur t est proportionnelle à t, ainsi qu'à la fréquence avec laquelle un service de longueur t a lieu.
On a donc $f_Z(t) = Ktf_Y(t)$ avec $K = \frac{1}{m}$ en intégrant des deux côtés, puis $t_r = \frac{1}{2}\int t f_Z(t)\,dt = \frac{1}{2m}\int t^2 f_Y(t)\,dt$.

Généralités sur les réseaux

Un réseau de files d'attente est un ensemble de M stations interconnectées. On peut classer les réseaux de files d'attente en deux catégories :

- les réseaux de files d'attente monoclasses, dans lesquels circulent une seule classe de clients,
- les réseaux de files d'attente multiclasses, dans lesquels circulent plusieurs classes de clients.

On fait un autre type de distinction :

- réseaux ouverts : les clients arrivent de l'extérieur, circulent dans le réseau à travers les stations, puis quittent le réseau.
- réseaux fermés : les clients sont en nombre constant. Il n'y a pas d'arrivée ni de départ de clients.

Notations générales

- M : nombre de stations dans le réseau (chaque station a sa propre file d'attente) ;
- n_i : nombre de clients dans la station i (service + attente) ;
- $n = \sum_{i=1}^{M} n_i$: nombre total de clients dans le réseau ;
- $\lambda(n)$: débit instantané, fonction de n, de clients arrivant de l'extérieur.

On néglige le temps de transport entre les stations ; les zones d'attente sont supposées de capacité infinie.

Détermination des routages

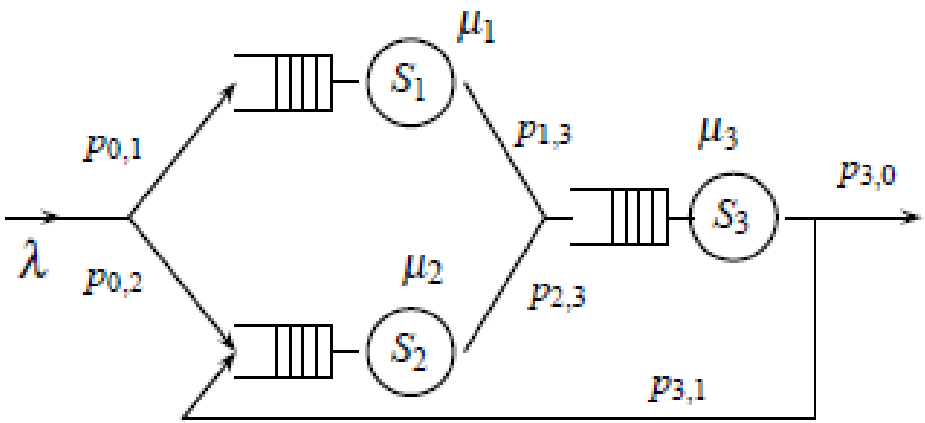
On suppose les routages probabilistes : quand un client a plusieurs destinations possibles à la fin d'un service, il fait son choix suivant une certaine distribution de probabilité et on note :

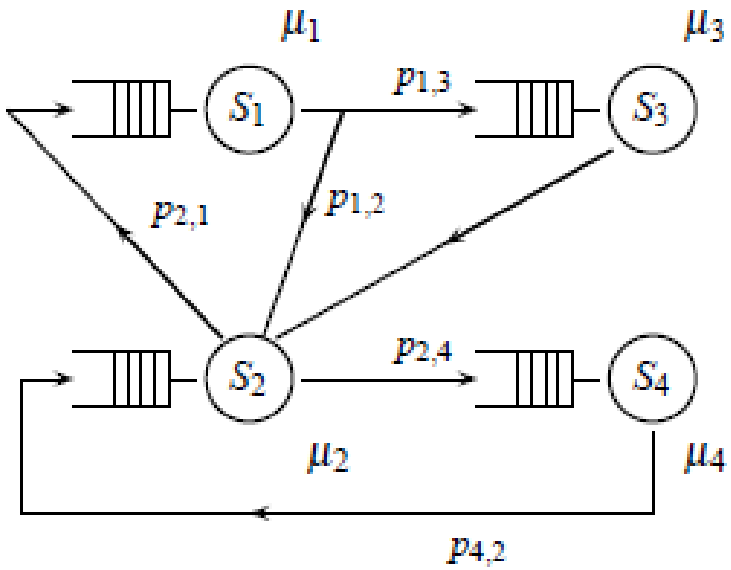
- $p_{0,i}$ la probabilité qu'un client qui arrive dans le système se rende à la station i ;
- $p_{i,j}$ la probabilité qu'un client qui termine son service à la station i se rende à la station j ;
- $p_{i,0}$ la probabilité qu'un client qui termine son service à la station i quitte le système.

On a, avec la convention $p_{0,0} = 0$:

$$\sum_{j=0}^{M} p_{i,j} = 1 \text{ pour } i = 0, \cdots, M.$$

Exemples de réseaux





Équation des flux, taux de visite

On note $\lambda_j(n)$ le taux d'arrivée à la station j : ce taux se décompose en $p_{0,j}\lambda(n)$ venant de l'extérieur et $p_{i,j}\lambda_i(n)$ venant de la station i, pour $i = 1, \cdots, M$.
On a donc les égalités :

$$\boxed{\lambda_j(n) = p_{0,j}\lambda(n) + \sum_{i=1}^{M} p_{i,j}\lambda_i(n), \qquad j = 1, \cdots, M}.$$

$\lambda_j(n)$ dépend de n et on peut poser, pour tout $j \in [\![1; M]\!]$,
$\boxed{\lambda_j(n) = e_j\lambda(n)}$ où les e_j, taux de visite dans chaque station j, sont les solutions du système :

$$\boxed{e_j = p_{0,j} + \sum_{i=1}^{M} p_{i,j}e_i, \qquad j = 1, \cdots, M.}$$

Cas Markovien

Les inter-arrivées de clients venant de l'extérieur sont exponentielles, de taux $\lambda(n)$, et les lois de service exponentielles, de taux respectifs $\mu_i(n_i)$.

Le processus aléatoire $\vec{N}(t) = (N_1(t), \cdots, N_M(t))$, où $N_i(t)$ est le nombre de clients dans la station i à l'instant t, est un processus de Markov, dont l'espace des états est $\mathbb{N}^M$.

Les transitions simultanées ont des probabilités infinitésimales négligeables.

Transitions à partir d'un état $\overrightarrow{n}$

Transitions à partir d'un état $\overrightarrow{n} = (n_1, \cdots, n_M)$:

- $0 \to i$: état final $\overrightarrow{n} + \overrightarrow{l_i}$; taux $\lambda(n) p_{0,i}$
- $i \to 0$: état final $\overrightarrow{n} - \overrightarrow{l_i}$ si $n_i \geq 1$; taux $\mu_i(n_i) p_{i,0}$
- $i \to j$: état final $\overrightarrow{n} - \overrightarrow{l_i} + \overrightarrow{l_j}$ si $n_i \geq 1$; taux $\mu_i(n_i) p_{i,j}$.

On en déduit le taux de sortie de l'état $\overrightarrow{n}$:

$$\begin{aligned} q(\overrightarrow{n}) &= \sum_{i=1}^{M} \lambda(n) p_{0,i} + \sum_{i=1}^{M} \mu_i(n_i) p_{i,0} + \sum_{i=1}^{M} \sum_{j \neq i} \mu_i(n_i) p_{i,j} \\ &= \lambda(n) + \sum_{i=1}^{M} \mu_i(n_i)(1 - p_{i,i}) \end{aligned}$$

Transitions vers un état $\overrightarrow{n}$

Transitions vers un état $\overrightarrow{n} = (n_1, \cdots, n_M)$:

- $0 \to i$: état initial $\overrightarrow{n} - \overrightarrow{l_i}$; $q(\overrightarrow{n} - \overrightarrow{l_i}, \overrightarrow{n}) = \lambda(n-1)p_{0,i}$
- $i \to 0$: état initial $\overrightarrow{n} + \overrightarrow{l_i}$; $q(\overrightarrow{n} + \overrightarrow{l_i}, \overrightarrow{n}) = \mu_i(n_i+1)p_{i,0}$
- $j \to i$: état initial $\overrightarrow{n} + \overrightarrow{l_j} - \overrightarrow{l_i}$; $q(\overrightarrow{n} + \overrightarrow{l_j} - \overrightarrow{l_i}, \overrightarrow{n}) = \mu_j(n_j+1)p_{j,i}$.

Équations d'équilibre

Soit $\pi(\vec{n})$ la probabilité asymptotique de l'état $\vec{n}$.
Les équations de transition à l'équilibre traduisant que le taux de sortie de l'état $\vec{n}$ est égal aux taux de transition vers l'état $\vec{n}$, s'écrivent :

$$\boxed{\pi(\vec{n})q(\vec{n}) = \sum_{\vec{n'} \neq \vec{n}} \pi(\vec{n'})q(\vec{n'}, \vec{n}) \text{ pour tout } \vec{n}}$$

soit,

$$\begin{aligned}\pi(\vec{n})\left(\sum_{i=1}^{M} \mu_i(n_i)(1-p_{i,i}) + \lambda(n)\right) &= \lambda(n-1)\sum_{i=1}^{M} p_{0,i}\pi(\vec{n} - \vec{l_i}) \\ &+ \sum_{i=1}^{M} \mu_i(n_i+1)p_{i,0}\pi(\vec{n} + \vec{l_i}) \\ &+ \sum_{i=1}^{M}\sum_{j\neq i} \mu_j(n_j+1)p_{j,i}\pi(\vec{n} - \vec{l_i} + \vec{l_j})\end{aligned}$$

Réseaux à forme produit

Définition

Un réseau de files d'attente sera dit à forme produit si et seulement si ses probabilités d'états $\pi(\vec{n})$ se mettent sous la forme

$$\pi(\vec{n}) = \frac{1}{G} \prod_{i=1}^{M} \varphi(n_i)$$

où G est une constante de normalisation assurant que $\sum_{\vec{n}} \pi(\vec{n}) = 1$.

Les réseaux monoclasses ouverts à taux constants

Ces réseaux, appelés aussi réseaux de Jackson ouverts remplissent les conditions suivantes :

- une seule classe de clients ;
- un seul serveur à chaque station ;
- une capacité de stockage illimitée à toutes les stations ;
- une discipline de service FIFO pour toutes les files ;
- processus d'arrivée des clients dans le système poissonien de taux λ, indépendant de n ;
- temps de service exponentiel à chaque station, de taux μ_i pour la station i, indépendant de n.

Stabilité et Taux de visite

En pratique :

- On calcule les taux de visite e_i à l'aide du système :

$$\boxed{e_i = p_{0,i} + \sum_{j=1}^{M} e_j p_{j,i} \text{ pour } i = 1, \cdots, M}$$

  à M équations à M inconnues, et on en déduit les taux d'arrivée aux différentes stations : $\boxed{\lambda_i = \lambda e_i}$ pour $i = 1, \cdots, M$;

- On vérifie la condition de stabilité du réseau : $\boxed{\lambda_i < \mu_i}$, $i = 1, \cdots, M$, sinon l'étude ne peut être poursuivie.

Analyse du régime permanent

Un réseau de Jackson, décrit par le processus $(\vec{N}(t))_{t\geq 0}$ où $\vec{N}(t) = (N_1(t), \cdots, N_M(t))$ avec $N_i(t)$ nombre de clients présents dans la station i au temps t, possède une solution extrêmement simple, donnée par la propriété suivante, connue sous le nom de théorème de Jackson :

Théorème

La probabilité stationnaire du réseau possède la "forme produit" suivante :

$$\pi(\vec{n}) = \prod_{i=1}^{M} \pi_i(n_i) = \prod_{i=1}^{M} (1-\rho_i)\rho_i^{n_i} \text{ où } \rho_i = \frac{\lambda_i}{\mu_i} = \frac{\lambda e_i}{\mu_i}$$

(π_i est la probabilité stationnaire d'une file $M/M/1$ ayant un taux d'arrivée λ_i et un taux de service μ_i).

Calcul des paramètres de performances

Les paramètres de performance peuvent être calculés par file ou pour l'ensemble du réseau :

	Station i	Réseau
Débit moyen	d_i	d
Nombre moyen de clients	L_i	L
Temps moyen de séjour	W_i	W

- Les paramètres de performance de chaque station se déduisent de la décomposition en files $M/M/1$:

$$d_i = \lambda_i = \lambda e_i \; ; \; L_i = \frac{\rho_i}{1-\rho_i} \text{ où } \rho_i = \frac{\lambda_i}{\mu_i} \; ; \; W_i = \frac{L_i}{d_i} = \frac{1}{\mu_i - \lambda_i}$$

- Les paramètres de performance du réseau s'en déduisent :

$$d = \lambda \; ; \; L = \sum_{i=1}^{M} L_i \text{ et } W = \frac{L}{d} = \frac{L}{\lambda} = \sum_{i=1}^{M} e_i W_i \; .$$

Extension au cas de stations multiserveurs

Ici, la station i dispose de C_i serveurs identiques, dont le service est exponentiel et de taux μ_i. La condition de stabilité du réseau est

$\boxed{\lambda_i < C_i \mu_i \text{ pour } i = 1, \cdots, M}$.

Théorème

La probabilité stationnaire du réseau possède la "forme produit"

$$\boxed{\pi(\vec{n}) = \prod_{i=1}^{M} \pi_i(n_i)}$$

où π_i est la probabilité stationnaire d'une file $M/M/C$ ayant un taux d'arrivée $\lambda_i = e_i \lambda$, un taux de service μ_i et comportant C_i serveurs.

L'analyse du réseau se réduit donc, comme dans le cas monoserveur, à l'analyse de M files simples qui sont ici des $M/M/C$.

Les réseaux monoclasses fermés à taux constants

Dans un réseau fermé, les clients circulent sans jamais en sortir et sans qu'aucun client de l'extérieur n'y rentre. Cela revient à supposer $p_{i,0} = 0$ et $p_{0,i} = 0$ pour $i = 1, \cdots, M$.

Dans un réseau fermé, il n'y a aucun problème de stabilité : pour toute station i, $N_i(t) \leq N$ pour tout t.

Comme dans le cas ouvert, on s'intéresse ici aux réseaux de files d'attente fermés comportant :

- une seule classe de clients
- un seul serveur à chaque station
- un temps de service exponentiel à chaque station μ_i ;
- des files FIFO.

Ces réseaux sont appelés réseaux de Jackson fermés.

Problème des taux de visite

Dans un réseau fermé, le nombre absolu de fois qu'un client passe par une station peut être infini. On s'intéresse donc au nombre relatif e_i de passages à une station i entre deux passages par une station de référence.
De même que dans le cas ouvert, les e_i sont solutions du système d'équations :

$$e_i = \sum_{j=1}^{M} e_j p_{j,i} \text{ pour } i = 1, \cdots, M.$$

Ce système admettant une infinité de solutions, on convient de poser $e_1 = 1$: les autres taux de visite se déduisent du système sans ambiguïté et s'interprètent comme le nombre moyen de passages par les différentes stations du réseau entre 2 passages par la station 1.

Analyse du régime permanent

Le réseau est encore décrit avec le processus $(\vec{N}(t))_{t\geq 0}$ où $\vec{N}(t) = (N_1(t), \cdots, N_M(t))$ mais ici $\boxed{\sum_{i=1}^{M} N_i(t) = N}$.

L'ensemble $E(M,N)$ de tous les états possibles du système est $\boxed{E(M,N) = \{\vec{n} = (n_1, \cdots, n_M) \ ; \ \sum_{i=1}^{M} n_i = N\}}$ et $\boxed{\text{card}(E(M,N)) = \binom{N+M-1}{M-1}}$ (nombre de façons de répartir les N clients dans les M stations du réseau).

Les équations d'états s'écrivent ici :

$$\boxed{\pi(\vec{n}) \sum_{i;n_i>0} \sum_{j=1}^{M} \mu_i p_{i,j} = \sum_{i;n_i>0} \sum_{j=1}^{M} \pi(\vec{n} - \vec{l_i} + \vec{l_j}) \mu_j p_{j,i} \text{ pour } \vec{n} \in E(M,N)}.$$

Distribution stationnaire

Ce système d'équations possède, comme dans le cas ouvert, une solution très simple.

Proposition

La probabilité stationnaire du réseau possède la "forme produit" suivante :

$$\pi(\vec{n}) = \frac{1}{G_{M,N}} \prod_{i=1}^{M} f_i(n_i)$$

où $f_i(n_i) = \left(\frac{e_i}{\mu_i}\right)^{n_i}$ *et* $G_{M,N}$ *est la constante de normalisation.*

Algorithme de Buzen pour le calcul de la constante de normalisation

La constante de normalisation d'un réseau contenant M stations dans lequel circulent N clients est

$$G_{M,N} = \sum_{\vec{n} \in E(M,N)} \prod_{i=1}^{M} \left(\frac{e_i}{\mu_i}\right)^{n_i}.$$

Pour la calculer, on résoud la récurrence suivante :

$$G(m,n) = G(m-1,n) + \rho_m G(m,n-1)$$

où $\rho_i = \frac{e_i}{\mu_i}$, m concerne la station, n est le nombre de clients. On calcule ainsi toutes les constantes $G(m,n)$ pour $m = 1, \cdots, M$ et $n = 0, \cdots, N$ en partant des conditions initiales $G(1,n) = \rho_1^n$ pour $n = 0, \cdots, N$ et $G(m,0) = 1$ pour $m = 1, \cdots, M$. Enfin, $G_{M,N} = G(M,N)$.

Paramètres de performances en fonction des constantes de normalisation

Pour $i = 1, \cdots, M$ et $k = 0, \cdots, N$,

$$\boxed{\pi_i(k) = \sum_{\vec{n} \in E(M,N); n_i = k} \pi(\vec{n}) = \left(\frac{e_i}{\mu_i}\right)^k \frac{G_i(M-1,N-k)}{G(M,N)}}.$$

Les paramètres de performances de chaque station s'en déduisent alors immédiatement :

- $U_i = 1 - \pi_i(0) = \frac{e_i}{\mu_i} \frac{G(M,N-1)}{G(M,N)} = \frac{d_i}{\mu_i}$;
- $d_i = \sum_{k=1}^{N} \pi_i(k)\mu_i = (1 - \pi_i(0))\mu_i = e_i \frac{G(M,N-1)}{G(M,N)}$;
- $L_i = \sum_{k=1}^{N} k\pi_i(k) = \frac{1}{G(M,N)} \sum_{k=1}^{N} k \left(\frac{e_i}{\mu_i}\right)^k G_i(M-1, N-k)$;
- $W_i = \frac{L_i}{d_i} = \frac{1}{e_i G(M,N-1)} \sum_{k=1}^{N} k \left(\frac{e_i}{\mu_i}\right)^k G_i(M-1, N-k)$

Algorithme de convolution

Initialisation :

$G(1,n) = f_1(n)$ pour $n = 0, \cdots, N$

$G(m,0) = 1$ pour $m = 1, \cdots, M$

$G_i(M-1,0) = 1$ pour $i = 1, \cdots, M$

Pour m variant de 2 à M, faire

Pour n variant de 1 à N, faire

$$G(m,n) = G(m-1,n) + \sum_{k=1}^{n} f_m(k) G(m-1,n-k)$$

Pour i variant de 1 à M, faire

Pour n variant de 1 à N, faire

$$G_i(M-1,n) = G(M,n) - \sum_{k=1}^{n} f_i(k) G_i(M-1,n-k)$$

Calculer les paramètres de performances moyens à l'aide des relations précédentes.

Algorithme MVA

Si seuls les paramètres de performances moyens sont requis, il existe un algorithme récursif simple et performant, qui permet d'éviter le calcul des constantes de normalisation.
L'algorithme repose sur la relation récursive, exprimant le temps moyen de séjour d'un client à la station i, dans le réseau contenant n clients, en fonction du nombre moyen de clients de la station i, dans le réseau contenant $n-1$ clients :

Proposition

On a les relations suivantes

$$W_i(n) = \frac{1}{\mu_i}(1 + L_i(n-1)) \; ; \; d(n) = \frac{n}{\sum_{i=1}^{M} e_i W_i(n)}$$

puis $d_i(n) = e_i d(n)$ *et* $L_i(n) = d_i(n) W_i(n)$, *avec* $L_i(0) = 0$

Algorithme MVA

Initialisation : $\pi_i(0,0) = 1$ pour $i = 1, \cdots, M$

Pour n variant de 1 à N, faire

$$W_i(n) = \sum_{k=1}^{n} \frac{k}{\mu_i(k)} \pi_i(k-1, n-1) \text{ pour } i = 1, \cdots, M$$

$$d(n) = \frac{n}{\sum_{i=1}^{M} e_i W_i(n)}$$

$$d_i(n) = e_i d(n) \text{ pour } i = 1, \cdots, M$$

$$L_i(n) = W_i(n) d_i(n) \text{ pour } i = 1, \cdots, M$$

$$\pi_i(k,n) = \frac{d_i(n)}{\mu_i(k)} \pi_i(k-1, n-1) \text{ pour } i = 1, \cdots, M \text{ et } k = 1, \cdots, n$$

$$\pi_i(0,n) = 1 - \sum_{k=1}^{n} \pi_i(k,n) \text{ pour } i = 1, \cdots, M$$

Extension au cas de stations multiserveurs

Comme dans le cas ouvert, on peut étendre le théorème de Jackson fermé au cas de stations multiserveurs, chaque station i comporte C_i serveurs identiques. Toutes les autres hypothèses sont conservées.

Proposition

La probabilité stationnaire du réseau possède la "forme produit" suivante :

$$\pi(\vec{n}) = \frac{1}{G(M,N)} \prod_{i=1}^{M} f_i(n_i)$$

où $f_i(n_i) = \dfrac{1}{\prod_{k=1}^{n_i} \min(k, C_i)} \left(\dfrac{e_i}{\mu_i}\right)^{n_i} = \begin{cases} \dfrac{1}{n_i!}\left(\dfrac{e_i}{\mu_i}\right)^{n_i} & \text{si } n_i < C_i \\ \dfrac{1}{C_i! C_i^{n_i - C_i}}\left(\dfrac{e_i}{\mu_i}\right)^{n_i} & \text{si } n_i \geq C_i \end{cases}$ *et*

$G(M,N)$ *est la constante de normalisation.*

ISM-AG 2
UE MI0B903T
Processus stochastiques (partie 2)

Claudie Hassenforder-Chabriac
Agnès Lagnoux

Introduction à l'apprentissage par renforcement

Introduction à l'apprentissage par renforcement

Sources :

- Livre collectif en français. Processus décisionnels de Markov et Intelligence Artificielle, Hermès, 2008. Editeurs O. Sigaud et O. Buffet.
- Cours de Rémi Munos "Introduction à l'apprentissage par renforcement"

  http ://researchers.lille.inria.fr/∼munos/master-mva
- Cours de David Silver "Introduction to reinforcement learning"

  https ://www.youtube.com/watch ?v=2pWv7GOvuf0
- Cours de Jacques A. Ferland "Modèles stochastiques Processus de décisions markoviens"
- Mémoire de DEA d'Adriana TAPUS "Utilisation de processus de décision markoviens pour la planification et l'exécution d'actions par un robot mobile". 2002.

Objectifs de l'A/R

- Acquisition automatisée de compétences pour la prise de décisions (actions ou contrôle) en milieu complexe et incertain.
- Apprendre par l'expérience une stratégie comportementale (appelée politique) en fonction des échecs ou succès constatés (les renforcements ou récompenses).
- Exemples : jeu du chaud-froid, apprentissage sensori-moteur, jeux (backgammon, échecs, poker, go), robotique mobile autonome, gestion de portefeuille, recherche opérationnelle, ...

Naissance du domaine

Rencontre fin années 1970 entre

- Neurosciences computationnelles. Renforcement des poids synaptiques des transmissions neuronales (règle de Hebbs, modèles de Rescorla et Wagner dans les années 60, 70). Renforcement = corrélations activités neuronales.
- Psychologie expérimentale. Modèles de conditionnement animal : renforcement de comportement menant à une satisfaction (recherches initiées vers 1900 par Pavlov, Skinner et le courant béhavioriste). Renforcement = satisfaction, plaisir ou inconfort, douleur. Cadre mathématique adéquat : Programmation dynamique de Bellman (années 50, 60), en théorie du contrôle optimal. Renforcement = critère à maximiser.

Liens avec la psychologie expérimentale

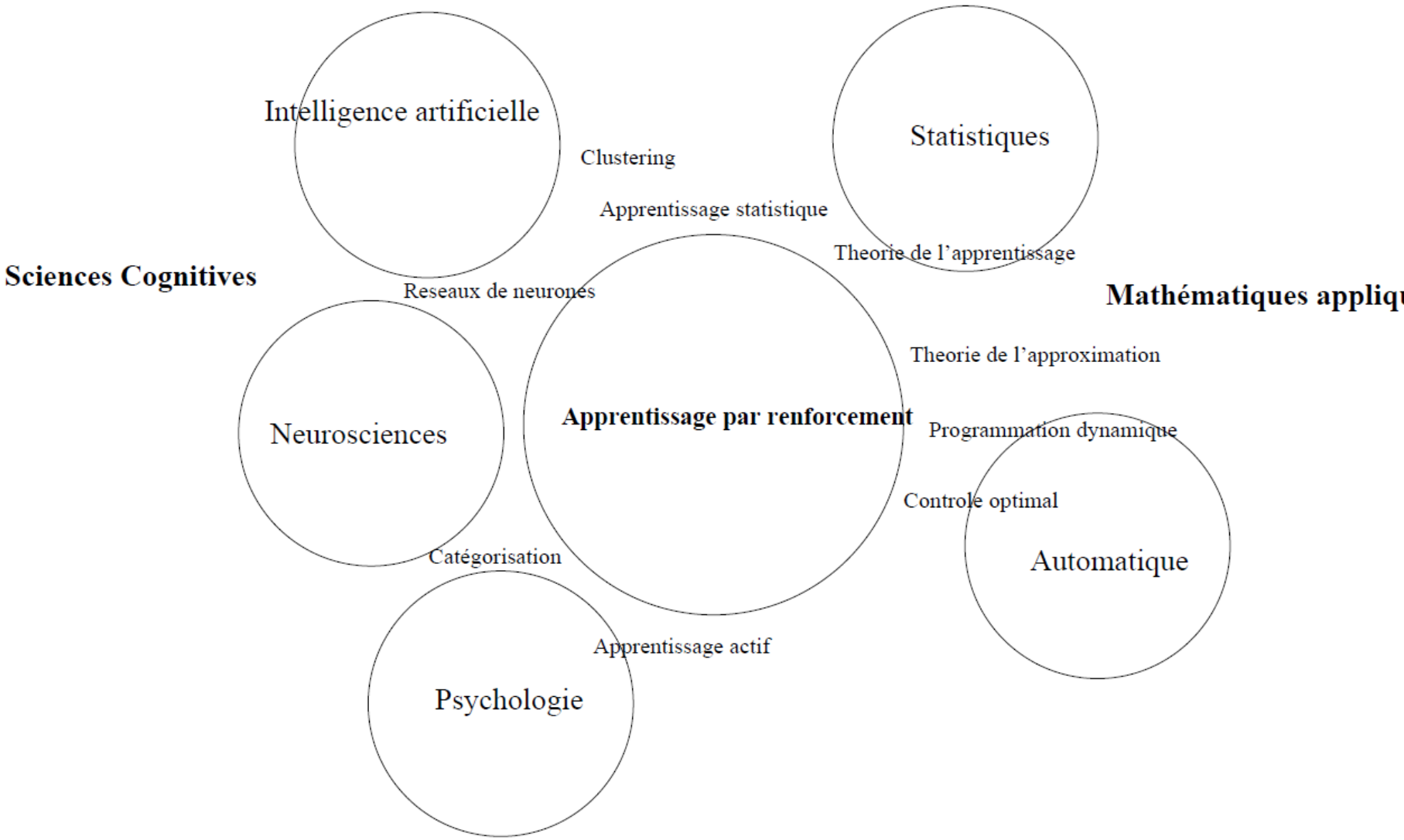
Thorndike (1911) - Loi des effets

> "Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur ; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond."

Préhistoire de l'A/R computationnel

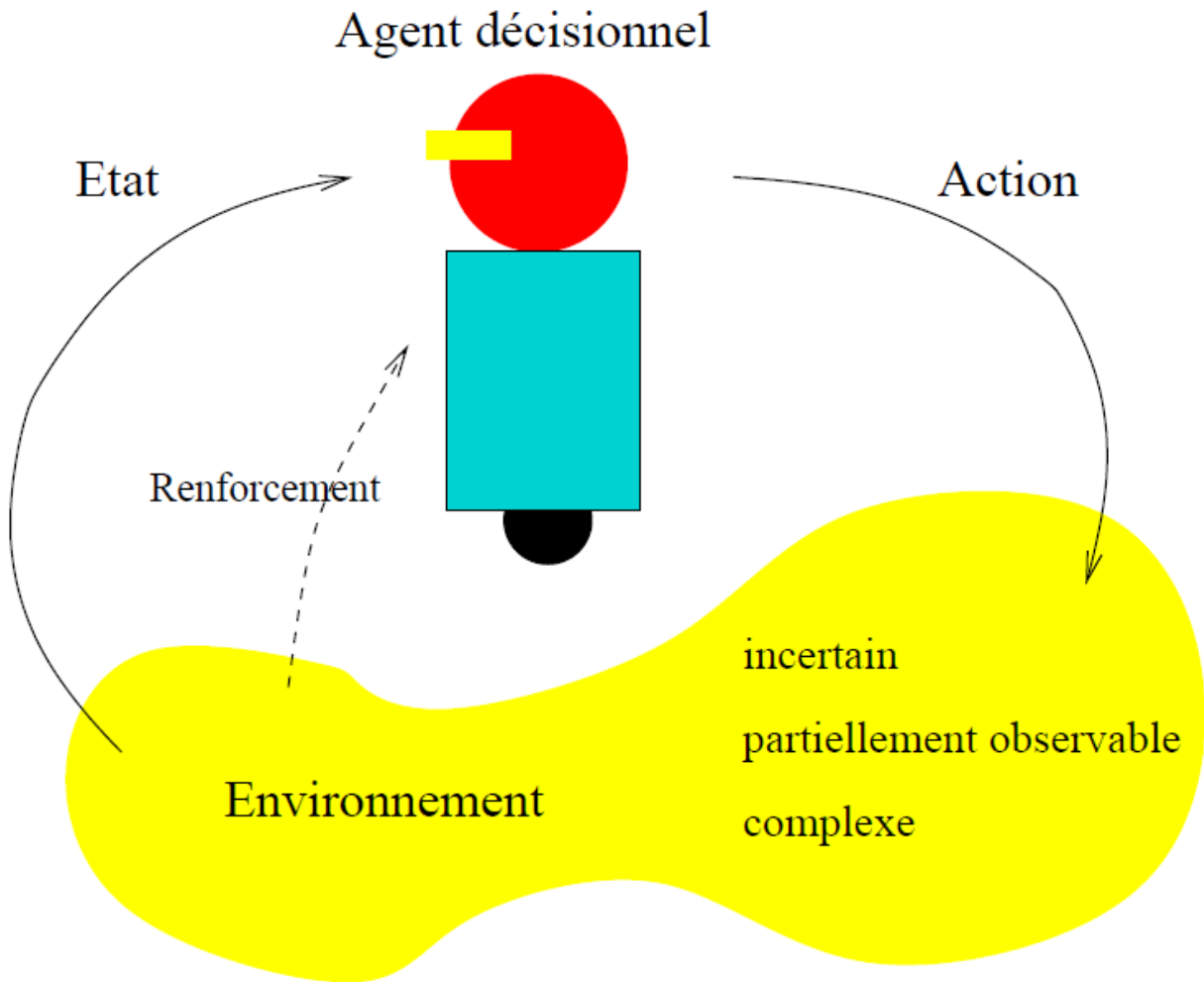
- Shannon 1950 : Programming a computer for playing chess.
- Minsky 1954 : Theory of Neural-Analog Reinforcement Systems.
- Samuel 1959 : Studies in machine learning using the game of checkers.
- Michie 1961 : Trial and error. -> joueur de tic-tac-toe.
- Michie et Chambers 1968 : Adaptive control -> pendule inversé.
- Widrow, Gupta, Maitra 1973 : Punish/reward : learning with a critic in adaptive threshold systems -> règles neuronales.
- Sutton 1978 : Théories d'apprentissage animal : règles dirigées par des modifications dans prédictions temporelles successives.
- Barto, Sutton, Anderson 1983 : règles neuronales Actor-Critic pour le pendule inversé.
- Sutton 1984 : Temporal Credit Assignment in Reinforcement Learning.
- Klopf 1988 : A neuronal model of classical conditioning.
- Watkins 1989 : Q-learning.
- Tesauro 1992 : TD-Gammon

Domaine pluridisciplinaire



Differents types d'apprentissage

- Apprentissage supervisé : à partir de l'observation de données $(X_i, Y_i)_i$ où $Y_i = f(X_i) + \varepsilon_i$ et f est la fonction cible (inconnue), estimer f afin de faire des prédictions de $f(x)$;
- Apprentissage non-supervisé : à partir de données $(X_i)_i$, trouver des structures dans ces données (ex. des classes), estimer des densités, ...
- Apprentissage par renforcement


Agent décisionnel
Etat
Action
Renforcement
Environnement
incertain
partiellement observable
complexe

L'environnement

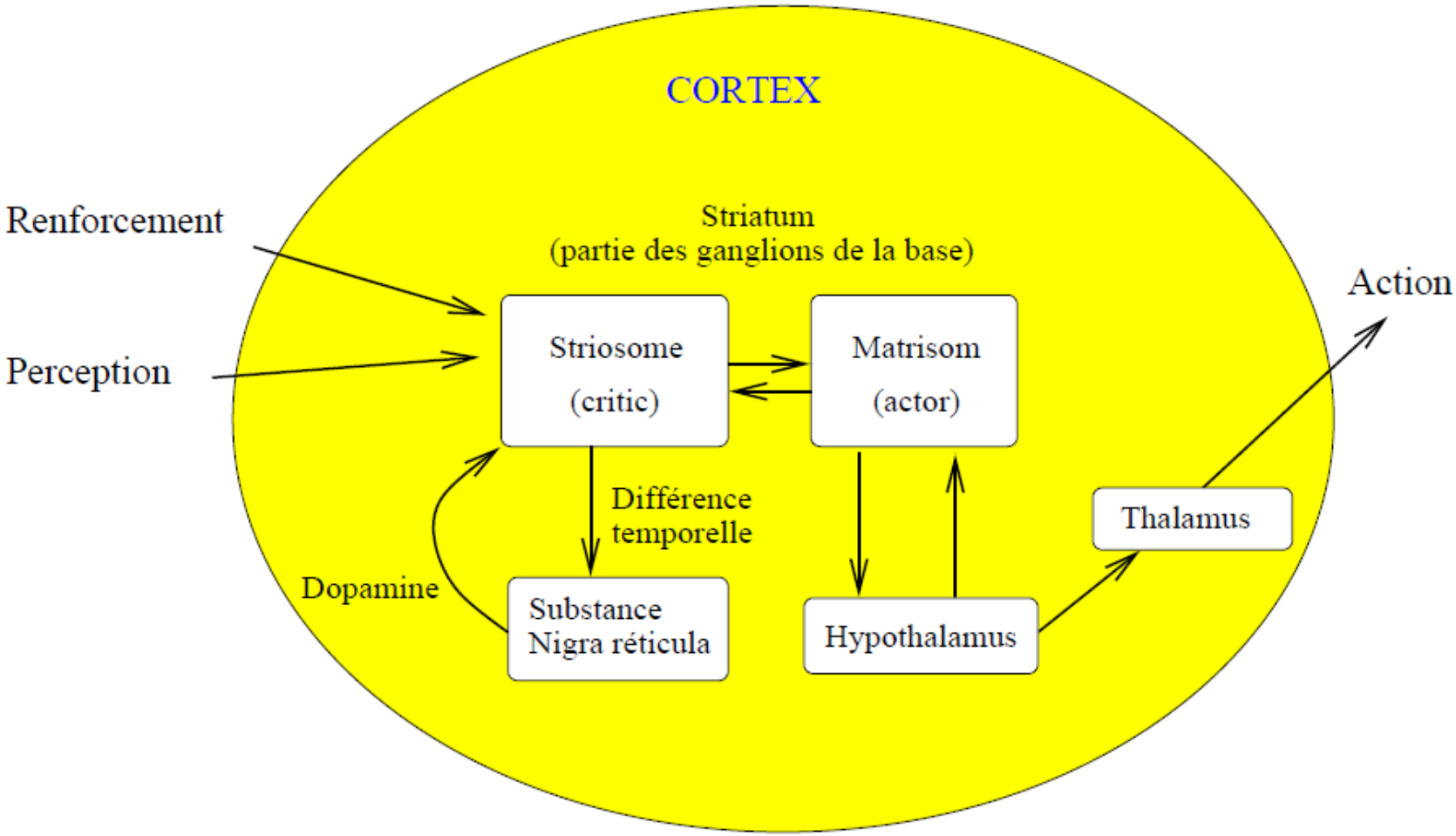
- Déterministe ou stochastique (ex : backgammon)
- Hostile (ex : jeu d'échecs) ou non (ex : jeu Tétris)
- Partiellement observable (ex : robotique mobile)
- Connu ou inconnu (ex : vélo) de l'agent décisionnel

Le renforcement

- Peut récompenser une séquence d'actions → problème du "credit-assignment" : quelles actions doivent être accréditées pour un renforcement obtenu au terme d'une séquence de décisions ?
- Comment sacrifier petit gain à court terme pour privilégier meilleur gain à long terme ?

Lien avec les neurosciences

- Théorie des émotions. Lien entre juste appréciation des émotions en fonction de la situation vécue et capacités de prises de décisions adéquates [Damasio, L'erreur de Descartes, la raison des émotions, 2001].
- Neurotransmetteurs du renforcement : dopamine → surprise.
- Modèle des ganglions de la base (inspiré de [Doya, 1999]).
- ...


CORTEX
Renforcement
Striatum
(partie des ganglions de la base)
Action
Perception
Striosome
(critic)
Matrisom
(actor)
Différence
temporelle
Thalamus
Dopamine
Substance
Nigra réticula
Hypothalamus

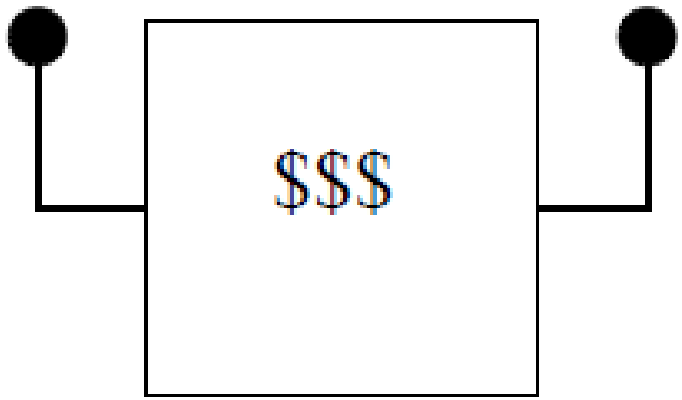
Quelques problématiques de l'A/R

- A/R = résoudre de manière adaptative un problème de contrôle optimal lorsque les dynamique d'état ou les récompenses sont partiellement inconnues. Deux approches possibles :
  - A/R indirect : apprentissage préalable d'un modèle des dynamiques (forme d'apprentissage supervisé), puis utilisation du modèle pour faire de la planification
  - A/R direct : apprentissage direct d'une stratégie d'action sans étape préliminaire de modélisation (peut être intéressant quand les dynamiques d'état sont complexes alors que le contrôleur est simple).
- Même si les dynamiques sont connues, le problème de planification peut être très complexe ! On cherche alors une solution approchée (programmation dynamique avec approximation), ex : le programme TD-gammon.

Dilemme Exploration / Exploitation

Exploiter (agir en maximisant) la connaissance actuelle, ou explorer (améliorer notre connaissance).

Exemple simple : Le bandit à 2 bras



A chaque instant t, le joueur choisit un bras k ($k = 1$ ou 2), reçoit récompense $r_t \sim \nu_k$, où les lois ν_k (une loi pour chaque bras) sont inconnues.

Dilemme Exploration / Exploitation

Objectif : maximiser $\sum_t r_t$.

Ex : récompenses déjà reçues : 6\$; 7\$; 5\$; 4\$ pour le bras gauche, 5\$; 0\$ pour le bras droit. Quel bras choisir ?

Propriété : Il ne faut jamais s'arrêter d'explorer, mais il faut explorer de moins en moins fréquemment $(\log n/n)$.

Différentes stratégies : ε-greedy, Upper-Confidence-Bounds, règles bayésiennes, échantillonnage de Gibbs, indices de Gittings, ...

$\Rightarrow$ A/R = bandit avec dynamique sur l'état.

Quelques réalisations

- TD-Gammon. [Tesauro 1992-1995] : jeu de backgammon. Produit le meilleur joueur mondial !
- KnightCap [Baxter et al. 1998] : jeu d'échec ('2500 ELO)
- Robotique : jongleurs, balanciers, acrobots, ... [Schaal et Atkeson, 1994]
- Robotique mobile, navigation : robot guide au musée Smithonian [Thrun et al., 1999], ...
- Commande d'une batterie d'ascenseurs [Crites et Barto, 1996], Routage de paquets [Boyan et Littman, 1993],
- Ordonnancement de tâches [Zhang et Dietterich, 1995],
- Maintenance de machines [Mahadevan et al., 1997],
- Computer poker (calcul d'un équilibre de Nash avec bandits adversarials), [Alberta, 2008]
- Computer go (algorithmes de bandits hiérarchiques), [Mogo, 2006]

Introduction aux processus de Markov décisionnel

Introduction aux PDM

Les problèmes de décision de ce chapitre sont communément appelés problèmes de décision séquentielle dans l'incertain.

1. Ce pb s'inscrit dans la durée et ce n'est pas en fait un, mais plusieurs pb de décisions en séquence qu'un agent (ou décideur ou encore acteur) doit résoudre ; chaque décision courante influençant la résolution des pb qui suivent. Cf. IA et méthodes de plus court chemin dans un graphe.
2. Ce pb est lié à l'incertitude des conséquences mêmes de chacune des décisions possibles. Ainsi, l'agent ne sait pas à l'avance précisément quels seront les effets des décisions qu'il prend. Cf. théorie classique de maximisation de l'utilité espérée.

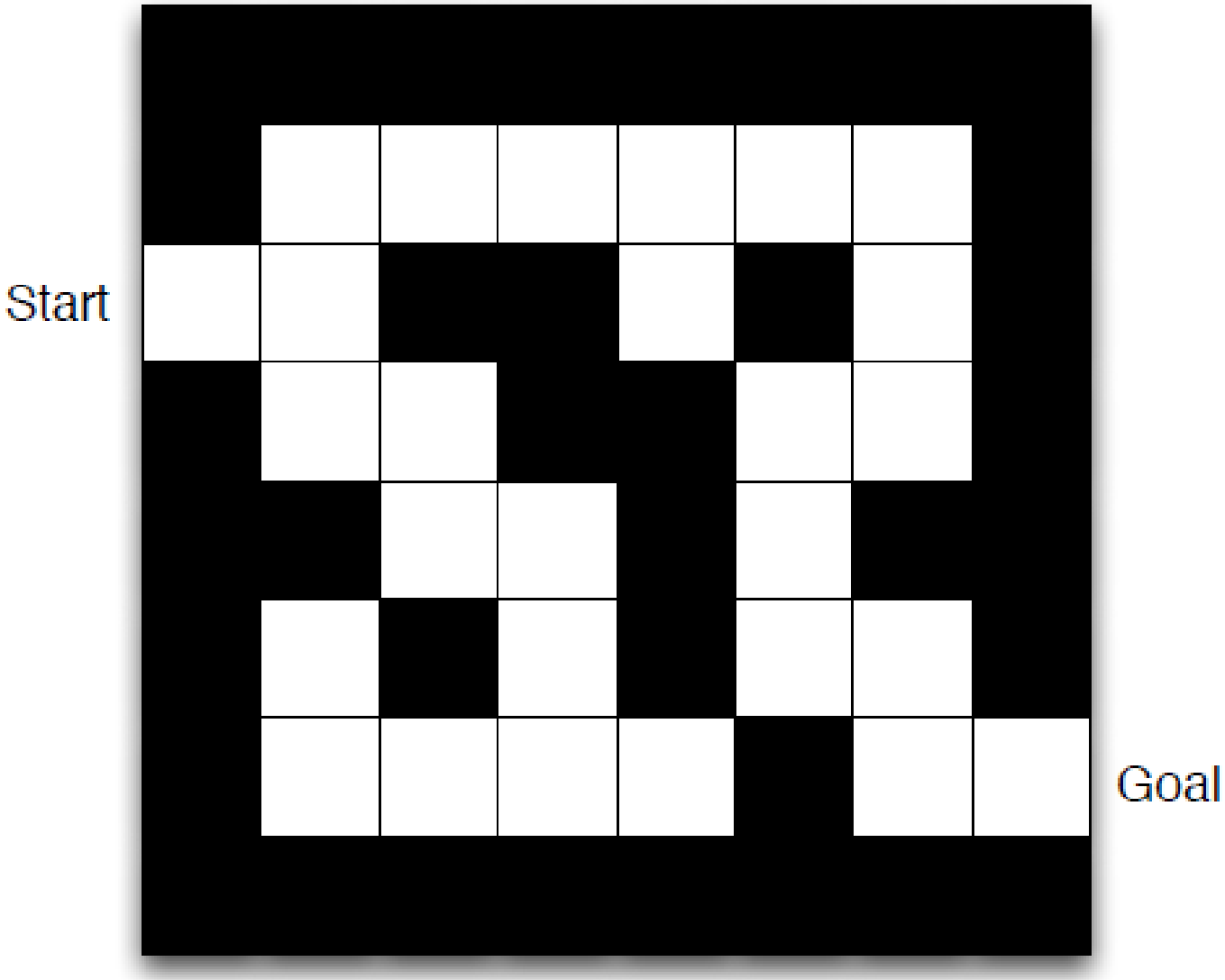
Exempe de l'entretien d'une voiture

Quelle est la meilleure stratégie (ne rien faire, remplacer préventivement, réparer, changer de voiture, etc.) pour minimiser le coût de l'entretien sur le long terme en fonction de l'état de la voiture (présence de panne, usure, âge, etc.) ?

Si on fait l'hypothèse que l'on connaît les conséquences et le coût des différentes actions pour chaque état (par exemple on connaît la probabilité qu'un moteur lâche si on ne répare pas une fuite d'huile) alors on peut modéliser ce problème comme un PDM dont la solution nous donnera, en fonction de l'état de la voiture, l'action optimale.

Ainsi, la suite des actions prises au fur et à mesure de l'évolution de l'état de la voiture permettra, en moyenne, de minimiser son coût d'entretien.

Exemple du labyrinthe



Exemple du labyrinthe

En tout moment, le système se retrouve dans un des $(M+1)$ états possibles : $S = \{0, ..., M\}$
$\Rightarrow$ ici les positions de l'agent, c-à-d. les cases blanches.

À chaque fois que nous observons le système (processus), il faut prendre une décision, et cette décision fait partie d'un ensemble de décisions disponibles $A = \{1, ..., K\}$
$\Rightarrow$ ici les actions possibles sont : gauche, droite, haut, bas.

Notons que pour certains états du processus, certaines des décisions $A = \{1, ..., K\}$ ne peuvent s'appliquer
$\Rightarrow$ ici le seul mouvement possible quand l'agent est sur la case départ est d'aller vers la droite.

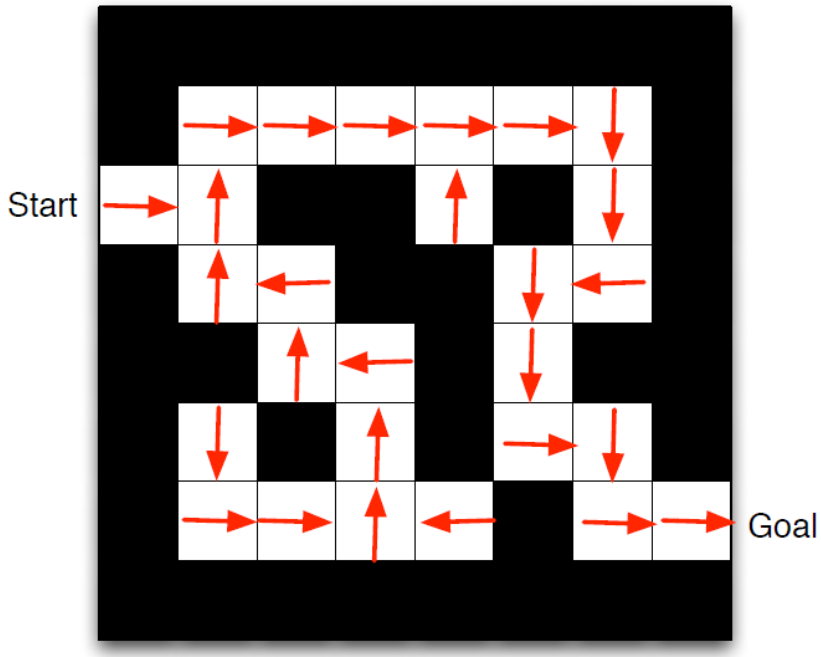
Exemple du labyrinthe

Considérons que le système est dans l'état i au moment de l'observation. Supposons que l'action ou la décision prise est dénotée par $a_i = k$.
Les conséquences de cette décision sont les suivantes :

- coût découlant de cette décision : C_k. Exemple de coût souvent utilisé : coût moyen par unité de temps ;
  $\Rightarrow$ ici nous pouvons par ex. supposer que un pas coûte -1 ;
- nouvelles probabilités de transition entre les états du sytème.

Exemple du labyrinthe

Une politique de décision π est spécification de la décison à prendre pour chaque état possible du système
$\Rightarrow$ ici une politique est donnée par les flèches rouges



Objectif Déterminer une politique optimale relativement à un critère de coût qui doit être spécifié par l'agent de décision

$\Rightarrow$ ici nous cherchons un chemin qui conduit à la sortie et qui minimise le coût total (= le nombre total de pas de l'agent).

Rappel sur les chaînes de Markov

Une chaîne de Markov est un système dynamique à temps discret $(X_t)_{t\in\mathbb{N}} \in \mathscr{X}$, où $\mathscr{X}$ est l'espace d'états (supposé fini ici) tel que

$$\mathbb{P}(X_{t+1} = x | X_t, X_{t-1}, \ldots, X_0) = \mathbb{P}(X_{t=1} = x | X_t).$$

Ainsi, toute l'information pertinente pour la prédiction du futur est contenue dans l'état présent (propriété de Markov).
Une chaîne de Markov sur $\mathscr{X}$ est définie par un état initial x_0 et les probabilités de transition :

$$p(y|x) = \mathbb{P}(X_{t+1} = y | X_t = x).$$

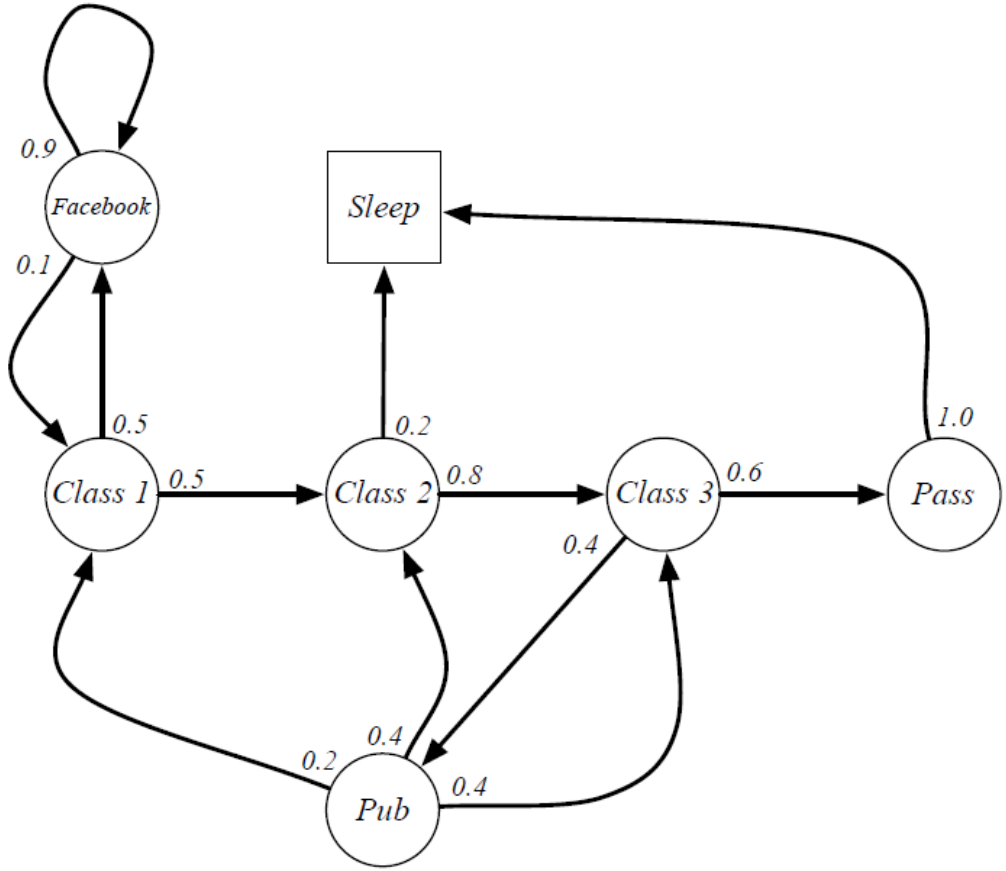
Exemple état X_t pour le vélo = toutes les variables pertinentes pour la prédiction de l'état suivant (position, vitesse, angles, vitesses angulaires...).

Le dilemme de l'étudiant ISM-AG

Voici un exemple de chaîne de Markov. Les états sont :

Facebook (FB),
Class 1 (C1),
Class 2 (C2),
Class 3 (C3),
Pub,
Pass,
Sleep.

L'état Sleep est absorbant (terminal donc).



Le dilemme de l'étudiant ISM-AG

Des exemples de trajectoires de cette chaîne sont en partant de $s_1 = C1$:

- C1 C2 C3 Pass Sleep
- C1 FB FB C1 C3 Sleep
- C1 C2 C3 Pub C2 C3 Pass Sleep
- C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep
- ...

Le dilemme de l'étudiant ISM-AG

Les transitions sont données par la matrice suivante :

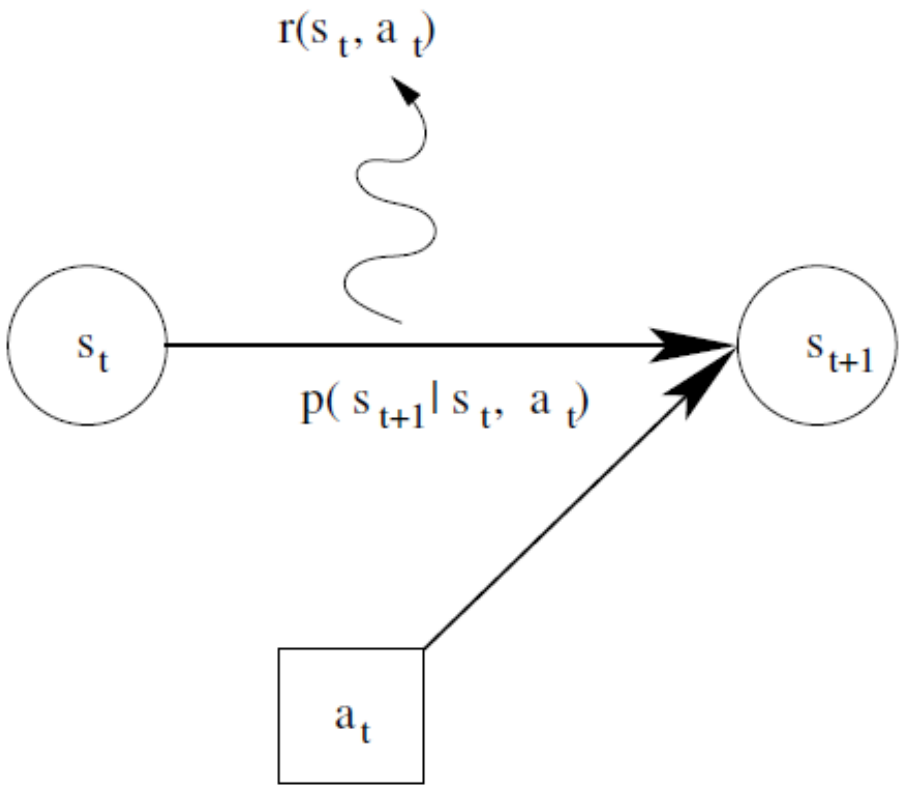
	C1	*C2*	*C3*	*Pass*	*Pub*	*FB*	*Sleep*
C1		0.5				0.5	
C2			0.8				0.2
C3				0.6	0.4		
Pass							1.0
Pub	0.2	0.4	0.4				
FB	0.1					0.9	
Sleep							1

Définition d'un PDM

Les processus décisionnels de Markov sont définis comme des processus stochastiques contrôlés satisfaisant la propriété de Markov, assignant des récompenses aux transitions d'états. On les définit par un quintuplet : (S, A, T, p, r) où :

- S est l'espace d'états dans lequel évolue le processus ;
- A est l'espace des actions qui contrôlent la dynamique de l'état ;
- T est l'espace des temps, ou axe temporel ;
- p sont les probabilités de transition entre états ;
- r est la fonction de récompense sur les transitions entre états.

Définition d'un PDM



Représentation d'un PDM sous la forme d'un diagramme d'influence. A chaque instant t de T, l'action a_t est appliquée dans l'état courant s_t, influençant le processus dans sa transition vers l'état s_{t+1}. La récompense r_t est émise au cours de cette transition.

Définition d'un PDM - Le temps, les états et les actions

Le temps Le domaine T des étapes de décision est un ensemble discret, assimilé à un sous ensemble de $\mathbb{N}$, qui peut être fini ou infini (on parle d'horizon fini ou d'horizon infini).

Les états et les actions Les domaines S et A sont supposés finis, même si de nombreux résultats peuvent être étendus aux cas où S et A sont dénombrables ou continus. Dans le cas général, l'espace A peut être dépendant de l'état courant (A_s pour $s \in S$). De même, S et A peuvent être fonction de l'instant t (S_t et A_t).

Nous nous limiterons ici au cas classique où S et A sont constants tout au long du processus.

Définition d'un PDM - Les transitions

Les probabilités de transition caractérisent la dynamique de l'état du système. Pour une action a fixée, $p(s'|s,a)$ représente la probabilité que le système passe dans l'état s' après avoir exécuté l'action a dans l'état s. On impose classiquement que pour tous s et a,

$$\sum_{s'} p(s'|s,a) = 1.$$

Par ailleurs, on utilise classiquement une représentation matricielle de ces probabilités de transition, en notant P_a la matrice de dimension $|S| \times |S|$ dont les éléments sont pour tous s, s', $P_a(s,s') = p(s'|s,a)$. Les probabilités décrites par p se décrivent donc par $|A|$ matrices P_a, chacune des lignes de ces matrices ayant pour somme 1 : les P_a sont des matrices stochastiques.
Les distributions p vérifient la propriété fondamentale qui donne son nom aux processus décisionnels de Markov considérés ici. Si on note h_t l'historique à la date t du processus,

$$h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t),$$

Définition d'un PDM - La récompense ou le coût

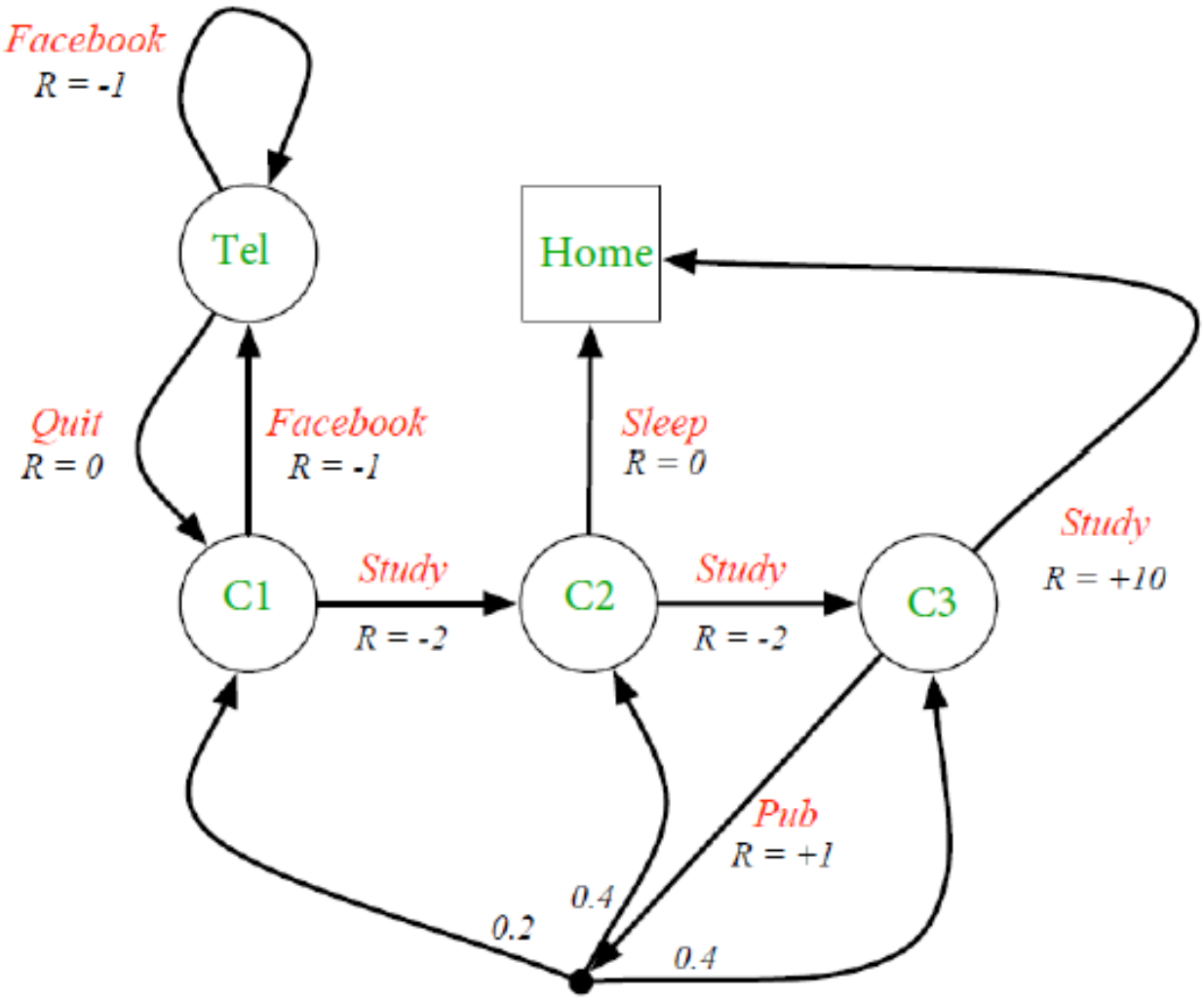
Comme résultat d'avoir choisi l'action a dans l'état s à l'instant t, l'agent décideur reçoit une récompense, ou revenu, $r_t = r(s,a) \in \mathbb{R}$.

Les valeurs de r_t positives peuvent être considérées comme des gains et les valeurs négatives comme des coûts.

Cette récompense peut être instantanément perçue à la date t, ou accumulée de la date t à la date $t+1$, l'important est qu'elle ne dépende que de l'état et de l'action choisie à l'instant courant. La représentation vectorielle de la fonction de récompense $r(s,a)$ consiste en $|A|$ vecteurs r_a de dimension $|S|$.

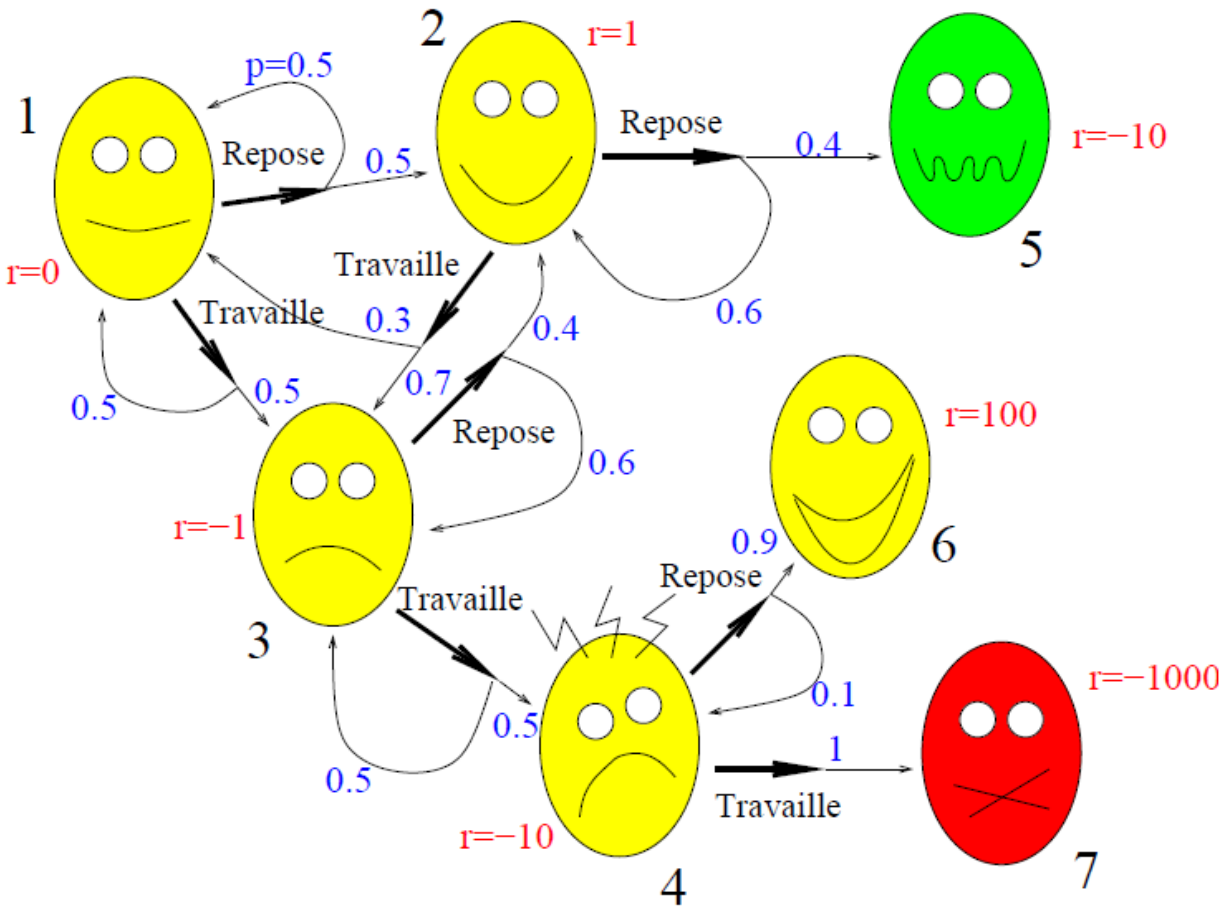
Exemple - Le dilemme de l'étudiant ISM-AG

Traduction en PDM (actions en rouge, états en vert) + ex de récompenses (en noir).



Exemple - Travail ou repos ? ? ?

L'humeur d'une personne oscille entre 7 états. Les états 5, 6, et 7 sont des "états terminaux". Dans les autres états, elle choisit de se reposer ou de travailler.



Exemple - Travail ou repos ? ? ?

Objectif : maximiser la somme des récompenses jusqu'à atteindre un état terminal.

Supposons que la personne connaisse les probabilités de transition et les fonctions récompenses, comment résoudre ce problème ?

Exemple - Maintenance d'un stock

Le responsable d'un entrepot dispose d'un stock x_t d'une marchandise. Il doit satisfaire la demande D_t des clients.

- Pour cela, il peut, tous les mois, décider de commander une quantité a_t supplémentaire à son fournisseur.
- Il paye un coût de maintenance du stock $h(x)$, un coût de commande du produit $C(a)$.
- Il reçoit un revenu $f(q)$ où q est la quantité vendue.
- Si la demande est supérieure au stock actuel, le client va s'approvisionner ailleurs.
- Le stock restant à la fin procure un revenu $g(x)$.
- Contrainte : l'entrepôt à une capacité limitée M.

Exemple - Maintenance d'un stock

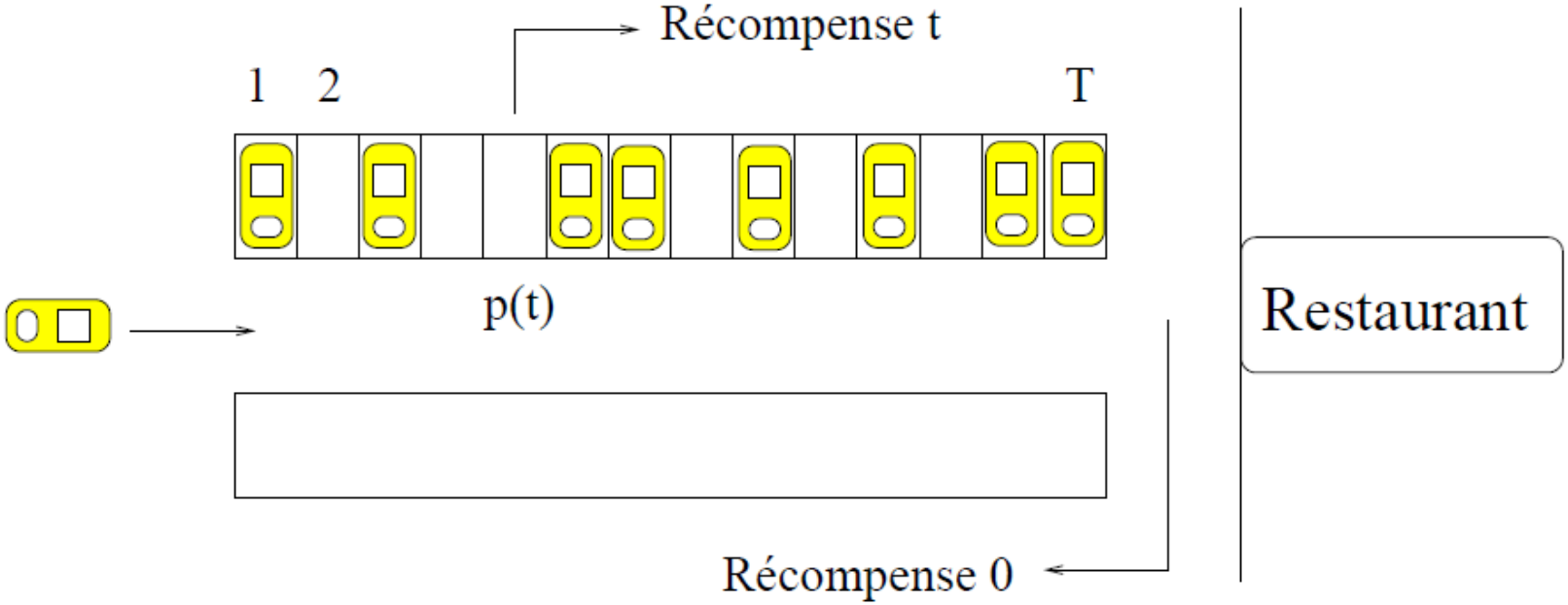
Objectif : maximiser le profit sur une durée donnée T.
Modèle simplifié :

- La demande D_t est une variable aléatoire i.i.d.
- Etats : $x_t \in X = \{0, 1, \ldots, M\}$ quantité (discrète) de produit en stock.
- Décisions : $a_t \in A_{x_t} = \{0, 1, \ldots, M - x_t\}$ commande supplémentaire du produit (ici l'ensemble des actions disponibles à chaque instant dépend de l'état).
- Dynamique : $x_{t+1} = (x_t + a_t - D_t)^+$; ce qui définit les probabilités de transition $p(x_{t+1}|x_t, a_t)$.
- Récompense : $r_t = -C(a_t) - h(x_t + a_t) + f((x_t + a_t - x_{t+1})^+)$.
- Critère à maximiser :

$$\mathbb{E}\left[\sum_{t=1}^{T-1} r_t + g(x_T)\right].$$

Exemple - Problème du parking

Un conducteur souhaite se garer le plus près possible du restaurant.

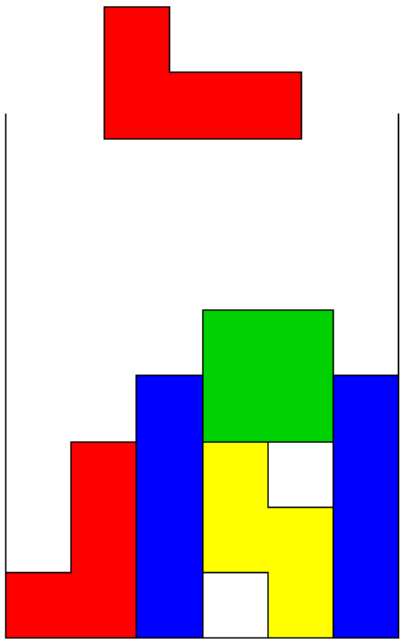


Exemple - Problème du parking

- A chaque instant, l'agent possède 2 actions : continuer ou arrêter.
- Chaque place i est libre avec une probabilité $p(i)$.
- Le conducteur ne peut voir si la place est libre que lorsqu'il est devant. Il décide alors de se garer ou de continuer.
- La place t procure une récompense t. Si le conducteur ne se gare pas, sa récompense est nulle.

Quelle stratégie maximise le gain espéré ?

Exemple - Tétris



- Etats : configuration du mur + nouvelle pièce
- Actions : positions possibles de la nouvelle pièce sur le mur
- Récompense : nombre de lignes supprimées
- Etat suivant : nouvelle configuration du mur + aléa sur la nouvelle pièce.

Pour toute stratégie, le jeu se finit avec probabilité 1. Donc l'espérance de la somme des récompenses à venir est finie.

Difficulté de ce problème : espace d'états très grand (ex : 10^{61} pour hauteur 20, largeur 10, et 7 pièces différentes).

Les règles de décision et les politiques d'actions

Les processus décisionnels de Markov permettent de modéliser la dynamique de l'état d'un système soumis au contrôle d'un agent, au sein d'un environnement stochastique. On nomme alors politique ou stratégie ou plan (notée π), la séquence de règles de décision suivie par l'agent pour choisir à chaque instant l'action à exécuter. Deux distinctions sont essentielles ici.

- Tout d'abord, une politique peut déterminer précisément l'action à effectuer (politique déterministe), ou simplement définir une distribution de probabilité selon laquelle cette action doit être sélectionnée (politique aléatoire).
- Ensuite, une politique peut se baser sur l'historique h_t du processus (politique histoire dépendante), ou peut ne simplement considérer que l'état courant s_t (politique markovienne).

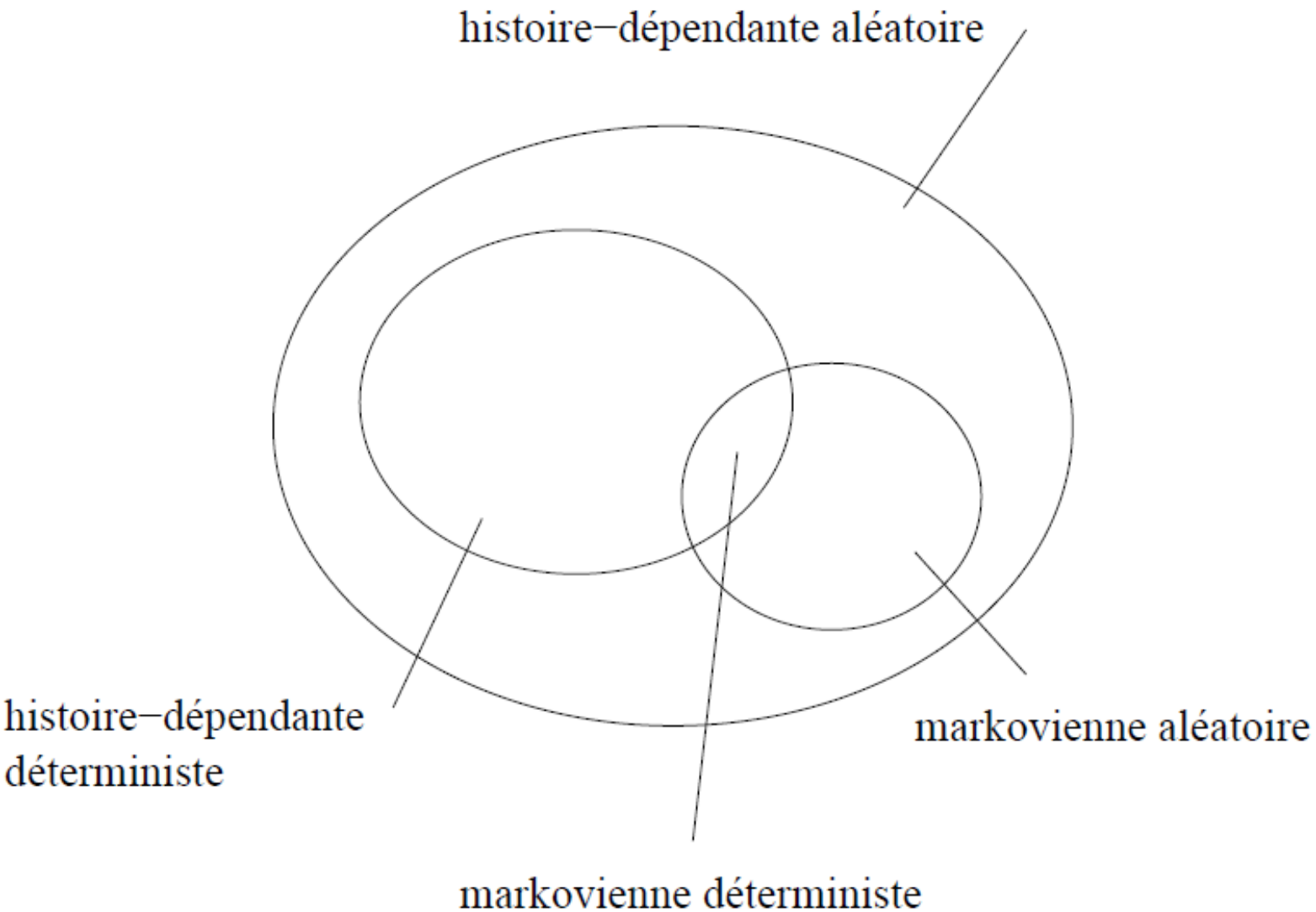
Les règles de décision et les politiques d'actions

On obtient ainsi le tableau suivant :

Politique π_t	Déterministe	Aléatoire
Markovienne	$s_t \mapsto a_t$	$(s_t, a_t) \mapsto [0,1]$
Histoire dépendante	$h_t \mapsto a_t$	$(s_t, h_t) \mapsto [0,1]$

Pour une politique déterministe, $\pi_t(s_t)$ ou $\pi_t(h_t)$ définit l'action a choisie à l'instant t. Pour une politique aléatoire, $\pi_t(a, s_t)$ ou $\pi_t(a, h_t)$ représente la probabilité de sélectionner a.

Les règles de décision et les politiques d'actions



Les règles de décision et les politiques d'actions

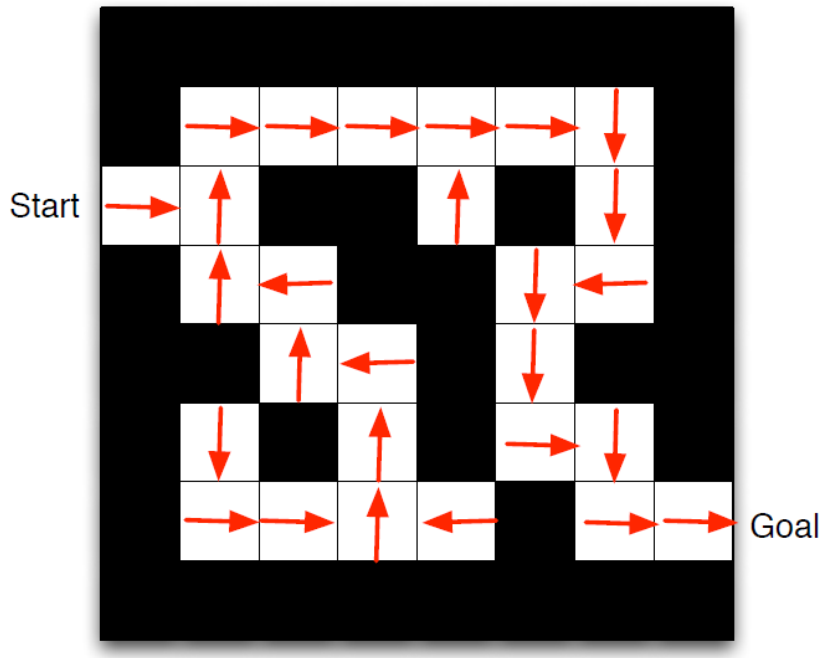
Indépendamment de cela et comme pour le processus décisionnel de Markov lui-même, la définition des politiques peut ou non dépendre explicitement du temps. Ainsi, une politique est stationnaire si pour tout t, $\pi_t = \pi$. Parmi ces politiques stationnaires, les politiques markoviennes déterministes sont centrales dans l'étude des PDM :

$$\pi : s \in S \mapsto \pi(s) \in A.$$

Il s'agit du modèle le plus simple de stratégie décisionnelle.

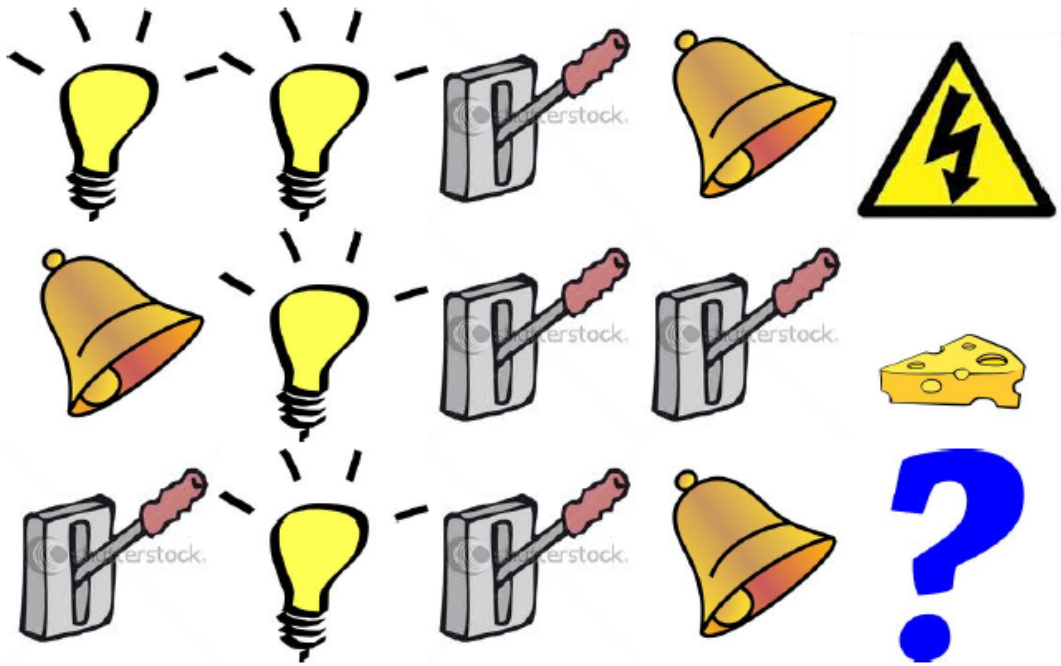
L'exemple du labyrinthe

La politique est stationnaire (ne dépend pas du temps), déterministe (une action par état) et markovienne (ne dépend que de l'état courant et pas de l'hisorique).



L'exemple du rat

Imaginons que le rat a vécu les deux premières séquences d'actions.
Quelle va être son action pour la dernière séquence ?



L'exemple du rat

- Que se passe-t-il si l'état de l'agent est constitué des trois derniers items de la séquence ?
- Que se passe-t-il si l'état de l'agent compte les lumières, les cloches et les leviers ?
- Que se passe-t-il si l'état de l'agent est constitué de toute la séquence ?

Critères de performance

Se poser un problème décisionnel de Markov = rechercher parmi une famille de politiques celles qui optimisent un critère de performance.
Ce critère a pour ambition de caractériser les politiques qui permettront de générer des séquences de récompenses les plus importantes possibles
$\Rightarrow$ mesure du cumul espéré des récompenses instantanées le long d'une trajectoire :

- le critère fini :
$$\mathbb{E}[r_0 + r_1 + r_2 + \ldots + r_{T-1} | s_0];$$
- le critère γ-pondéré :
$$\mathbb{E}[r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots + \gamma^t r_t + \ldots | s_0];$$
- le critère total :
$$\mathbb{E}[r_0 + r_1 + r_2 + \ldots + r_t + \ldots | s_0];$$
- le critère moyen :
$$\lim_{t \to \infty} \frac{1}{t} \mathbb{E}[r_0 + r_1 + r_2 + \ldots + r_{t-1} | s_0].$$

Critères de performance

Les deux caractéristiques communes à ces quatre critères sont

- d'une part, leur formule additive en r_t, qui est une manière simple de résumer l'ensemble des récompenses reçues le long d'une trajectoire
- d'autre part, l'espérance qui est retenue pour résumer la distribution des récompenses pouvant être reçues le long des trajectoires, pour une même politique et un même état de départ.

Critères de performance

Souvent en pratique, on utilise un critère pondéré. Quel est l'intérêt ?

- Mathématiquement pratique pour réduire les récompenses.
- Évite les retours infinis dans les processus de Markov cycliques.
- L'incertitude quant à l'avenir pourrait ne pas être pleinement représentée.
- Si la récompense est financière, les récompenses immédiates ont plus d'intérêt par rapport aux récompenses différées
- Le comportement animal/humain montre une préférence pour la récompense immédiate

Lorsque toutes les séquences se terminent, on peut prendre $\gamma = 1$, la récompense n'est plus pondérée et on obtient le critère total.

Critères de performance

Ce choix d'un cumul espéré est bien sûr important, car il permet d'établir le principe d'optimalité de Bellman (■ les sous-politiques de la politique optimale sont des sous-politiques optimales ■), à la base des nombreux algorithmes de programmation dynamique permettant de résoudre efficacement les PDM.

Dans la suite de ce chapitre, nous allons successivement caractériser les politiques optimales et présenter les algorithmes permettant d'obtenir ces politiques optimales pour chacun des précédents critères.

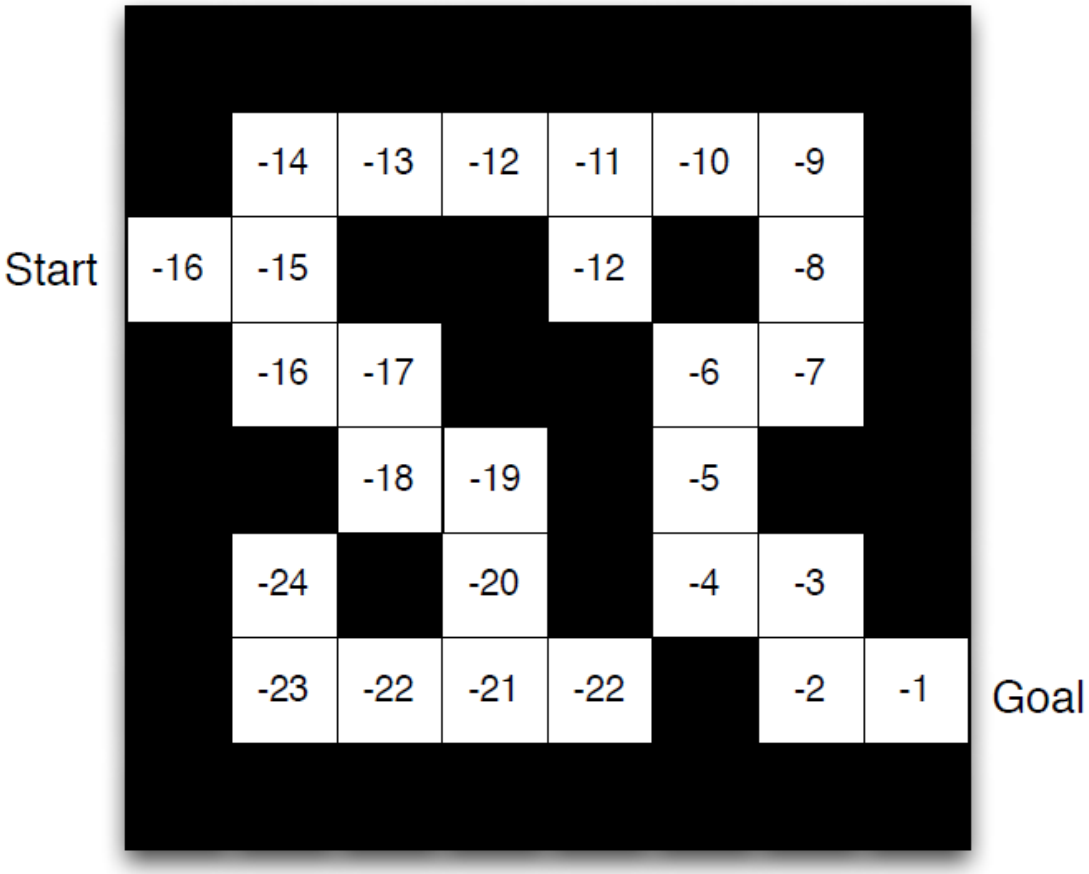
La fonction valeur

La fonction valeur V est la prédiction de la récompense future. Elle attribue à chaque état ce que l'on peut espérer de mieux en moyenne si on est dans cet état. Elle permet donc d'évaluer si un état est bon (prometteur) ou pas.

La valeur $V(s)$ en un état dépend de la récompense immédiate, de la valeur des états résultants $V^{\pi}(s')$ mais aussi de la politque π. On notera donc la fonction valeur V^{π}.

La fonction valeur - Exemple du labyrinthe

Les nombres représentent les valeurs de la fonction $V^{\pi}(s)$ pour chaque état s associées à la politique π donnée précédemment. On voit donc bien quelle est l'action à choisir à chaque pas !



La fonction valeur - Définition et exemples

Plus formellement, la fonction valeur est le gain espéré en partant de l'état s et en suivant la politique π et s'écrit donc

$$V^{\pi}(t,s) = \mathbb{E}[G_t | s_t = s, \pi],$$

où G est le gain. Voici différentes fonctions valeurs possibles (correspondant à différents gains/critètes G).

La fonction valeur - Définition et exemples

- Horizon temporel fini :

$$V^{\pi}(t,s) = \mathbb{E}\Big[\sum_{t'=t}^{T-1} r(s_{t'}, \pi_{t'}(s_{t'})) + R(s_T) | s_t = s, \pi\Big],$$

où R est une fonction récompense terminale. C'est le critère choisi pour l'exemple de la maintenance de stock et le parking.

- Horizon temporel infini avec critère actualisé :

$$V^{\pi}(s) = \mathbb{E}\Big[\sum_{t=0}^{+\infty} \gamma^t r(s_t, \pi_t(s_t)) | s_0 = s, \pi\Big],$$

où $\gamma \in [0,1[$ est un coefficient d'actualisation.

La fonction valeur - Définition et exemples

- Horizon temporel infini avec critère non actualisé :

$$V^{\pi}(s) = \mathbb{E}\Big[\sum_{t=0}^{T} r(s_t, \pi_t(s_t)) | s_0 = s, \pi\Big],$$

où T est le premier instant (aléatoire) où l'on atteint un état absorbant.

- Horizon temporel infini avec critère moyen :

$$V^{\pi}(s) = \lim_{T \to +\infty} \frac{1}{T} \mathbb{E}\Big[\sum_{t=0}^{T-1} r(s_t, \pi_t(s_t)) | s_0 = s, \pi\Big].$$

C'est critère choisi pour l'exemple de système de production plus loin dans le cours.

Exemple du dilemme de l'étudiant ISM-AG

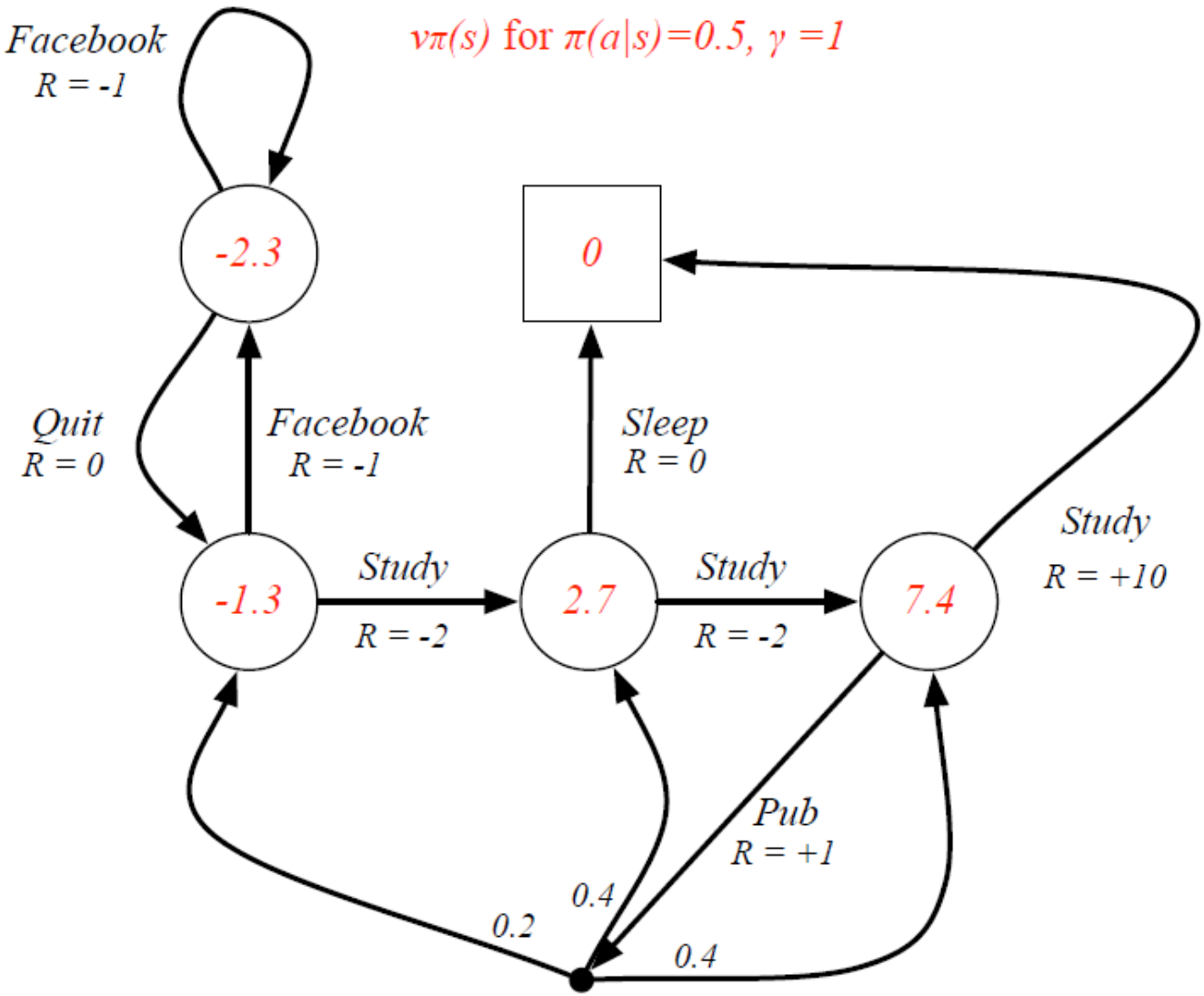
Les nombres représentent les valeurs de la fonction $V^\pi(t,s)$ pour chaque état s pour

- la politique uniforme π où, en chaque état, on choisit l'une des deux actions possibles avec probabilités $1/2$;
- la fonction valeur choisie est

$$\begin{aligned} V^\pi(t,s) &= \mathbb{E}[G_t | s_t = s, \pi] \\ &= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \ldots | s_t = s, \pi] \\ &= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^T R_T | s_t = s, \pi] \end{aligned}$$

où T est le temps aléatoire de fin de trajectoire.

Exemple du dilemme de l'étudiant ISM-AG



La fonction valeur

On remarque que la fonction valeur se décompose en deux parties :

- la récompense immédiate ;
- la fonction valeur de l'état suivant pondérée.

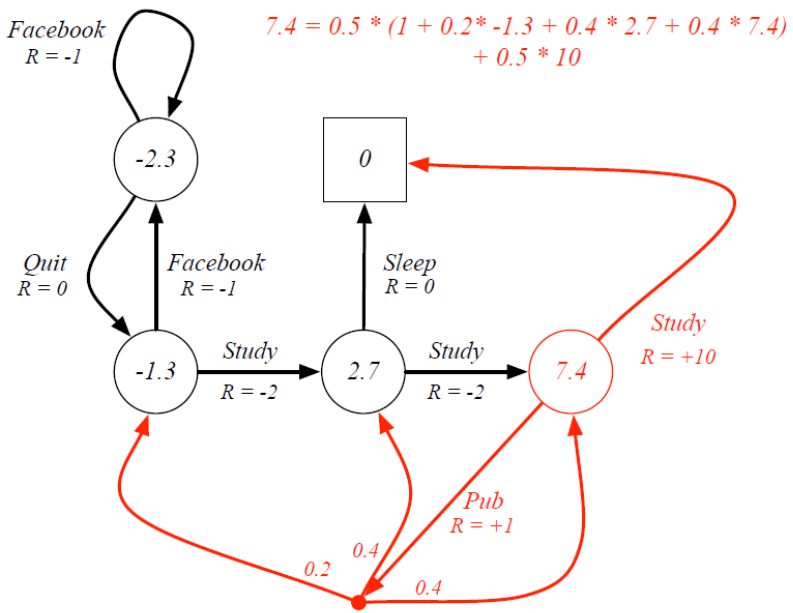
En d'autres termes :

$$V^{\pi}(t,s) = \mathbb{E}[R_{t+1} + \gamma V^{\pi}(t+1, s_{t+1}) | s_t = s, \pi].$$

☞ C'est l'équation de Bellman que nous reverrons plus loin.

La fonction valeur - Dilemme de l'étudiant ISM-AG

$$V^{\pi}(C3) = \frac{1}{2}\Big[R(Pub) + p(C1|C3,Pub) * V^{\pi}(C1) + p(C2|C3,Pub) * V^{\pi}(C2)$$
$$+ p(C3|C3,Pub) * V^{\pi}(C3)\Big] \quad \text{si l'action est Pub}$$
$$+ \frac{1}{2}[R(Study) + p(Sleep|C3,Study) * V^{\pi}(Sleep)] \quad \text{si l'action est Study}$$
$$= \frac{1}{2}(1 + 0.2 * (-1.3) + 0.4 * 2.7 + 0.4 * 7.4) + \frac{1}{2} * 10 = 7.4.$$



Problèmes à horizon temporel fini

Problèmes à horizon temporel fini

Considérons un horizon temporel T. Pour une politique $\pi = (\pi_0, \ldots, \pi_{T-1})$ donnée, le gain en partant de s à l'instant $t \in \{0, \ldots, T\}$ est :

$$V^{\pi}(t,s) = \mathbb{E}\Big[\sum_{t'=t}^{T-1} r(s_{t'}, \pi_{t'}(s_{t'})) + R(s_T) | s_t = s, \pi\Big].$$

Définitions :

- La fonction valeur optimale est

$$V^*(t,s) = \max_{\pi} V^{\pi}(t,s).$$

- Une politique π^* est dite optimale si

$$V^{\pi^*}(t,s) = V^*(t,s).$$

Problèmes à horizon temporel fini

Proposition

(i) Pour une politique π markovienne et déterministe, $\pi = (\pi_t, \ldots, \pi_{T-1})$, la fonction valeur V^π satisfait l'équation de Bellman :

$$V^\pi(t,s) = r(s,\pi_t(s)) + \sum_{s' \in S} p(s'|s,\pi_t(s)) V^\pi(t+1,s')$$

$$V^\pi(T,s) = R(s).$$

(ii) La fonction valeur optimale $V^*(t,s)$ est la solution de l'équation de Bellman optimale :

$$V^*(t,s) = \max_{a \in A} \left\{ r(s,a) + \sum_{s' \in S} p(y|x,a) V^*(t+1,s') \right\}, \quad \text{pour } 0 \leqslant t < T$$

$$V^*(T,s) = R(s)$$

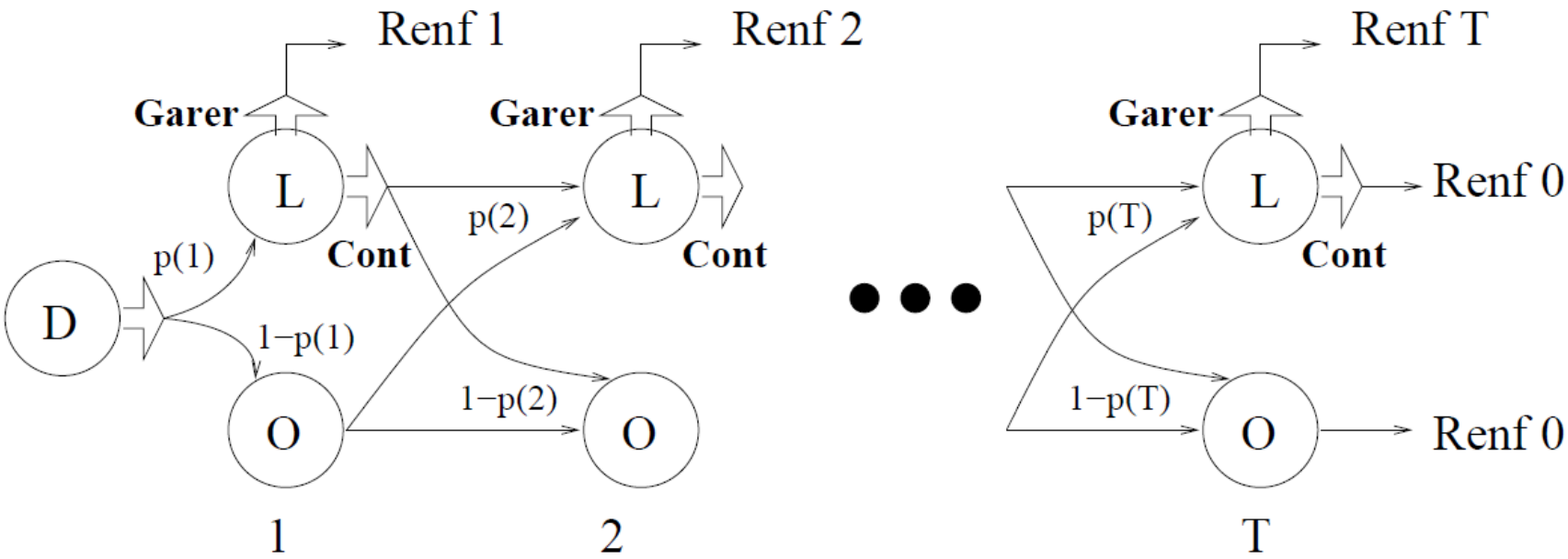
De plus, la politique définie par

$$\pi_t^*(s) \in \arg\max_{a \in A} \left\{ r(s,a) + \sum_{s' \in S} p(s'|s,a) V^*(t+1,s') \right\}, \quad \text{pour } 0 \leqslant t < T$$

est une politique optimale.

Pb à horizon temp. fini - Fct valeur opt. pour le parking

Modélisation du parking par un PDM : L = libre, O = occupé



Pb à horizon temp. fini - Fct valeur opt. pour le parking

Soient $V^*(t,L)$ et $V^*(t,O)$ les récompenses maximales moyennes à la position t lorsque la place est Libre et Occupée.Alors au temps T, on a

$$V^*(T,L) =$$
$$V^*(T,O) =$$

puis au temps $T-1$,

$$V^*(T-1,L) =$$
$$=$$
$$V^*(T-1,O) =$$

et plus généralement, au temps t,

$$V^*(t,L) =$$
$$V^*(t,O) =$$

Enfin, une politique optimale est donnée par l'argument du max.

Problèmes à horizon temporel infini et critère actualisé

Problèmes à horizon temporel infini et critère actualisé

Soit $\pi = (\pi_0, \pi_1, \ldots)$ une politique. Considérons la fonction valeur pour la politique π donnée par

$$V^{\pi}(x) = \mathbb{E}\Big[\sum_{t=0}^{+\infty} \gamma^t r(s_t, \pi_s(s_t)) | s_0 = s, \pi\Big],$$

où $0 \leqslant \gamma < 1$ est un coefficient d'actualisation (ce qui garantit la convergence de la série).
Définissons la fonction valeur optimale

$$V^*(x) = \max_{\pi=(\pi_0,\pi_1,\ldots)} V^{\pi}(x).$$

Problèmes à horizon temporel infini et critère actualisé

Proposition

(i) Pour une politique π stationnaire, i.e. $\pi = (\pi, \pi, \ldots)$, la fonction valeur V^π satisfait l'équation de Bellman :

$$V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V^\pi(s').$$

(ii) La fonction valeur optimale V^* satisfait l'équation de programmation dynamique ou encore équation de Bellman optimale :

$$V^*(s) = \max_{a \in A} \left\{ r(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) V^*(s') \right\}.$$

Problèmes à horizon temporel infini et critère actualisé

☞ On peut écrire les équations de Bellman en chaque état matriciellement (pusiqu'on a un système linéaire) et donc de manière plus concise :

$$V^\pi = R^\pi + \gamma P^\pi V^\pi,$$

où R^π est le vecteur des récompenses de la politique π ; qui se résout en

$$V^\pi = (I - \gamma P^\pi)^{-1} R^\pi.$$

En pratique, on peut aussi procèder itérativement jusqu'à converger.

☞ Les équations de Bellman optimales sont non-linéaires et il n'existe pas de solution close (en général). Plusieurs techniques de résolution sont alors envisageables : value-iteration, policy-iteration, Q-learning, Sarsa, ...

Pb à horizon temp infini et critère actualisé - Fct valeur et politique optimales pour l'exemple ISM-AG

On a d'abord $V^*(Home) =$, puis d'après l'équation de Bellman optimale (ici on a choisi $\gamma = 1$) :

$$V^*(C3) =$$

$$=$$

$$V^*(C2) =$$

$$=$$

$$V^*(C1) =$$

$$=$$

$$V^*(Tel) =$$

$$=$$

Pb à horizon temp infini et critère actualisé - Fct valeur et politique optimales pour l'exemple ISM-AG

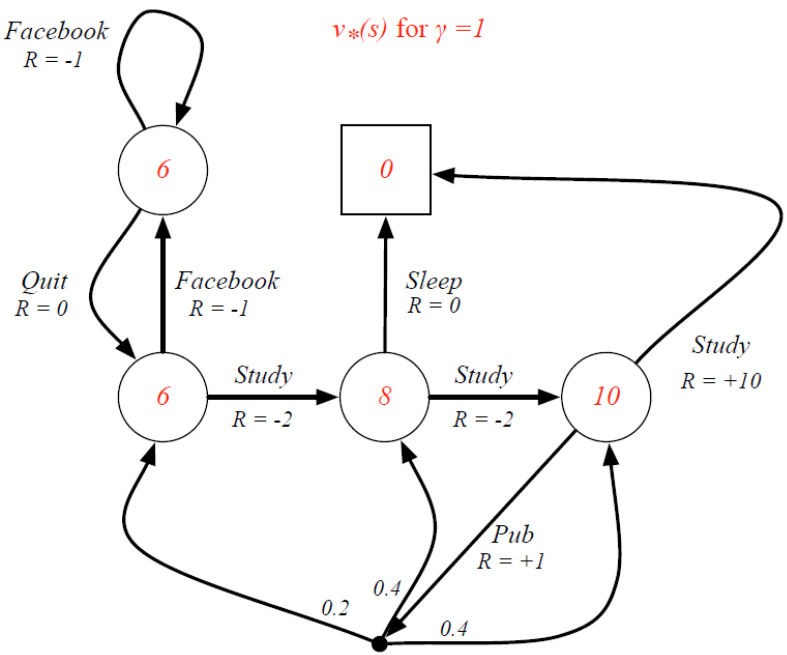
En étudiant tous les chemins possibles, on arrive à

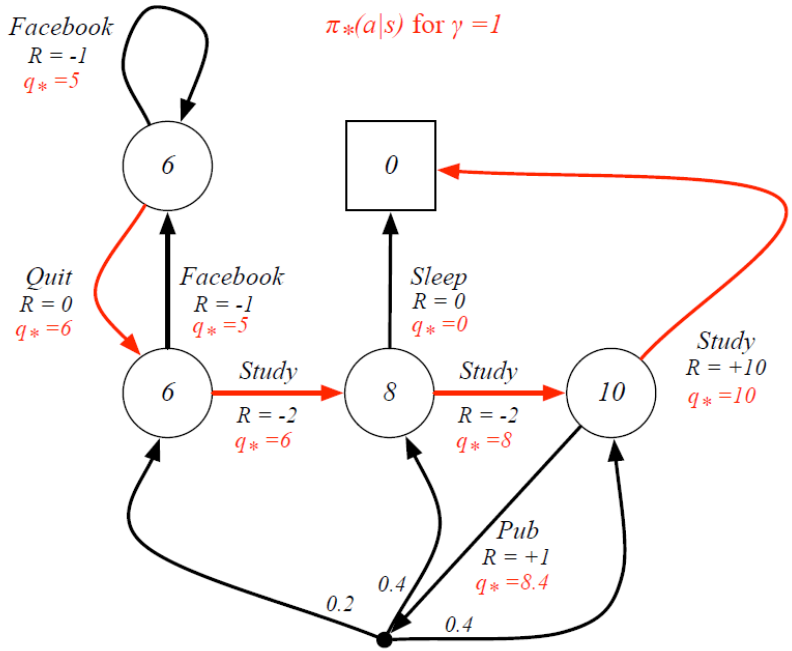
$$V^*(C3) =$$
$$V^*(C2) =$$
$$V^*(C1) =$$
$$V^*(Tel) =$$

On en déduit la politique (intuitive) optimale :

π^*(C1)= , π^*(C2)= , π^*(C3)= et π^*(Tel)= .

Pb à horizon temp infini et critère actualisé - Fct valeur et politique optimales pour l'exemple ISM-AG



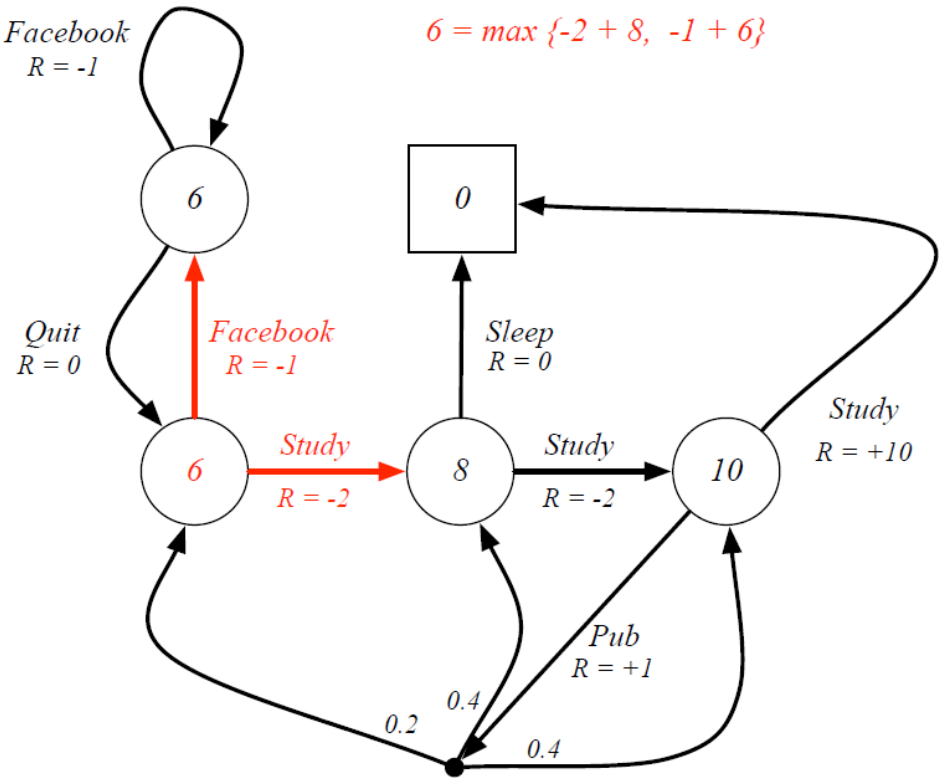


Pb à horizon temp infini et critère actualisé - Fct valeur et politique optimales pour l'exemple ISM-AG

On peut vérifier l'équation de Bellman optimale sur l'état C1 :

$$V^*(C1) =$$

$$=$$



Pb à horizon temp infini et critère actualisé - Fct valeur et politique optimales pour l'exemple ISM-AG

Conclusion

☞ En pratique, nous chercherons la meilleure politique (au sens des critères précédents). Deux alternatives pour cela :

1. on détermine la fonction valeur optimale puis on en déduit la politique optimale - value-based method ;
2. on détermine dirctement la politique optimale sans passer par la fonction valeur - policy-based method.

Un exemple de système de production pour le critère moyen

Système de production pour le critère moyen - Etats

Les états possibles sont les suivants :

Etats	Condition de la machine
0	comme une neuve
1	utilisable avec détérioration mineure
2	utilisable avec détérioration majeure
3	inutilisable

Système de production pour le critère moyen - Probabilités de transition

En utilisant des données historiques, nous sommes en mesure de spécifier les transitions suivantes entre les états d'une semaine à l'autre.

$$\begin{matrix}\text{Etats} \\ \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix}\end{matrix} \quad P= \quad \begin{matrix}\begin{bmatrix} 0 & 1 & 2 & 3 \end{bmatrix} \\ \begin{pmatrix} 0 & \frac{7}{8} & \frac{1}{16} & \frac{1}{16} \\ 0 & \frac{3}{4} & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix}\end{matrix}$$

Puisque nous avons une chaîne de Markov discrète, en invoquant la perte de mémoire du passé, nous en déduisons que les transitions ne dépendent pas des semaines antérieures.

Système de production pour le critère moyen - Décisions

Considérons que lors d'une observation 3 décisions différentes peuvent être prises :

Décision k	Action correspondante	Etats où la décision est applicable
1	Rien faire	0, 1, 2
2	Mise au point (retour à l'état 1)	2
3	Remplacement de machine (retour à l'état 0)	1, 2, 3

Dans cet exemple, les décisions ne dépendent pas du temps.
Les conséquences de la décision k sont les suivantes :

- coût découlant de cette décision : C_k ;
- nouvelles probabilités de transition entre les états du sytème.

Système de production pour le critère moyen - Coûts

Coûts découlant des décisions :

a) Si nous décidons de ne rien faire (décision 1), alors le coût moyen de perte par semaine pour les produits défectueux depend de l'état de la machine
   - coût moyen des produits défectueux de l'état 0 = 0€
   - coût moyen des produits défectueux de l'état 1 = 1000€
   - coût moyen des produits défectueux de l'état 2 = 3000€

b) Coût de maintenance :
   - coût de mise au point = 2000€
   - coût de remplacement d'une machine = 4000€

c) Coût de perte de production par semaine :
   - lors d'une mise au point = 2000€
   - lors du remplacement d'une machine = 2000€

Système de production pour le critère moyen - Coûts totaux

Les coûts totaux par semaine sont donc

Décision	Etat	Produits défectueux	Maintenance	Perte de production	Coût total par semaine
1 = ne rien faire	0 1 2				
2 = mise au point	2				
3 = remplacement	1, 2, 3				

Système de production - Comparaison de 4 politiques

Considérons quatre politiques de décision différentes :

Politique	Description	$d_0(\pi)$	$d_1(\pi)$	$d_2(\pi)$	$d_3(\pi)$
π_a	Remplacer dans l'état 3	1	1	1	3
π_b	Remplacer dans l'état 3, mise au point dans l'état 2	1	1	2	3
π_c	Remplacer dans les états 2 et 3	1	1	3	3
π_d	Remplacer dans les états 1, 2 et 3	1	3	3	3

Dans cet exemple, les politiques ne dépendent pas du temps : elles sont donc stationnaires. De plus, étant donné l'état du système, la décision est unique donc ces politiques sont déterministes. Chaque politique R entraîne de nouvelles probabilités de transitions entre les états du sytème.

Système de production - Comparaison de 4 politiques

(a) La politique π_a entraîne les nouvelles probabilités de transitions suivantes :

Un exemple de système de production pour le critère moyen

(b) La politique π_b entraîne de nouvelles probabilités de transitions entre les états du sytème suivantes :

Système de production - Comparaison de 4 politiques

(c) La politique π_c entraîne les nouvelles probabilités de transitions suivantes :

Système de production - Comparaison de 4 politiques

(d) La politique π_d entraîne les nouvelles probabilités de transitions suivantes :

Système de production - Comparaison de 4 politiques

Nous sommes dans un contexte markovien puisque étant donnés l'état actuel du systéme et la décision prise, toute affirmation sur le futur du système n'est pas affectée par l'information passée (le processus est sans mémoire) :

- les nouvelles probabilités de transition dépendent uniquement de l'état actuel et de la décision prise ;
- le coût moyen (à long terme) dépend uniquement de l'état actuel et de la décision prise.

Système de production - Comparaison de 4 politiques

Résolution Identification d'une politique déterministe optimale par résolution exhaustive : évaluer le coût de chaque politique et choisir celle ayant la plus petite valeur. Considérons le critère défini par le coût moyen par unité de temps :

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} C(X_t)\right].$$

En utilisant que

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=1}^{T} p_{ij}^{k} = \mu_j,$$

on peut démontrer que le coût moyen à long terme par unité de temps est donné par

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} C(X_t) = \sum_{j=0}^{M} \mu_j C(j).$$

Système de production - Comparaison de 4 politiques

Ici les probabilités d'équilibre μ_j sont obtenues en considérant les nouvelles probabilités de transition résultant de l'exécution de la politique et sont telles que $\mu = \mu^\top P$ c-à-d.

$$\begin{cases} \mu_j = \sum_{i=0}^{M} \mu_i p_{ij} \text{ pour tout } j = 0, \ldots, M \\ \sum_{i=0}^{M} \mu_j = \mu_0 + \mu_1 + \cdots + \mu_M = 1 \end{cases} \tag{1}$$

et les coûts $C(j)$ sont les coûts d'exécution associés aux états j. **Considérons la politique π_b.** En traduisant le système dans ce cas, on obtient

$$\begin{cases} \mu_0 = \\ \mu_1 = \\ \mu_2 = \\ \mu_3 = \\ \end{cases} \qquad \Rightarrow \mu_0 = \quad \mu_1 = \quad \mu_2 = \quad \mu_3 =$$

Système de production - Comparaison de 4 politiques

Au niveau des coûts, on a

$$C_0(\pi_b) = \qquad C_1(\pi_b) = \qquad C_2(\pi_b) = \qquad C_3(\pi_b) =$$

Ainsi le coût moyen à long terme pour la politique π_b est donné par

En répétant les étapes pour les trois autres politiques, nous obtenons :

Politique	Probabilités stat	$\mathbb{E}[C]$ en milliers d'euros
π_a		
π_b		
π_c		
π_d		

Système de production - Comparaison de 4 politiques

Conclusion La meilleure politique (celle qui est optimale parmi les quatre politiques proposées et pour ce coût) est

Système de production - Identification politique optimale

Dans l'exemple précédent, nous avons déterminé la politique optimale parmi les quatre politiques en faisant une résolution exhaustive puisqu'il n'y avait que quatre politiques, c-à-d. que nous avons calculé le coût pour TOUTES les politiques. Plus généralement, à partir d'un contexte donné, nous pouvons identifier la politique stationnaire et déterministe optimale par programmation linéaire.

Système de production - Identification politique optimale

Pour ce faire, nous représentons la politique de décision π à l'aide d'un tableau de décisions $D(\pi)$:

$$\begin{array}{cc} & \begin{array}{cccc} & & \text{Décisions } k & \\ [1 & 2 & \ldots & \text{K }] \end{array} \\ \text{Etats } i \quad \begin{bmatrix} 0 \\ 1 \\ \vdots \\ \text{M} \end{bmatrix} & \begin{pmatrix} D_{01} & D_{02} & \ldots & D_{0K} \\ D_{11} & D_{12} & \ldots & D_{1K} \\ \vdots & \vdots & & \vdots \\ D_{M1} & D_{M2} & \ldots & D_{MK} \end{pmatrix} \end{array}$$

Système de production - Identification politique optimale

Dans l'exemple du système de production, la politique π_b est représentée par :

Système de production - Identification politique optimale

Un exemple de politique stationnaire probabiliste pourrait être :

$$\begin{array}{c} \\ \text{états } i \end{array} \begin{array}{c} \text{Décisions } k \\ \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \\ \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 0 & 0.5 \\ 0.3 & 0.2 & 0.5 \\ 0 & 0 & 1 \end{bmatrix}$$

$$P(\text{dec. } 1 \mid \text{état } 1) = P(\text{dec. } 3 \mid \text{état } 1) = 0.5$$
$$P(\text{dec. } 1 \mid \text{état } 2) = 0.3$$
$$P(\text{dec. } 2 \mid \text{état } 2) = 0.2$$
$$P(\text{dec. } 3 \mid \text{état } 2) = 0.5$$

Bien sûr, la somme des lignes vaut 1 : $\sum_{k=1}^{K} D_{ik} = 1$.
Pour formuler le problème de programmation linéaire, nous utilisons la représentation de la matrice $D(\pi)$ pour représenter la politique de décision. Or puisque la programmation linéaire travaille avec des variables continues, alors nous considérons le problème de déterminer une politique probabiliste optimale.

Système de production - Identification politique optimale

Les variables de décisions sont

$$y_{ik} = \mathbb{P}(\text{état} = i \text{ et décision} = k).$$

Or, d'après la définition des probabilités conditionnelles :

$$\begin{aligned} y_{ik} &= \mathbb{P}(\text{état} = i \text{ et décision} = k) \\ &= \mathbb{P}(\text{état} = i)\mathbb{P}(\text{décision} = k|\text{état} = i) \\ &= \mu_i D_{ik}. \end{aligned}$$

De plus, puisque $\sum_{k=1}^{K} D_{ik} = 1$, on a

$$\sum_{k=1}^{K} y_{ik} = \sum_{k=1}^{K} \mu_i D_{ik} = \mu_i \sum_{k=1}^{K} D_{ik} = \mu_i,$$

qui conduit à

$$D_{ik} = \frac{y_{ik}}{\mu_i} = \frac{y_{ik}}{\sum_{k=1}^{K} y_{ik}}.$$

Système de production - Identification politique optimale

Contraintes du problème de programmation linéaire

1. Puisque μ est une probabilité, $\sum_{i=0}^{M} \mu_i = 1$ donc

2. D'après la définition de la probabilité stationnaire : $\mu = \mu^\top P$, on a $\mu_j = \sum_{i=0}^{M} \mu_i p_{ij}$ pour tout $j = 0, \ldots, M$ et donc

   où les $p_{ij}(k)$ sont les probabilités de transitions suite à la décision k.

3. Enfin, on doit avoir $y_{ik} \geqslant 0$ pour $i = 0, \ldots, M$ et $k = 1, \ldots, K$.

Coût

On choisit comme coût le coût moyen à long terme :

$$\begin{aligned}
\mathbb{E}[C] &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} C(X_t) \\
&= \sum_{i=0}^{M} \mu_i C(i) = \sum_{i=0}^{M} \sum_{k=1}^{K} \mu_i C_{ik} D_{ik} \\
&= \sum_{i=0}^{M} \mu_i C(i) \\
&= \sum_{i=0}^{M} \sum_{k=1}^{K} C_{ik} y_{ik}.
\end{aligned}$$

Formulation du problème de programmation linéaire

Ainsi il s'agit de minimiser

$$\mathbb{E}[C] = \sum_{i=0}^{M} \sum_{k=1}^{K} C_{ik} y_{ik}$$

sous les contraintes

$$\begin{cases} [\text{C1}] : \sum_{i=0}^{M} \sum_{k=1}^{K} y_{ik} = 1 & \\ [\text{C2}] : \sum_{k=1}^{K} y_{jk} = \sum_{k=1}^{K} \sum_{i=0}^{M} y_{ik} p_{ij}(k) & \text{pour } j = 0, \ldots, M \\ [\text{C3}] : y_{ik} \geqslant 0 & \text{pour } i = 0, \ldots, M \text{ et } k = 1, \ldots, K. \end{cases}$$

Système de production - Identification politique optimale

Ainsi nous avons $K(M+1)$ variables (positives) et $M+2$ contraintes. Or $\mu_j = \sum_{i=0}^{M} \mu_i p_{ij}$ pour tout $j = 0, \ldots, M$ comporte une contrainte redondante puisque $\sum_{i=0}^{M} \mu_i = 1$ donc finalement, il y a seulement $M+1$ contraintes linéairement indépendantes. Il s'ensuit qu'il y a $(M+1)$ variables de base dans toute solution de base.

Remarque

Pour tout indice $i = 0, \ldots, M$ (i.e. pour tout état i), il doit nécessairement exister au moins un indice $k = 1, \ldots, K$ tel que $y_{ik} > 0$. Sinon, $\sum_{k=1}^{K} y_{ik} = 0$ et D_{ik} ne serait pas défini (division par 0).
Supposons que cet indice est unique et appelons-le k_i. On a donc $y_{ik} = 0$ pour tout $k = 1, \ldots, K$ sauf pour $k = k_i$. Calculons D_{ik_i} :

$$D_{ik_i} = \frac{y_{ik_i}}{\sum_{k=1}^{K} y_{ik}} = \frac{y_{ik_i}}{y_{ik_i}} = 1$$

et la politique optimale est bien déterministe.

Système de production - Identification politique optimale

Il y a 4 états et 3 actions donc potentiellement 12 variables. Puisque

- la décision 2 ne s'applique qu'à l'état 2,
- la décision 3 ne s'applique pas à l'état 0,
- les décisions 1 et 2 ne s'appliquent pas à l'état 3,

Il ne reste donc que 7 variables. Ensuite le coût est donné par

$$\mathbb{E}[C] = \sum_{i=0}^{M} \sum_{k=1}^{K} C_{ik} y_{ik}$$

$$=$$

Système de production - Identification politique optimale

La contrainte [C1] donne :

$$\sum_{i=0}^{M}\sum_{k=1}^{K} y_{ik} =$$

Pour $j = 0$, la contrainte [C2] : $\sum_{k=1}^{K} y_{0k} = \sum_{k=1}^{K}\sum_{i=0}^{M} y_{ik} p_{i0}(k)$ donne

$$\begin{aligned} y_{01} + y_{02} + y_{03} =& (y_{01}p_{00}(1) + y_{02}p_{00}(2) + y_{03})p_{00}(3)) \\ &+ (y_{11}p_{10}(1) + y_{12}p_{10}(2) + y_{13})p_{10}(3)) \\ &+ (y_{21}p_{20}(1) + y_{22}p_{20}(2) + y_{23})p_{20}(3)) \\ &+ (y_{31}p_{30}(1) + y_{32}p_{30}(2) + y_{33})p_{30}(3)) \end{aligned}$$

qui se réduit à

Système de production - Identification politique optimale

en utilisant les tables suivantes :

$$y_{02} = y_{12} = y_{32} = 0$$
$$y_{03} = 0$$
$$y_{31} = y_{32} = 0$$

états

$$P = \begin{bmatrix} \boxed{0} & \frac{7}{8} & \frac{1}{16} & \frac{1}{16} \\ \boxed{0} & \frac{3}{4} & \frac{1}{8} & \frac{1}{8} \\ \boxed{0} & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$k = 3$ appliquée à état 3

états

$$P = \begin{bmatrix} 0 & \frac{7}{8} & \frac{1}{16} & \frac{1}{16} \\ 0 & \frac{3}{4} & \frac{1}{8} & \frac{1}{8} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \boxed{1} & 0 & 0 & 0 \end{bmatrix}$$

$k = 2$ appliquée à état 2

états

$$P = \begin{bmatrix} 0 & \frac{7}{8} & \frac{1}{16} & \frac{1}{16} \\ 0 & \frac{3}{4} & \frac{1}{8} & \frac{1}{8} \\ \boxed{0} & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$k = 3$ appliquée à l'état 2

états

$$P = \begin{bmatrix} 0 & \frac{7}{8} & \frac{1}{16} & \frac{1}{16} \\ 0 & \frac{3}{4} & \frac{1}{8} & \frac{1}{8} \\ \boxed{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$k = 3$ appliquée à l'état 1

états

$$P = \begin{bmatrix} 0 & \frac{7}{8} & \frac{1}{16} & \frac{1}{16} \\ \boxed{1} & 0 & 0 & 0 \\ 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Système de production - Identification politique optimale

Pour $j = 1$, la contrainte [C2] : $\sum_{k=1}^{K} y_{1k} = \sum_{k=1}^{K} \sum_{i=0}^{M} y_{ik} p_{i1}(k)$ donne

$$y_{11} + y_{13} = \frac{7}{8} y_{01} + \frac{3}{4} y_{11} + y_{22}.$$

Pour $j = 2$, la contrainte [C2] : $\sum_{k=1}^{K} y_{2k} = \sum_{k=1}^{K} \sum_{i=0}^{M} y_{ik} p_{i2}(k)$ donne

$$y_{21} + y_{22} + y_{23} = \frac{1}{16} y_{01} + \frac{1}{8} y_{11} + \frac{1}{2} y_{21}.$$

Pour $j = 3$, la contrainte [C2] : $\sum_{k=1}^{K} y_{3k} = \sum_{k=1}^{K} \sum_{i=0}^{M} y_{ik} p_{i3}(k)$ donne

$$y_{33} = \frac{1}{16} y_{01} + \frac{1}{8} y_{11} + \frac{1}{2} y_{21}.$$

Système de production - Identification politique optimale

La solution optimale est alors

$$y_{01} = \frac{2}{21}, \quad (y_{11}, y_{13}) = \left(\frac{5}{7}, 0\right), \quad (y_{21}, y_{22}, y_{23}) = \left(0, \frac{2}{21}, 0\right), \quad y_{33} = \frac{2}{21}.$$

Il reste à utiliser

$$\begin{cases} y_{02} = y_{12} = y_{32} = 0 \\ y_{03} = 0 \\ y_{31} = y_{32} = 0 \end{cases} \qquad \text{et} \quad D_{ik} = \frac{y_{ik}}{\mu_i} = \frac{y_{ik}}{\sum_{k=1}^{K} y_{ik}}$$

pour obtenir la solution optimale

$$(y_{01}, y_{02}, y_{03}) = \left(\frac{2}{21}, 0, 0\right) \qquad \Rightarrow (D_{01}, D_{02}, D_{03}) = (1, 0, 0)$$

$$(y_{11}, y_{12}, y_{13}) = \left(\frac{5}{7}, 0, 0\right) \qquad \Rightarrow (D_{11}, D_{12}, D_{13}) = (1, 0, 0)$$

$$(y_{21}, y_{22}, y_{23}) = \left(0, \frac{2}{21}, 0\right) \qquad \Rightarrow (D_{21}, D_{22}, D_{23}) = (0, 1, 0)$$

$$(y_{31}, y_{32}, y_{33}) = \left(0, 0, \frac{2}{21}\right) \qquad \Rightarrow (D_{31}, D_{32}, D_{33}) = (0, 0, 1)$$

qui s'avère être la politique π_b trouvée précédemment !