UE Apprentissage par renforcement - séance 4

Valentin Emiya

M2 IAAA

18 décembre 2019

Plan

Programme de l'UE

En 7 séances de 4h :

1. Bandits (1/2) : notions et strategies de base, UCB

   Exploration/exploitation
2. Bandits (2/2) : Thomson sampling, bandits contextuels

   contexte
3. Monte-Carlo Tree Search

   Environnement connu, simulations
4. **Processus de décision de Markov**

   Cas général d'environnements et de stratégies
5. TD learning
6. Miniprojet
7. Miniprojet

Programme de la séance

Séance consacrée aux processus de décision de Markov (MDP) :

- thème : modélisation des problèmes d'AR
  - comment modéliser des problèmes réels ?
  - comment formaliser les aspects aléatoires (environnement et agent) ?
  - comment formuler l'objectif du problème ?

    Notion de retour (gain à long terme)
  - comment analyser les stratégies ?

    Notions de fonctions d'évaluation, équations de Bellman
  - comment caractériser les stratégies optimales ?

    Équations d'optimalité de Bellman
  - remarque : pas d'algorithme de résolution cette semaine !

    $\rightarrow$ *on prépare la semaine consacrée à l'algorithme TD-learning.*
- source principale : Chapitre 3 du livre de Richard S. Sutton et Andrew G. Barto, *Reinforcement Learning : An Introduction*, Second Edition, MIT Press, Cambridge, MA, 2018
- en mode cours/TD (exercices sans ordi)

Plan

Variable alétaoire discrète

Loi de probabilité, espérance

Une **variable aléatoire** (v.a.) $X \in \mathcal{X}$ définie sur un ensemble fini $\mathcal{X}$ est caractérisée par sa loi de probabilité p_X : pour $x \in \mathcal{X}$, $p_X(x)$ est la probabilité que la v.a. X soit égale à une *valeur* donnée x. $x \mapsto p_X(x)$ est une loi de probabilité si $\forall x \in \mathcal{X}, p_X(x) \geq 0$ et $\sum_{x \in \mathcal{X}} p_X(x) = 1$.

Notations : suivant le contexte, on notera de façon équivalente $p_X(x)$, $p(x)$ ou $\mathbb{P}(X = x)$.

Soit $f : \mathcal{X} \to \mathbb{R}$. L'**espérance** de $f(X)$ est $\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x) p_X(x)$.

Exemple

Pour $\mathcal{X} = \{\text{pile}, \text{face}\}$, X peut être le résultat du lancer d'une pièce déséquilibrée, avec

$$\begin{cases} \mathbb{P}(X = \text{pile}) & = 0.6 \\ \mathbb{P}(X = \text{face}) & = 0.4 \end{cases}$$

Si $f(\text{pile}) = 1$ et $f(\text{face}) = 0$, alors $\mathbb{E}[f(X)] = ?$.
Si $g(\text{pile}) = 1$ et $g(\text{face}) = -1$, alors $\mathbb{E}[g(X)] = ?$.

Variable alétaoire discrète

Loi de probabilité, espérance

Une **variable aléatoire** (v.a.) $X \in \mathcal{X}$ définie sur un ensemble fini $\mathcal{X}$ est caractérisée par sa loi de probabilité p_X : pour $x \in \mathcal{X}$, $p_X(x)$ est la probabilité que la v.a. X soit égale à une *valeur* donnée x. $x \mapsto p_X(x)$ est une loi de probabilité si $\forall x \in \mathcal{X}, p_X(x) \geq 0$ et $\sum_{x \in \mathcal{X}} p_X(x) = 1$.

Notations : suivant le contexte, on notera de façon équivalente $p_X(x)$, $p(x)$ ou $\mathbb{P}(X = x)$.

Soit $f : \mathcal{X} \to \mathbb{R}$. L'**espérance** de $f(X)$ est $\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x) p_X(x)$.

Exemple

Pour $\mathcal{X} = \{\text{pile}, \text{face}\}$, X peut être le résultat du lancer d'une pièce déséquilibrée, avec

$$\begin{cases} \mathbb{P}(X = \text{pile}) & = 0.6 \\ \mathbb{P}(X = \text{face}) & = 0.4 \end{cases}$$

Si $f(\text{pile}) = 1$ et $f(\text{face}) = 0$, alors $\mathbb{E}[f(X)] = 1 \times 0.6 + 0 \times 0.4 = 0.6$.
Si $g(\text{pile}) = 1$ et $g(\text{face}) = -1$, alors $\mathbb{E}[g(X)] = ?$.

Variable alétaoire discrète

Loi de probabilité, espérance

Une **variable aléatoire** (v.a.) $X \in \mathcal{X}$ définie sur un ensemble fini $\mathcal{X}$ est caractérisée par sa loi de probabilité p_X : pour $x \in \mathcal{X}$, $p_X(x)$ est la probabilité que la v.a. X soit égale à une *valeur* donnée x. $x \mapsto p_X(x)$ est une loi de probabilité si $\forall x \in \mathcal{X}, p_X(x) \geq 0$ et $\sum_{x \in \mathcal{X}} p_X(x) = 1$.

Notations : suivant le contexte, on notera de façon équivalente $p_X(x)$, $p(x)$ ou $\mathbb{P}(X = x)$.

Soit $f : \mathcal{X} \to \mathbb{R}$. L'**espérance** de $f(X)$ est $\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x) p_X(x)$.

Exemple

Pour $\mathcal{X} = \{\text{pile}, \text{face}\}$, X peut être le résultat du lancer d'une pièce déséquilibrée, avec

$$\begin{cases} \mathbb{P}(X = \text{pile}) & = 0.6 \\ \mathbb{P}(X = \text{face}) & = 0.4 \end{cases}$$

Si $f(\text{pile}) = 1$ et $f(\text{face}) = 0$, alors $\mathbb{E}[f(X)] = 1 \times 0.6 + 0 \times 0.4 = 0.6$.
Si $g(\text{pile}) = 1$ et $g(\text{face}) = -1$, alors $\mathbb{E}[g(X)] = 1 \times 0.6 - 1 \times 0.4 = 0.2$.

Plusieurs v.a.

Lois jointe/conditionnelle/marginale, Bayes

Soient $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$ deux variables aléatoires.

Loi jointe : (X, Y) est une variable aléatoire sur $\mathcal{X} \times \mathcal{Y}$ et pour un couple de valeurs fixées $(x, y) \in \mathcal{X} \times \mathcal{Y}$, la probabilité que $X = x$ et $Y = y$ s'écrit $p_{X,Y}(x, y)$, $p(x, y)$ ou $\mathbb{P}(X = x, Y = y)$.

Loi conditionnelle : $X|Y$ est une variable aléatoire sur $\mathcal{X}$ [1] et on note $p_{X|Y=y}(x)$, $p(x|y)$ ou $\mathbb{P}(X = x|Y = y)$ sa loi, dite loi de X conditionnellement à Y.
L'**espérance conditionnelle** est $\mathbb{E}[f(X)|Y = y] = \sum_{x \in \mathcal{X}} f(x)p(x|y)$.

Théorème de Bayes :

?

Loi marginale : $p(x)$ s'obtient en *marginalisant* la loi jointe $p(x, y)$ par rapport à y :

$$p(x) = ?$$

1. En particulier, on a $\forall y \in \mathcal{Y}, \sum_{x \in \mathcal{X}} p(x|y) = 1$.

Plusieurs v.a.

Lois jointe/conditionnelle/marginale, Bayes

Soient $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$ deux variables aléatoires.

Loi jointe : (X, Y) est une variable aléatoire sur $\mathcal{X} \times \mathcal{Y}$ et pour un couple de valeurs fixées $(x, y) \in \mathcal{X} \times \mathcal{Y}$, la probabilité que $X = x$ et $Y = y$ s'écrit $p_{X,Y}(x, y)$, $p(x, y)$ ou $\mathbb{P}(X = x, Y = y)$.

Loi conditionnelle : $X|Y$ est une variable aléatoire sur $\mathcal{X}$ [1] et on note $p_{X|Y=y}(x)$, $p(x|y)$ ou $\mathbb{P}(X = x|Y = y)$ sa loi, dite loi de X conditionnellement à Y.
L'**espérance conditionnelle** est $\mathbb{E}[f(X)|Y = y] = \sum_{x \in \mathcal{X}} f(x)p(x|y)$.

Théorème de Bayes :

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

Loi marginale : $p(x)$ s'obtient en *marginalisant* la loi jointe $p(x, y)$ par rapport à y :

$$p(x) = ?$$

1. En particulier, on a $\forall y \in \mathcal{Y}, \sum_{x \in \mathcal{X}} p(x|y) = 1$.

Plusieurs v.a.

Lois jointe/conditionnelle/marginale, Bayes

Soient $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$ deux variables aléatoires.

Loi jointe : (X, Y) est une variable aléatoire sur $\mathcal{X} \times \mathcal{Y}$ et pour un couple de valeurs fixées $(x, y) \in \mathcal{X} \times \mathcal{Y}$, la probabilité que $X = x$ et $Y = y$ s'écrit $p_{X,Y}(x, y)$, $p(x, y)$ ou $\mathbb{P}(X = x, Y = y)$.

Loi conditionnelle : $X|Y$ est une variable aléatoire sur $\mathcal{X}$ [1] et on note $p_{X|Y=y}(x)$, $p(x|y)$ ou $\mathbb{P}(X = x|Y = y)$ sa loi, dite loi de X conditionnellement à Y.
L'**espérance conditionnelle** est $\mathbb{E}[f(X)|Y = y] = \sum_{x \in \mathcal{X}} f(x)p(x|y)$.

Théorème de Bayes :

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

Loi marginale : $p(x)$ s'obtient en *marginalisant* la loi jointe $p(x, y)$ par rapport à y :

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

1. En particulier, on a $\forall y \in \mathcal{Y}, \sum_{x \in \mathcal{X}} p(x|y) = 1$.

Exemple

Soit $X \in \{\text{p(ile)}, \text{f(ace)}\}$ une v.a. désignant le résultat du lancer de pièces.
Soit $Y \in \{\text{é(quilibrée)}, \text{d(éséquilibrée)}\}$ une v.a. indiquant si la pièce lancée est équilibrée ou pas.
Il y a une chance sur cinq de lancer une pièce déséquilibrée et on obtient alors pile dans 60% des cas.
Donnez les lois de X, de Y, de $X|Y$, de $Y|X$ et de (X, Y).

Exemple

Soit $X \in \{\text{p(ile)}, \text{f(ace)}\}$ une v.a. désignant le résultat du lancer de pièces.
Soit $Y \in \{\text{é(quilibrée)}, \text{d(éséquilibrée)}\}$ une v.a. indiquant si la pièce lancée est équilibrée ou pas.
Il y a une chance sur cinq de lancer une pièce déséquilibrée et on obtient alors pile dans 60% des cas.
Donnez les lois de X, de Y, de $X|Y$, de $Y|X$ et de (X, Y).

y	e	d
$p(y)$	0.8	0.2

x	p	f
$p(x)$	0.52	0.48

$p(x\|y)$	$y = e$	$y = d$
$x = p$	0.5	0.6
$x = f$	0.5	0.4

$p(y\|x)$	$y = e$	$y = d$
$x = p$	$\frac{0.4}{0.52} \approx 0.77$	$\frac{0.12}{0.52} \approx 0.23$
$x = f$	$\frac{0.4}{0.48} \approx 0.83$	$\frac{0.08}{0.48} \approx 0.17$

$p(x, y)$	$y = e$	$y = d$
$x = p$	0.4	0.12
$x = f$	0.4	0.08

Plan

Processus de décision de Markov fini

Définition

Modèle d'un agent dans un environnement, un MDP fini est défini par :

- un ensemble fini d'états de l'environnement $\mathcal{S} = \{s_i, 0 \leq i < S\}$
- un ensemble fini d'actions de l'agent $\mathcal{A} = \{a_i, 0 \leq i < A\}$
- un ensemble fini de récompenses immédiates $\mathcal{R} = \{r_i \in \mathbb{R}, 0 \leq i < R\}$
- la loi

$$p(s', r|s, a) = p(S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a)$$

caractérisant la probabilité qu'ayant effectué l'action a dans l'état s de l'environnement à l'instant $t-1$, l'agent reçoive la récompense immédiate r et l'environnement se retrouve dans l'état s' à l'instant t.

Remarque : dans un cadre plus général, les actions possibles peuvent dépendre de l'état courant s, ce que l'on note $\mathcal{A}(s) = \{a_{s,i}, 0 \leq i < A_s\}$.

Processus de décision de Markov fini

Exemple

Exemple

On définit un MDP fini dans lequel un environnement est composé de quatre pièces équilibrées marquées d'une lettre e et une pièce déséquilibrée portant la lettre d,dont la probabilité d'obtenir pile est notée α ; une des pièces est sélectionnée à chaque instant, ce qui constitue l'état ; un agent peut choisir entre garder la pièce courante (action g) ou mélanger les cinq pièces et en tirer une au hasard (action m) ; une fois l'action choisie, la pièce est lancée et rapporte une récompense $r = 1$ si le résultat est pile, $r = 0$ sinon. Que vaut $p(s', r|s, a)$ dans les cas suivants :

- $(s', r, s, a) = (e, 1, e, g)$?
- $(s', r, s, a) = (d, 1, d, g)$?
- $(s', r, s, a) = (d, 0, d, g)$?
- $(s', r, s, a) = (d, 1, e, g)$?
- $(s', r, s, a) = (d, 1, d, m)$?
- $(s', r, s, a) = (e, 0, d, m)$?

Combien de cas faut-il énumérer pour caractériser le MDP ?

Processus de décision de Markov fini

Exemple

Exemple

On définit un MDP fini dans lequel un environnement est composé de quatre pièces équilibrées marquées d'une lettre e et une pièce déséquilibrée portant la lettre d,dont la probabilité d'obtenir pile est notée α ; une des pièces est sélectionnée à chaque instant, ce qui constitue l'état ; un agent peut choisir entre garder la pièce courante (action g) ou mélanger les cinq pièces et en tirer une au hasard (action m) ; une fois l'action choisie, la pièce est lancée et rapporte une récompense $r = 1$ si le résultat est pile, $r = 0$ sinon. Que vaut $p(s', r|s, a)$ dans les cas suivants :

- $(s', r, s, a) = (e, 1, e, g)$? $p(e, 1|e, g) = 0.5$
- $(s', r, s, a) = (d, 1, d, g)$? $p(d, 1|d, g) = \alpha$
- $(s', r, s, a) = (d, 0, d, g)$? $p(d, 0|d, g) = 1 - \alpha$
- $(s', r, s, a) = (d, 1, e, g)$? $p(d, 1|e, g) = 0$
- $(s', r, s, a) = (d, 1, d, m)$? $p(d, 1|d, m) = 0.2\alpha$
- $(s', r, s, a) = (e, 0, d, m)$? $p(e, 0|d, m) = 0.4$

Combien de cas faut-il énumérer pour caractériser le MDP ? $S^2RA = 16$

Remarques sur $p(s', r|s, a)$

- Plusieurs conventions sont possibles pour le passage de t à $t+1$. Dans la définition choisie

$$p(s', r|s, a) = p(S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a),$$

le passage de t à $t+1$ se fait entre le choix de l'action et la réception de la récompense. On observe donc un enchainement

$$S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, \ldots$$

2

- $p(s', r|s, a)$ est une simple fonction

$$p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$$
$$(s', r, s, a) \mapsto p(s', r|s, a)$$

- Dans le cas présent d'un MDP fini, on peut donc stocker cette fonction comme un tableau à quatre dimensions $[S, R, S, A]$ dont chaque élément est un réel de $[0, 1]$.
- Pour tout $s \in \mathcal{S}$ et $a \in \mathcal{A}$, on a

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

2. Autre convention possible : $S_0, A_0, R_0, S_1, A_1, R_1, \ldots, S_t, A_t, R_t, S_{t+1}, A_{t+1}, \ldots$

Remarques sur $p(s', r|s, a)$

- Plusieurs conventions sont possibles pour le passage de t à $t+1$. Dans la définition choisie

$$p(s', r|s, a) = p(S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a),$$

le passage de t à $t+1$ se fait entre le choix de l'action et la réception de la récompense. On observe donc un enchainement

$$S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, \ldots$$

[2]

- $p(s', r|s, a)$ est une simple fonction

$$p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$$
$$(s', r, s, a) \mapsto p(s', r|s, a)$$

- Dans le cas présent d'un MDP fini, on peut donc stocker cette fonction comme un tableau à quatre dimensions $[S, R, S, A]$ dont chaque élément est un réel de $[0, 1]$.
- Pour tout $s \in \mathcal{S}$ et $a \in \mathcal{A}$, on a

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) = 1.$$

2. Autre convention possible : $S_0, A_0, R_0, S_1, A_1, R_1, \ldots, S_t, A_t, R_t, S_{t+1}, A_{t+1}, \ldots$

Autres quantités utiles

$p(s', r|s, a)$ caractérise entièrement le MDP et on peut en déduire les quantités suivantes

- les probabilités de transition entre états : la probabilité de passer d'un état $s \in \mathcal{S}$ à un état $s' \in \mathcal{S}$ via une action $a \in \mathcal{A}$ est

$$p(s'|s, a) = ?$$

- l'espérance de la récompense immédiate lorsque l'on choisit une action $a \in \mathcal{A}$ dans un état $s \in \mathcal{S}$ est

$$r(s, a) = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = ?$$

- l'espérance de la récompense immédiate lorsque l'on choisit une action $a \in \mathcal{A}$ dans un état $s \in \mathcal{S}$ et que l'on se retrouve dans un état $s' \in \mathcal{S}$ est

$$r(s, a, s') = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'] = ?$$

Autres quantités utiles

$p(s', r|s, a)$ caractérise entièrement le MDP et on peut en déduire les quantités suivantes

- les probabilités de transition entre états : la probabilité de passer d'un état $s \in \mathcal{S}$ à un état $s' \in \mathcal{S}$ via une action $a \in \mathcal{A}$ est

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

- l'espérance de la récompense immédiate lorsque l'on choisit une action $a \in \mathcal{A}$ dans un état $s \in \mathcal{S}$ est

$$r(s, a) = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = ?$$

- l'espérance de la récompense immédiate lorsque l'on choisit une action $a \in \mathcal{A}$ dans un état $s \in \mathcal{S}$ et que l'on se retrouve dans un état $s' \in \mathcal{S}$ est

$$r(s, a, s') = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'] = ?$$

Autres quantités utiles

$p(s', r|s, a)$ caractérise entièrement le MDP et on peut en déduire les quantités suivantes

- les probabilités de transition entre états : la probabilité de passer d'un état $s \in \mathcal{S}$ à un état $s' \in \mathcal{S}$ via une action $a \in \mathcal{A}$ est

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

- l'espérance de la récompense immédiate lorsque l'on choisit une action $a \in \mathcal{A}$ dans un état $s \in \mathcal{S}$ est

$$r(s, a) = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r|s, a)$$

- l'espérance de la récompense immédiate lorsque l'on choisit une action $a \in \mathcal{A}$ dans un état $s \in \mathcal{S}$ et que l'on se retrouve dans un état $s' \in \mathcal{S}$ est

$$r(s, a, s') = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'] = ?$$

Autres quantités utiles

$p(s', r|s, a)$ caractérise entièrement le MDP et on peut en déduire les quantités suivantes

- les probabilités de transition entre états : la probabilité de passer d'un état $s \in \mathcal{S}$ à un état $s' \in \mathcal{S}$ via une action $a \in \mathcal{A}$ est

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

- l'espérance de la récompense immédiate lorsque l'on choisit une action $a \in \mathcal{A}$ dans un état $s \in \mathcal{S}$ est

$$r(s, a) = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r|s, a)$$

- l'espérance de la récompense immédiate lorsque l'on choisit une action $a \in \mathcal{A}$ dans un état $s \in \mathcal{S}$ et que l'on se retrouve dans un état $s' \in \mathcal{S}$ est

$$r(s, a, s') = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathcal{R}} r \frac{p(s', r|s, a)}{p(s'|s, a)}$$

Exemples applicatifs de MDP

Bio-réacteur

On souhaite fabriquer des produits issus d'une réaction bio-chimique :

- environnement : des capteurs mesurent la quantité de bactéries, de nutriments, de produit fabriqué, et d'autres données, éventuellement avec une certaine latence ;
- actions : à chaque instant, l'agent peut agir à la fois sur un système de chauffage et sur un moteur en fixant une température cible et une vitesse cible de mélange de la réaction ;
- récompense immédiate : elle est donnée par la quantité de produit obtenu à chaque instant.

Exemples applicatifs de MDP

Bras articulé

Un bras articulé doit déplacer des objets d'un point à un autre de façon répétée :

- actions : à chaque instant, l'agent choisit la tension à appliquer à chaque moteur du bras ;
- environnement : des capteurs mesurent la vitesse et l'angle de chaque articulation du bras ;
- récompense immédiate : elle est de $+1$ lorsqu'un objet est déposé à l'endroit souhaité et est négative à chaque à-coup du bras afin d'obtenir un mouvement fluite.

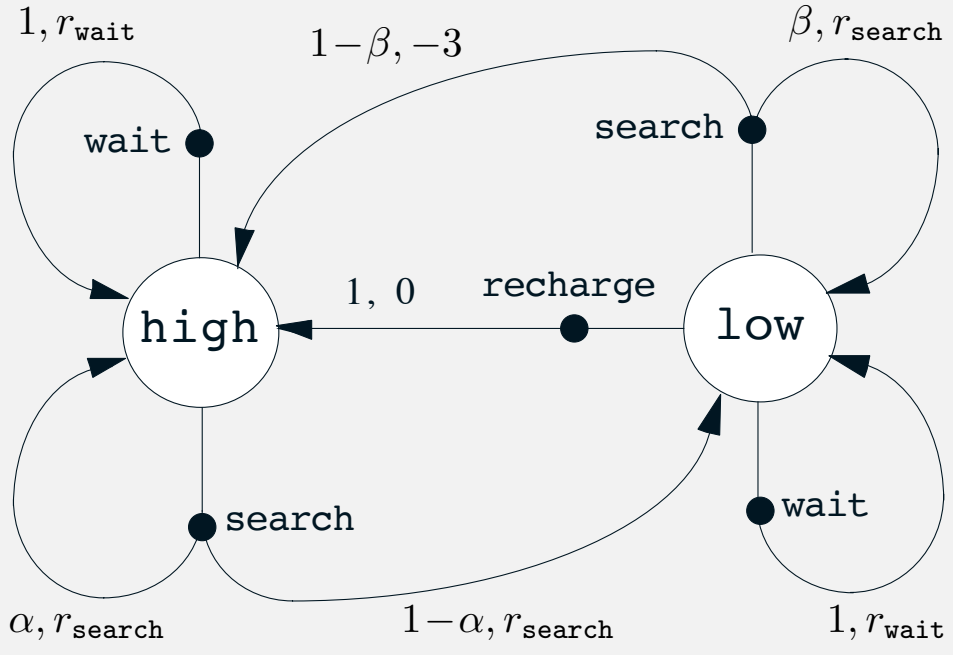
Example 3.3 Recycling Robot

A mobile robot has the job of collecting empty soda cans in an office environment. It has sensors for detecting cans, and an arm and gripper that can pick them up and place them in an onboard bin; it runs on a rechargeable battery. The robot's control system has components for interpreting sensory information, for navigating, and for controlling the arm and gripper. High-level decisions about how to search for cans are made by a reinforcement learning agent based on the current charge level of the battery. To make a simple example, we assume that only two charge levels can be distinguished, comprising a small state set $\mathcal{S} = \{\texttt{high}, \texttt{low}\}$. In each state, the agent can decide whether to (1) actively `search` for a can for a certain period of time, (2) remain stationary and `wait` for someone to bring it a can, or (3) head back to its home base to `recharge` its battery. When the energy level is `high`, recharging would always be foolish, so we do not include it in the action set for this state. The action sets are then $\mathcal{A}(\texttt{high}) = \{\texttt{search}, \texttt{wait}\}$ and $\mathcal{A}(\texttt{low}) = \{\texttt{search}, \texttt{wait}, \texttt{recharge}\}$.

The rewards are zero most of the time, but become positive when the robot secures an empty can, or large and negative if the battery runs all the way down. The best way to find cans is to actively search for them, but this runs down the robot's battery, whereas waiting does not. Whenever the robot is searching, the possibility exists that its battery will become depleted. In this case the robot must shut down and wait to be rescued (producing a low reward). If the energy level is `high`, then a period of active search can always be completed without risk of depleting the battery. A period of searching that begins with a `high` energy level leaves the energy level `high` with probability α and reduces it to `low` with probability $1 - \alpha$. On the other hand, a period of searching undertaken when the energy level is `low` leaves it `low` with probability β and depletes the battery with probability $1 - \beta$. In the latter case, the robot must be rescued, and the battery is then recharged back to `high`. Each can collected by the robot counts as a unit reward, whereas a reward of -3 results whenever the robot has to be rescued. Let $r_{\texttt{search}}$ and $r_{\texttt{wait}}$, with $r_{\texttt{search}} > r_{\texttt{wait}}$, respectively denote the expected number of cans the robot will collect (and hence the expected reward) while searching and while waiting. Finally, suppose that no cans can be collected during a run home for recharging, and that no cans can be collected on a step in which the battery is depleted. This system is then a finite MDP, and we can write down the transition probabilities and the expected rewards, with dynamics as indicated in the table on the left:

s	a	s'	$p(s' \mid s, a)$	$r(s, a, s')$
high	search	high	α	$r_{\texttt{search}}$
high	search	low	$1 - \alpha$	$r_{\texttt{search}}$
low	search	high	$1 - \beta$	-3
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
high	wait	low	0	-
low	wait	high	0	-
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0
low	recharge	low	0	-

Note that there is a row in the table for each possible combination of current state, s, action, $a \in \mathcal{A}(s)$, and next state, s'. Some transitions have zero probability of occurring, so no expected reward is specified for them. Shown on the right is another useful way of summarizing the dynamics of a finite MDP, as a *transition graph*. There are two kinds of nodes: *state nodes* and *action nodes*. There is a state node for each possible state (a large open circle labeled by the name of the state), and an action node for each state–action pair (a small solid circle labeled by the name of the action and connected by a line to the state node). Starting in state s and taking action a moves you along the line from state node s to action node (s, a). Then the environment responds with a transition to the next state's node via one of the arrows leaving action node (s, a). Each arrow corresponds to a triple (s, s', a), where s' is the next state, and we label the arrow with the transition probability, $p(s' \mid s, a)$, and the expected reward for that transition, $r(s, a, s')$. Note that the transition probabilities labeling the arrows leaving an action node always sum to 1.



Exercise 3.4 Give a table analogous to that in Example 3.3, but for $p(s', r \mid s, a)$. It should have columns for s, a, s', r, and $p(s', r \mid s, a)$, and a row for every 4-tuple for which $p(s', r \mid s, a) > 0$. □

Remarques sur les exemples

Comparez les exemples en faisant attention au points suivants :

- les instants successifs où l'état est observé et une action choisie peuvent être réguliers ou irréguliers ;
- les actions peuvent être de bas niveau ou de haut niveau ; elles peuvent être de natures très différentes, souvent sous forme vectorielle ;
- les états observés peuvent être également de natures très différentes, souvent sous forme vectorielle ;
- les récompenses sont des scalaires ;
- la limite entre l'environnement et l'agent n'est pas forcément une limite physique (par exemple entre un robot et le monde qui l'entoure) ;
- dans un problème d'AR, l'agent peut connaitre le fonctionnement de l'environnement ou du système de récompense, mais il n'est pas en mesure de bien les contrôler.

Exercices

Exercice : gain et états avec politique aléatoire

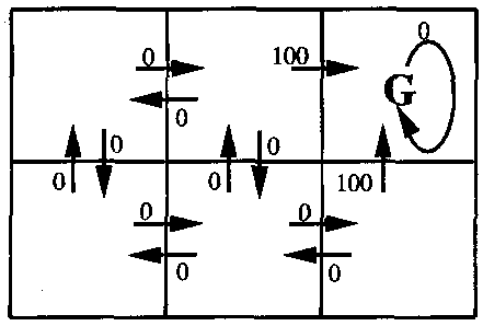
Si l'état courant est $S_t = s$ et si les actions sont sélectionnées au hasard par l'agent,

1. quelle est l'expression de l'espérance de la récompense R_{t+1} en fonction des probabilités $p(s', r|s, a)$?
2. quelle est la probabilité de se retrouver dans un état $S_{t+1} = s'$ à l'instant $t + 1$, en fonction des probabilités $p(s', r|s, a)$?

Exercices

Exercice : modèle déterministe

1. Reprenez l'exemple simple de la grille 2×3 du livre de Mitchell et définissez $p(s', r|s, a)$.



2. Dans le cas d'un environnement déterministe caractérisé par une fonction de transition $\delta(s, a)$ qui renvoie l'état s' obtenu en effectuant l'action a dans l'état s et une fonction de récompense $r(s, a)$ qui renvoie la récompense associée, que vaut $p(s', r|s, a)$?

Plan

Récompense et retour

L'agent reçoit des **récompenses (immédiates)** R_t, R_{t+1}, R_{t+2}. Son objectif est de maximiser un **retour**, ou **récompense à long terme**, noté G_t que l'on peut définir de plusieurs façons :

- sur un horizon fini $T < +\infty$, par exemple pour le cas d'épisodes :

$$G_t = R_{t+1} + R_{t+2} + \ldots + R_T = \sum_{k=t+1}^{T} R_k$$

- avec une dévaluation $\gamma \in [0, 1[$, par exemple dans le cas perpétuel :

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=t+1}^{+\infty} \gamma^{k-t-1} R_k$$

Remarque :

- on peut unifier les deux cas en notant $G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ où l'on peut avoir $T = +\infty$ ou $\gamma = 1$ (mais pas simultanément).
- on a $G_t = R_{t+1} + \gamma G_{t+1}$ pour $t < T$ (et l'on peut poser $G_T = 0$).

Objectif vs. récompense

En AR, l'objectif à atteindre se traduit par la maximisation de l'espérance des récompenses cumulées :

- on se ramène donc à une quantité scalaire : cela parait simple, mais offre une flexibilité qui permet de couvrir un grand nombre de situations.
- bien définir le système de récompense pour qu'il corresponde à l'objectif à atteindre.
- penser le système de récompense pour indiquer *quoi* atteindre, et non *comment* l'atteindre.

Exercices

Labyrinthe

On veut qu'un robot apprenne à sortir d'un labyrinthe. On décide de lui donner une récompense $+1$ lorsqu'il franchit la sortie et 0 le reste du temps. Il est naturel de procéder par épisodes, et on décide donc de maximiser le retour total $G_t = \sum_{k=1}^{T} R_{t+k}$. Après avoir fait tourner un algorithme d'apprentissage, vous constatez que l'agent n'a rien appris. Pourquoi et comment faire ?

Exercices

Calcul de G_t, cas fini.

On fixe $\gamma = 0.5$ et $T = 5$. L'agent reçoit les récompenses suivantes :

$$R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, R_5 = 2.$$

Calculez $G_0, \ldots, G_5$.

Calcul de G_t, cas infini

On fixe $\gamma = 0.9$. L'agent reçoit la récompense $R_1 = 2$ puis une suite infinie $R_t = 7$ pour $t > 1$. Calculez G_0 et G_1.

Plan

Politique : modéliser un agent

- Une politique permet à un agent de choisir une action a à partir d'un état s.
- Avec un formalisme probabiliste, on définit une politique sous la forme d'une fonction

$$\pi(a|s) = \mathbb{P}(A_t = a|S_t = s)$$

que l'on peut donc stocker dans une matrice $A \times S$ et qui régit la loi de probabilité de $A_t|S_t$

- Choisir une action dans un état s consiste à faire une tirage selon cette loi $a \mapsto \pi(a|s)$

Interaction agent/environnement

Étant données

- une loi $p(s', r|s, a)$ caractérisant l'environnement,
- et une loi $\pi(a|s)$ caractérisant la politique de l'agent,

l'interaction entre l'agent et l'environnement consiste à :

- choisir un état initial : $s \leftarrow s_0$
- répéter pour $t = 0, 1, \ldots$
  - choisir une action : tirer $A_t = a$ selon $a \mapsto \pi(a|s)$
  - obtenir la récompense et le nouvel état : tirer $S_{t+1} = s', R_{t+1} = r$ selon $(s', r) \mapsto p(s', r|s, a)$
  - mettre à jour l'état courant $s \leftarrow s'$

Fonctions d'évaluation des états et des actions

Objectif : pour une politique π, évaluer dans quelle mesure il est intéressant d'être dans d'un état ou de choisir une action dans un état donné, au sens de l'espérance du retour obtenu à partir de cet état ou de cette action dans cet état.

Fonction d'évaluation d'un état pour une politique π

$$\forall s \in \mathcal{S}, v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{+\infty} \gamma^k R_{t+k+1} | S_t = s\right]$$

Fonction d'évaluation d'une action pour une politique π

$$\begin{aligned}\forall (s,a) \in \mathcal{S} \times \mathcal{A}, q_\pi(s,a) &= \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{+\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a\right]\end{aligned}$$

$\rightarrow$ Ces fonctions peuvent être estimées à partir des données obtenues en faisant interagir l'agent et l'environnement.

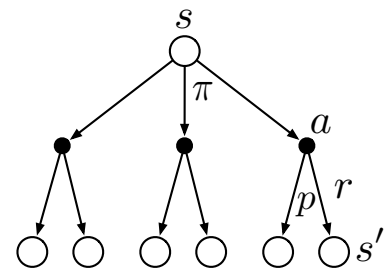
Équations de Bellman

Les fonctions d'évaluation sont définies à partir des récompenses reçues au cours de l'ensemble des itérations depuis un état $S_t = s$ $(t+1, t+2, \ldots)$. Les équations de Bellman établissent une relation récursive faisant intervenir les données d'une seule itération $(t \to t+1)$. Elles sont à l'origine des formules de rétropropagation permettant d'estimer v et q, et d'apprendre π.

Équation de Bellman pour v_π

Pour $s \in \mathcal{S}$,

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

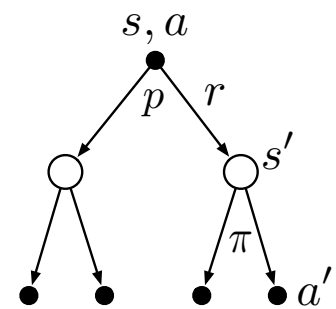


Backup diagram for v_π

Équation de Bellman pour q_π

Pour $(s, a) \in \mathcal{S} \times \mathcal{A}$,

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



q_π backup diagram

Démonstration des équations de Bellman

Preuve au tableau pour q_π.
Preuve à faire en exercice pour v_π

Example 3.5: Gridworld Figure 3.2 (left) shows a rectangular gridworld representation of a simple finite MDP. The cells of the grid correspond to the states of the environment. At each cell, four actions are possible: `north`, `south`, `east`, and `west`, which deterministically cause the agent to move one cell in the respective direction on the grid. Actions that would take the agent off the grid leave its location unchanged, but also result in a reward of -1. Other actions result in a reward of 0, except those that move the agent out of the special states A and B. From state A, all four actions yield a reward of $+10$ and take the agent to A′. From state B, all actions yield a reward of $+5$ and take the agent to B′.

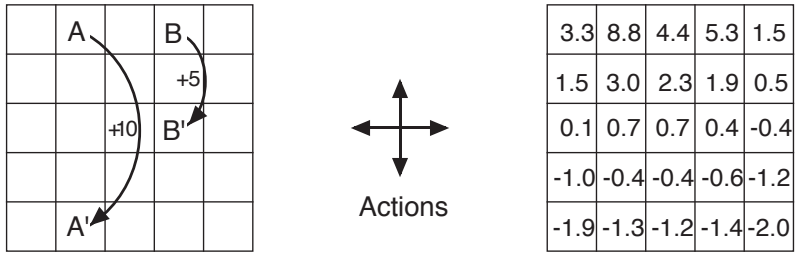


Figure 3.2: Gridworld example: exceptional reward dynamics (left) and state-value function for the equiprobable random policy (right).

Suppose the agent selects all four actions with equal probability in all states. Figure 3.2 (right) shows the value function, v_π, for this policy, for the discounted reward case with $\gamma = 0.9$. This value function was computed by solving the system of linear equations (3.14). Notice the negative values near the lower edge; these are the result of the high probability of hitting the edge of the grid there under the random policy. State A is the best state to be in under this policy, but its expected return is less than 10, its immediate reward, because from A the agent is taken to A′, from which it is likely to run into the edge of the grid. State B, on the other hand, is valued more than 5, its immediate reward, because from B the agent is taken to B′, which has a positive value. From B′ the expected penalty (negative reward) for possibly running into an edge is more than compensated for by the expected gain for possibly stumbling onto A or B. ■

Exercise 3.14 The Bellman equation (3.14) must hold for each state for the value function v_π shown in Figure 3.2 (right) of Example 3.5. Show numerically that this equation holds for the center state, valued at $+0.7$, with respect to its four neighboring states, valued at $+2.3$, $+0.4$, -0.4, and $+0.7$. (These numbers are accurate only to one decimal place.) □

Plan

Fonctions d'évaluations optimales

Définitions

Fonction d'évaluation optimale des états

$$\forall s \in \mathcal{S}, v_*(s) \triangleq \max_\pi v_\pi(s)$$

Fonction d'évaluation optimale des actions

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

Propriété : relation entre q_* et v_*

On a

$$\forall s \in \mathcal{S}, v_*(s,a) = \max_a q_*(s,a)$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma v_*\left(S_{t+1}\right) | S_t = s, A_t = a\right]$$

Remarque : à ce stade, seules les fonctions d'évaluation optimales sont définies, les politiques optimales ne sont pas (encore) définies.

Politiques optimales

Ordre partiel entre les politiques

On définit un ordre partiel entre les politiques : les politiques π et π' sont telles que $\pi \geq \pi'$ si et seulement si $\forall s \in \mathcal{S}, \pi(s) \geq \pi'(s)$.

Question : qu'est-ce qu'un ordre partiel et pourquoi en est-ce un ?

Existence de politiques optimales (admis)

Il y a toujours au moins une politique π_* qui est supérieure à toutes les politiques, c-à-d que $\forall \pi, \pi_* \geq \pi$. On les appelle *politiques optimales*. Leurs fonctions d'évaluation sont les mêmes, ce sont les fonctions d'évaluation optimales : $v_{\pi_*} = v_*$ et $q_{\pi_*} = q_*$.

Équations d'optimalité de Bellman

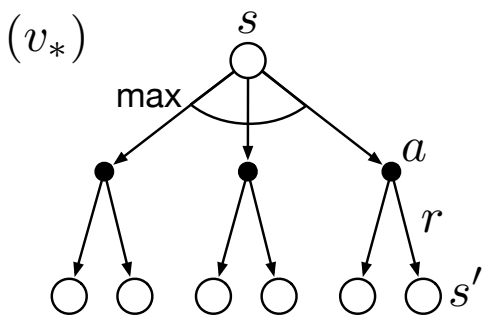
→ Comportement local des fonctions d'évaluation.

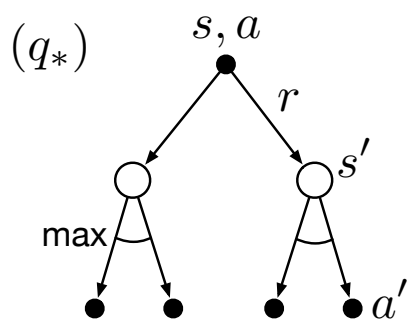
Les fonctions d'évaluation optimales vérifient les équations de Bellman suivantes : pour $s \in \mathcal{S}$,

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

et pour $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$q_*(s, a) = \sum_{s',r} p(s', r|s, a) \left[r + \gamma \max_{a'} q_*(s', a')\right].$$





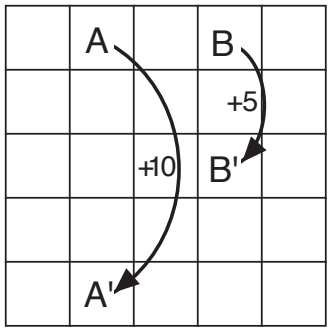
Preuves à faire

Remarques sur les solutions optimales

- Les équations d'optimalité de Bellman sont non-linéaires
- Pour un MDP fini, il existe une unique solution pour v_* : on peut l'obtenir si on connait l'environnement.
- À partir de v_*, on peut trouver π_* : à partir de s, choisir une action a pour laquelle le max est atteint dans l'équation d'optimalité : $a \in \text{argmax}_a \sum_{s',r} p(s', r|s, a) [r + \gamma v_*(s')]$; toute politique affectant des probabilités non-nulles uniquement à ces actions est optimale. $\rightarrow$ politique greedy par rapport à v_*.
- À partir de q_*, on peut trouver π_* facilement : $\pi_*(a|s) \geq 0$ seulement si $a \in \text{argmax}_{a'} q_*(s, a')$.
- Mais résoudre les équations de Bellman est en général trop coûteux (nombre d'états/actions élevé) + nécessité de connaitre l'environnement. $\rightarrow$ on essaie de trouver des solutions approximatives qui s'en rapprochent et s'en inspirent.

Exercice

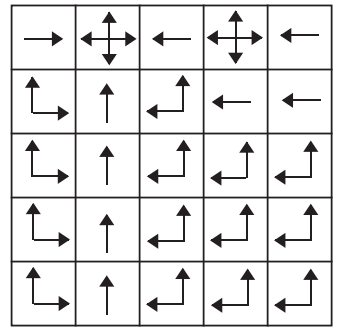
La solution optimale de l'exemple de la grille est donnée ci-dessous. On voit que le maximum de v_* vaut 24.4.
Vérifiez la cohérence numériquement de cette valeur en utilisant les résultats qui précèdent.



Gridworld

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

v_*



π_*

Example 3.9: Bellman Optimality Equations for the Recycling Robot Using (3.19), we can explicitly give the Bellman optimality equation for the recycling robot example. To make things more compact, we abbreviate the states `high` and `low`, and the actions `search`, `wait`, and `recharge` respectively by `h`, `l`, `s`, `w`, and `re`. Because there are only two states, the Bellman optimality equation consists of two equations. The equation for $v_*(\texttt{h})$ can be written as follows:

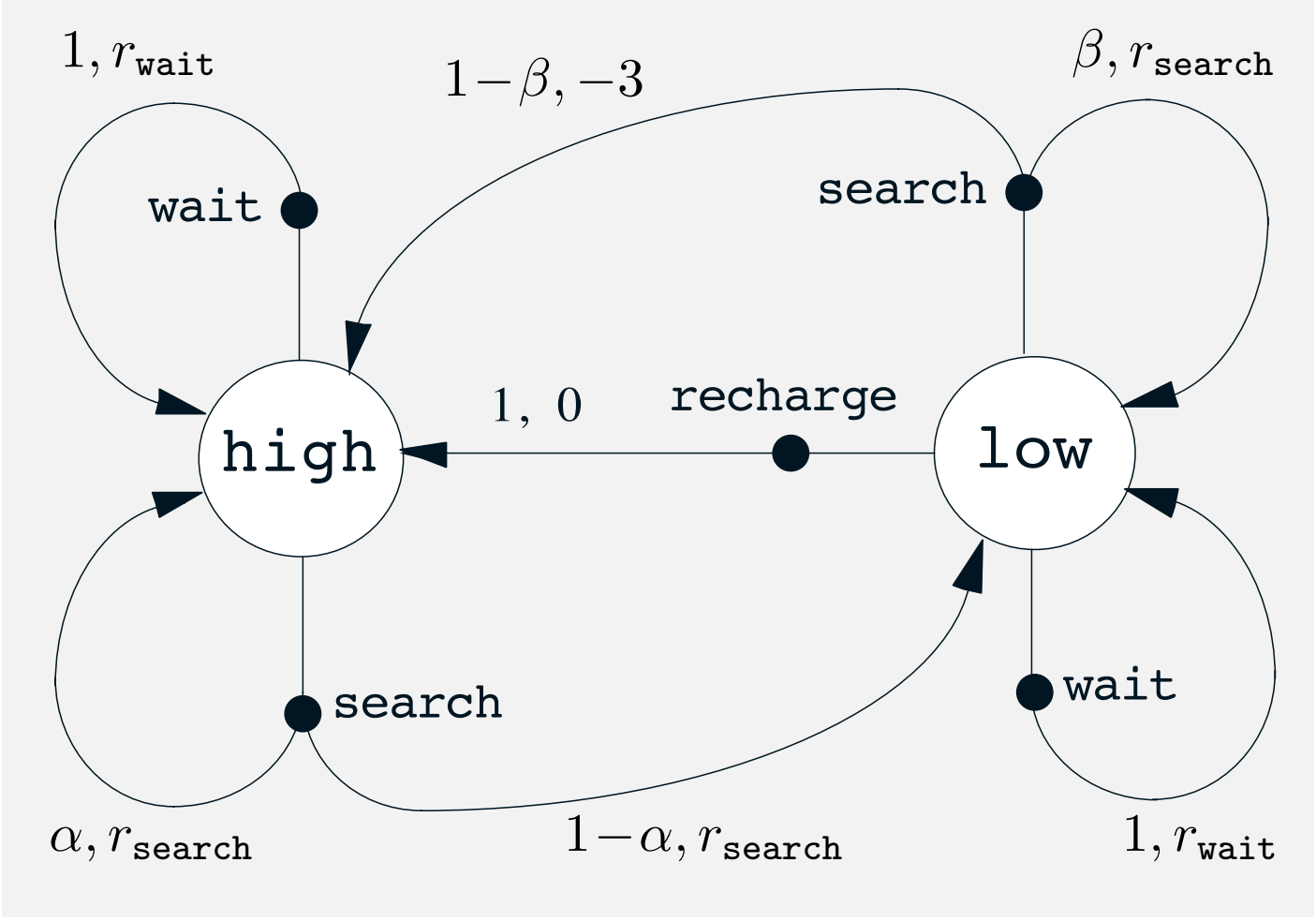
$$\begin{aligned} v_*(\texttt{h}) &= \max\left\{ \begin{array}{l} p(\texttt{h}|\texttt{h},\texttt{s})[r(\texttt{h},\texttt{s},\texttt{h}) + \gamma v_*(\texttt{h})] + p(\texttt{l}|\texttt{h},\texttt{s})[r(\texttt{h},\texttt{s},\texttt{l}) + \gamma v_*(\texttt{l})], \\ p(\texttt{h}|\texttt{h},\texttt{w})[r(\texttt{h},\texttt{w},\texttt{h}) + \gamma v_*(\texttt{h})] + p(\texttt{l}|\texttt{h},\texttt{w})[r(\texttt{h},\texttt{w},\texttt{l}) + \gamma v_*(\texttt{l})] \end{array} \right\} \\ &= \max\left\{ \begin{array}{l} \alpha[r_\texttt{s} + \gamma v_*(\texttt{h})] + (1-\alpha)[r_\texttt{s} + \gamma v_*(\texttt{l})], \\ 1[r_\texttt{w} + \gamma v_*(\texttt{h})] + 0[r_\texttt{w} + \gamma v_*(\texttt{l})] \end{array} \right\} \\ &= \max\left\{ \begin{array}{l} r_\texttt{s} + \gamma[\alpha v_*(\texttt{h}) + (1-\alpha) v_*(\texttt{l})], \\ r_\texttt{w} + \gamma v_*(\texttt{h}) \end{array} \right\}. \end{aligned}$$

Following the same procedure for $v_*(\texttt{l})$ yields the equation

$$v_*(\texttt{l}) = \max\left\{ \begin{array}{l} \beta r_\texttt{s} - 3(1-\beta) + \gamma[(1-\beta) v_*(\texttt{h}) + \beta v_*(\texttt{l})], \\ r_\texttt{w} + \gamma v_*(\texttt{l}), \\ \gamma v_*(\texttt{h}) \end{array} \right\}.$$

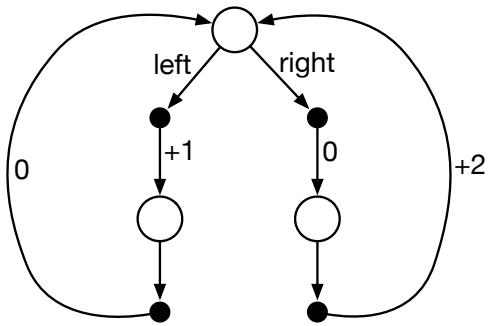
For any choice of $r_\texttt{s}$, $r_\texttt{w}$, α, β, and γ, with $0 \leq \gamma < 1$, $0 \leq \alpha, \beta \leq 1$, there is exactly one pair of numbers, $v_*(\texttt{h})$ and $v_*(\texttt{l})$, that simultaneously satisfy these two nonlinear equations. ■

Écrire les équations de Bellman pour q_*.



Écrire les équations de Bellman pour q_*.

Exercise 3.22 Consider the continuing MDP shown on to the right. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$? □



Optimialité et approximation

- Les résultats d'optimalité sont importants d'un point de vue théorique et comme idéal à atteindre.
- On peut difficilement obtenir une politique optimale en pratique.
- Même avec la connaissance de l'environnement, la complexité temporelle est trop importante en général.
- De même, complexité spatiale trop importante pour les méthodes tabulaires ($\equiv$ stockage des fonctions dans des tableaux)
  $\rightarrow$ utiliser des approximations avec des fonctions paramétriques plus compactes à manipuler.
- Une grande contribution de l'AR est de proposer des méthodes d'approximation pour approcher les politiques optimales.

Plan

Conclusions (1/2)

- comment modéliser des problèmes réels ?

  Les MDP permettent de modéliser de nombreuses situations pratiques.
  Les choix à faire ne sont pas triviaux : identifier environnement et agent, définir les états, actions, récompenses

- comment formaliser les aspects aléatoires (environnement et agent) ?

  Les MDP permettent de modéliser des environnements et des politiques aléatoires ou déterministes !

- comment formuler l'objectif du problème ?

  Notion de retour (gain à long terme)

- comment analyser les stratégies ?

  Les fonctions d'évaluations donnent la récompense à long terme pour une politique en fonction des états ou des couples (état, action).
  Les équations de Bellman permettent de caractériser ces fonctions d'évaluation.

- comment caractériser les stratégies optimales ?

  Équations d'optimalité de Bellman

Conclusions (2/2)

- Dans le cas fini, on peut se rapporter à des tableaux (qui atteignent vite les limites physiques des ordinateurs) pour les modèles d'environnement, les politiques, les fonctions d'évaluations.
- Comment apprendre des stratégies ?

  → *Les propriétés importantes des MDP (équations de Bellman, résultats d'optimalité) permettent d'analyser et résoudre ces problèmes, y compris approximativement.*

  → *Autant de raisons de passer du temps sur la modélisation des problèmes avant de les résoudre.*

  → *Algos de résolution : prochaine séance sur TD-learning.*

Joyeuses fêtes et bonnes vacances !