Chapitre 2:
Processus de Décision Markoviens (MDP).

MDP = sorte de chaîne de Markov couplée à un système de décision:
l'agent, par une suite de décisions cherche à optimiser une récompense "sur le long terme"

Exemple: une machine peut être dans l'un des états $1, 2, \ldots, 5$
(1 ↑ neuve, 5 ↑ en panne)

Chaque jour, on peut essayer d'améliorer l'état de la machine D_1, ne rien faire D_2 ou en acheter une neuve D_3.
On connaît la probabilité $T(s, a, s')$

de passer de l'état s à l'état s' pour chacune des décisions d. Il y a également un coût C_d pour chacune des décisions

- On note X_s la rentabilité de la machine en état s
- Le jour $j+1$, notre décision de la veille au soir a entraîné un nouvel état s' et une récompense

$$R_{d,s'} = C_d + X_{s'}$$

Question : Comment optimiser la décision que l'on doit prendre

Réponse : Cela dépend...

optimiser le gain (moyen) sur un jour n'est pas du tout la même chose que d'optimiser le gain sur 10 ans ...

Formellement :

Définition : MDP = (S, A, T, R)

- S = ens. des états du système
- A = ens. des décisions / actions possibles
- $T : S \times A \times S \longrightarrow [0,1]$

  $T(s, a, s')$ = Pr(le syst. passe de l'état s à l'état s' lorsque la décision a est prise)
- $R : S \times A \times S \longrightarrow \mathbb{R}$

  $R(s, a, s')$ = récompense immédiate lorsque l'action entreprise est a et que le syst. passe de s à s'

Comme indiqué dans le chapitre précédent, on note S_t et R_t les v.a. indiquant l'état à l'instant t et la récompense à l'instant $t+1$ resp.

R_t et S_t vérifient la propriété de Markov :

$$\Pr(S_{t+1}=s' \mid S_t=s \wedge S_{t-1}=s_{t-1} \wedge \dots \wedge S_0=s_0) = \Pr(S_{t+1}=s' \mid S_t=s)$$

idem pour R_t.

La dynamique du MDP est alors décrite par la fonction p :

$$p(s', r \mid s, a) = \Pr(S_{t+1}=s' \wedge R_{t+1}=r \mid S_t=s \wedge A_t=a)$$

$\forall\, s, s', r, a \in \mathcal{S} \times \mathcal{S} \times \mathcal{R} \times \mathcal{A}$.

En effet, si on connait $p(s', r | s, a)$:

- $p(s' | s, a) = \Pr(S_{t+1} = s' \mid S_t = s \wedge A_t = a)$

$$= \sum_{r \in \mathbb{R}} p(s', r | s, a)$$

$\Rightarrow$ Prob. de transitions d'état

- $r(s, a) = \mathbb{E}(R_{t+1} \mid S_t = s, A_t = a)$

$$= \sum_{r \in \mathbb{R}} r \sum_{s' \in S} \underbrace{p(s', r | s, a)}_{p(s' | s, a)}$$

$\Rightarrow$ Espérance de la récompense pour s et a

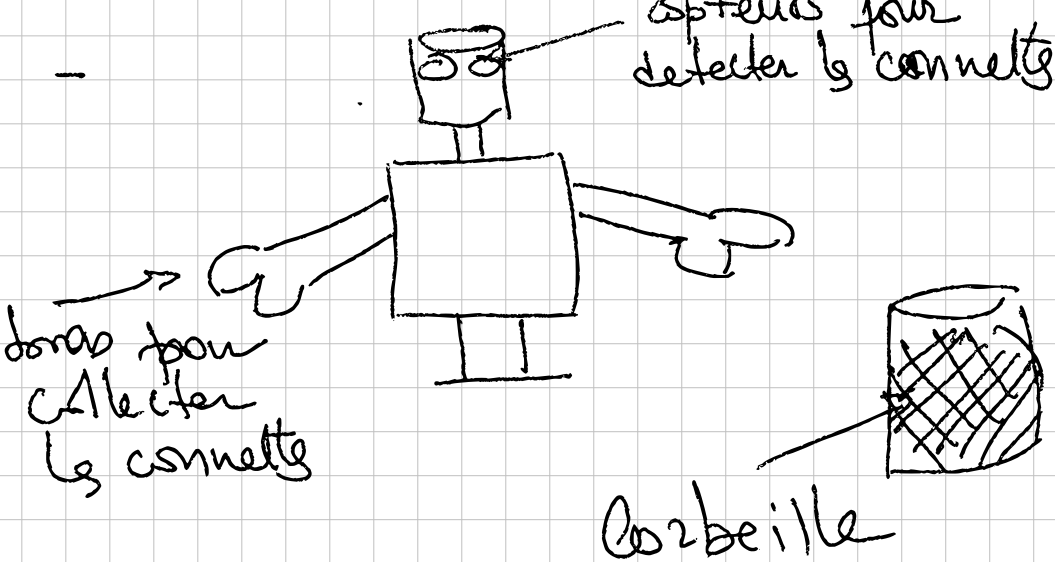
- $r(s, a, s') = \mathbb{E}(R_{t+1} \mid S_t = s \wedge A_t = a \wedge S_{t+1} = s')$

$$= \sum_{r \in \mathbb{R}} r \frac{p(s', r | s, a)}{p(s' | s, a)}$$

En gros, si on connait $p(s', r | s, a)$, on connait tout (du moins localement)

Exemple : Un robot pour le recyclage.

- Le robot collecte des cannettes vides dans un bureau.



En fonction du niveau de charge de la batterie, l'agent qui guide le comportement du robot peut choisir :

- chercher des cannettes pdt une période
- rester sur place et attendre que quelqu'un lui apporte une cannette
- retourner à sa base pour recharger la batterie

$= \{ hig$

$= \{ sea$

$\mathcal{A}(h) =$

Avant de continuer, un peu de formalisme

$\mathcal{S} = \{ high, low \}$

$\mathcal{A} = \{ search, wait, recharge \}$.

avec $\mathcal{A}(h) = \{ s, w \}$; $\mathcal{A}(l) = \{ s, w, r \}$

Le but du robot est de collecter un maximum de canette => "privilégier" la recherche.

Cela décharge la batterie alors que rester sur place la conserve.
La batterie peut se décharger complètement ce qui implique une intervention humaine

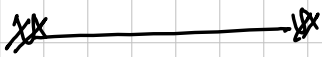
Hypothèses

- Si la batterie est "high" et que le robot fait une période de recherche, elle passe à "low" avec prob. α.
- Si la batterie est "low", alors elle y reste avec prob β.
- Chaque canette récupérée rapporte 1
- Si le robot doit être secouru alors la pénalité est de -3

On note

r_s = l'espérance du nombre de connectés reçus dans une période de recherche

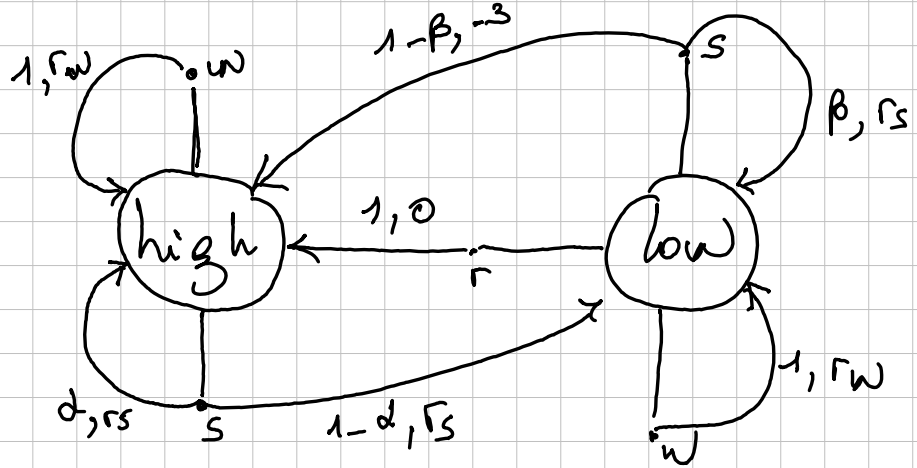
r_w = ~ ~ ~ ~ ~ ~ ~ ~ d'attente



MDP :

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
h	s	h	$1-\alpha$	r_s
h	s	l	α	r_s
h	w	h	1	r_w
l	s	l	β	r_s
l	s	h	$1-\beta$	-3
l	w	l	1	r_w
l	r	h	1	0

Graphe de transitions



Rappelons que le but de l'agent c'est d'optimiser sa récompense (récompense cumulée)

Soit $t \geq 0$, l'agent recevra des récompenses $R_{t+1}, R_{t+2}, \dots$

$\Rightarrow$ Il doit chercher à maximiser l'espérance G_t du gain.

$G_t = f(R_{t+1}, R_{t+2}, \dots)$

Exemple : $f = \Sigma$:

$$G_t = R_{t+1} + R_{t+2} + \dots + R_T$$

T est-ce le dernier pas, il faut être un épisode ou encore correspondre à un état final (ou absorbant)

⚠ Si la tâche ne peut pas être découpée en épisodes, la définition de G_t devient problématique car $T = \infty$ et G_t non convergente...

⇒ On introduit un facteur de dépréciation $0 \leq \gamma \leq 1$:

$$G_t = R_{t+1} + \gamma R_{t+2} + \dots = \sum_{k \geq 0} \gamma^k R_{t+k+1}$$

$$= R_{t+1} + \gamma G_{t+1}$$

Remarques :

- La nouvelle déf. de G_t englobe la précédente ($G_t = 0 \; \forall t > T$ et $\gamma = 1$)
- Si la récompense est finie et non nulle et si $\gamma < 1$, alors G_t est finie.

$$G_t = R_{t+1} + \gamma (R_{t+2} + \gamma G_{t+2})$$

$$= R_{t+1} + \gamma R_{t+2} + \gamma^2 G_{t+2}.$$

???

Politiques et Fonctions de valeurs

But ultime de tout algo de RL:
trouver une politique $\pi : S \longrightarrow A$ qui soit optimale.

i.e. une fct qui à chaque état s, indique l'action à exécuter

La politique π ainsi trouvée est censée maximiser les récompenses.

Rque Une politique peut également être probabiliste:

$$\pi : A, S \longrightarrow [0,1]$$
$$\pi(a,s) = Pr(A_t = a \mid S_t = s)$$

La fonction de valeur d'un état s, sous la politique π est le gain si on commence à l'état s et que l'on suit la politique π :

$$v_\pi(s) = \mathbb{E}(G_t \mid S_t = s)$$

$$= \mathbb{E}\left(\sum_{k \geq 0} \gamma^k R_{t+k+1} \mid S_t = s\right)$$

Fonction valeur des actions :

$$q(s,a) = \mathbb{E}(G_t \mid S_t = s \wedge A_t = a)$$

$$= \mathbb{E}\left(\sum_{k \geq 0} \gamma^k R_{t+k+1} \mid S_t = s\right)$$

De manière récursive :

$$v_\pi(s) = \mathbb{E}_\pi(G_t \mid S_t = s)$$

$$= \mathbb{E}_\pi(R_{t+1} + \gamma G_{t+1} \mid S_t = s)$$

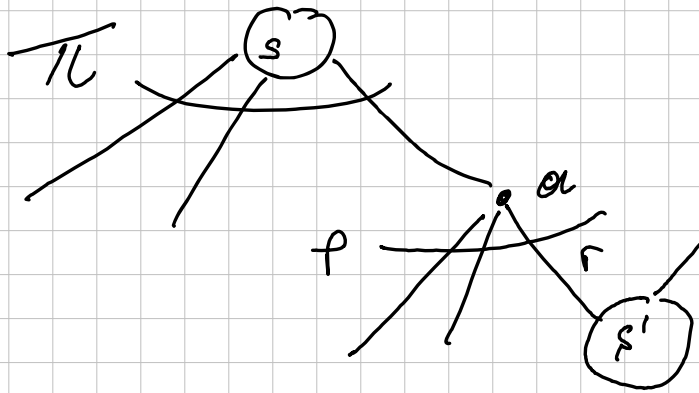
$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \mathbb{E}(G_{t+1} \mid S_{t+1} = s')\right]$$

$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a) \left(r + \gamma v_\pi(s')\right)$$

Eq. de Bellmann :

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left(r + \gamma v_\pi(s')\right)$$

back-up diagram



- Première séance arrêtée là. -
(le 11/10/2022)

Exemple : retour à l'exemple jouet.

Soit π la politique suivante :

$\pi = \{$ high → search ;
low → wait $\}$.

$$\begin{cases} v_\pi(h) = \alpha\left(r_s + \gamma v_\pi(h)\right) + (1-\alpha)\left(r_s + \gamma v_\pi(l)\right) \\ v_\pi(l) = 1\left(r_w + \gamma v_\pi(l)\right) \end{cases}$$

avec des variables : $x = v_\pi(h)$; $y = v_\pi(l)$

$$x = \alpha(r_s + \gamma x) + (1-\alpha)(r_s + \gamma y) \quad (1)$$

$$y = r_w + \gamma y \quad (2)$$

$$(2) \Rightarrow \boxed{y = r_w/(1-\gamma)}$$

Dans (1) : $x = r_s + \alpha\gamma x + (1-\alpha)\gamma \frac{r_w}{1-\gamma}$

$$\Rightarrow (1-\gamma)x = (1-\gamma)r_s + \alpha\gamma x + (1-\alpha)\gamma$$

$$\Rightarrow (1-\gamma-\alpha\gamma)x = (1-\gamma)r_s + (1-\alpha)\gamma$$

$$\Rightarrow \boxed{x = \frac{(1-\gamma)r_s + (1-\alpha)\gamma}{1-\gamma-\alpha\gamma}}$$

Application numérique :

$$\gamma = 10^{-1} \; ; \; \alpha = 1/2 \; ; \; \beta = 1/2$$

$$r_6 = 10 \; ; \; r_w = 2$$

$$\Rightarrow \frac{\vartheta}{\pi}(w) = x = \frac{(1-10^{-1})10 + (1-1/2)10^{-1}}{1 - 10^{-1} - \frac{1}{2}10^{-1}}$$

$$= \frac{(10-1) + .05}{.9 - .05} = \frac{9.05}{.85} = \frac{905}{85}$$

$$= 10.65$$

$$\frac{\vartheta}{\pi}(e) = y = \frac{2}{1-10^{-1}} = \frac{2}{.9} = \frac{20}{9} = 2.22$$

Politiques optimales / fonctions de val. optimale

Déf. π et π' deux politiques

$\pi \geq \pi' \Leftrightarrow v_\pi(s) \geq v_{\pi'}(s) \quad \forall s \in \mathcal{S}$.

$\exists$ au moins une politique π_* meilleure que toutes les autres politiques

π_* : politique optimale

A ces politiques est associée une fonction de valeur optimale

$$v_*(s) = \max_\pi v_\pi(s) \quad \forall s \in \mathcal{S}$$

On y associe également une fonction de valeur état-action optimale :

$$q_*(s,a) = \max_\pi q_\pi(s,a) \quad \forall s, a \in \mathcal{S} \times \mathcal{A}$$

Remarque,

$$q_*(s,a) = \mathbb{E}\left(R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s \wedge A_t = a\right)$$

Eq. de Bellman pour v_* :

$$v_*(s) = \max_a q_*(s,a)$$

$$= \max_a \mathbb{E}_\pi\left(G_t \mid S_t = s \wedge A_t = a\right)$$

$$= \max_a \mathbb{E}_\pi\left(R_{t+1} + \gamma G_{t+1} \mid S_t = s \wedge A_t = a\right)$$

$$= \max_a \mathbb{E}_\pi\left(R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s \wedge A_t = a\right)$$

$$v_*(s) = \max_a \sum_{s',r} p(s', r \mid s, a)\left(r + \gamma v_*(s')\right)$$

idem pour q_* :

$$q_*(s,a) = \sum_{s',r} p(s', r \mid s, a)\left(r + \max_{a'} q_*(s', a')\right)$$