Chapitre 7: Optimisation par apprentissage

INF889B — Algorithmes d'optimisation combinatoire

Alexandre Blondin Massé

Université du Québec à Montréal

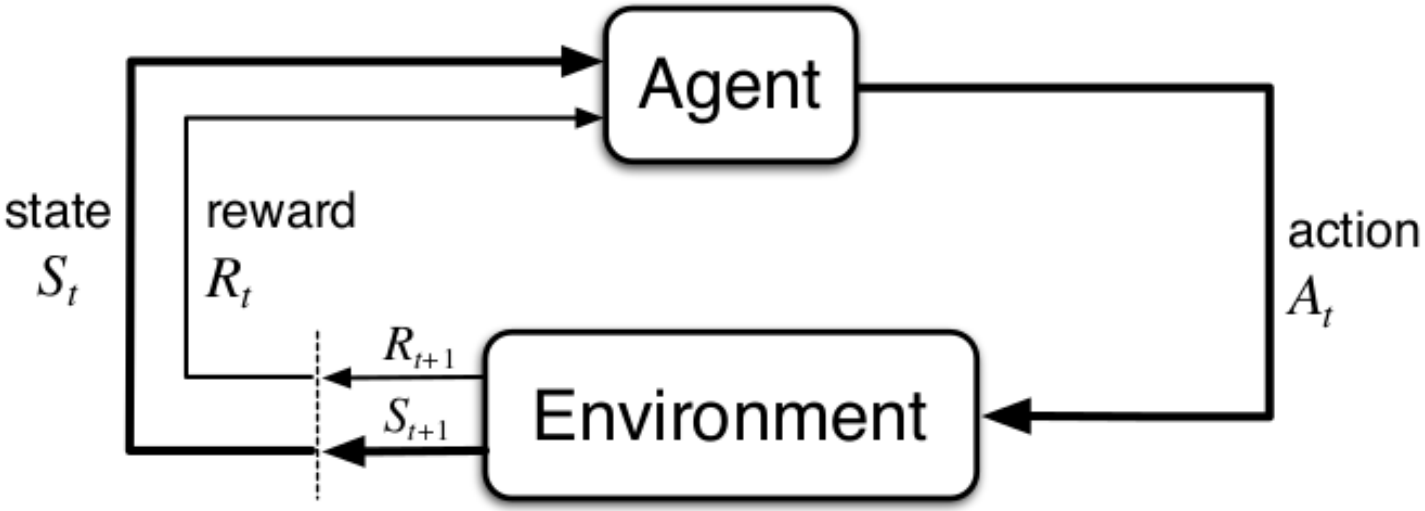
Hiver 2020

Partie d: Processus de décision markoviens

Plan

Définitions et notation

Interaction



(Source: Sutton et Barto, chapitre 3)

- t: **étape** ou **moment**
- $\mathcal{S}$: ensemble de tous les **états** possibles
- $\mathcal{R} \subseteq \mathbb{R}$: ensemble de toutes les **récompenses** possibles
- $\mathcal{A}$: ensemble de toutes les **actions** possibles
- On se concentre sur le cas **fini**: $|\mathcal{S}|, |\mathcal{R}|, |\mathcal{A}| < \infty$

Définition

Un **processus de décision markovien** fini est la donnée

- d'un ensemble fini d'**états** $\mathcal{S}$
- d'un ensemble fini d'**actions** $\mathcal{A}$
- d'un ensemble fini de **récompenses** $\mathcal{R}$
- d'une **suite d'états** $\{S_t\}_{t\geq 0}$
- d'une **suite d'actions** $\{A_t\}_{t\geq 0}$
- d'une **suite de récompenses** $\{R_t\}_{t\geq 1}$
- d'une fonction de **dynamique**

$$p : (\mathcal{S} \times \mathcal{R}) \times (\mathcal{S} \times \mathcal{A}) \to [0, 1]$$

définie par

$$p(s', r \mid s, a) = \mathbb{P}(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a)$$

- ayant la **propriété de Markov**

Propriété de Markov

Pour tout t, S_t et R_t ne **dépendent** que de S_{t-1} et A_{t-1}:

$$\begin{aligned} p(s', r \mid s, a) &= \mathbb{P}(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a) \\ &= \mathbb{P}(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a, \\ &\qquad S_{t-2} = s_{t-2}, A_{t-2} = a_{t-2}, \\ &\qquad \dots, \\ &\qquad S_1 = s_1, A_1 = a_1) \end{aligned}$$

Ainsi,

- pour **calculer** la probabilité de se trouver à S_t et d'avoir une récompense R_t
- nous n'avons **pas besoin** de connaître les états S_{t-2}, S_{t-3}, ... ni les récompenses R_{t-2}, R_{t-3}, ...

Remarques et notation

- Pour tous $s \in \mathcal{S}$ et $a \in \mathcal{A}$, on a

$$\sum_{s',r} p(s', r \mid s, a) = 1$$

- On **étend** la fonction de dynamique p:

$$p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$$

en posant

$$p(s' \mid s, a) = \mathbb{P}(S_t = s' \mid S_{t-1} = s, A_{t-1} = a) = \sum_r p(s', r \mid s, a)$$

Récompense espérée (1/2)

- On introduit une fonction indiquant la **récompense espérée**:

$$r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

définie par

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

- Par **définition** de l'espérance mathématique, on a donc

$$r(s, a) = \sum_r r \sum_s^{'} p(s', r \mid s, a)$$

Récompense espérée (2/2)

- La fonction r s'étend naturellement à

$$r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$$

en posant

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

- Par conséquent, on trouve

$$r(s, a, s') = \sum_r r \, \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

Exemples

Exemple 1: le robot recycleur

- Un robot qui se **déplace** dans une pièce
- De façon **autonome**
- **Unique tâche**: récupérer les canettes vides
- Il est alimenté par une **batterie rechargeable**
- On souhaite maximiser le nombre de canettes **récupérées**
- En minimisant le nombre de **recharges** manuelles et autonomes

États possibles

- `high`: la batterie est *chargée*
- `low`: la batterie est *faible*

Actions possibles

- `search`: le robot recherche une canette
- `wait`: le robot attend que quelqu'un lui donne une canette
- `recharge`: le robot recharge sa batterie de façon autonome

Paramètres

- α: probabilité de rester dans l'état `high`
- β: probabilité de rester dans l'état `low`
- $r_{\texttt{search}}$: la récompense espérée quand le robot cherche
- $r_{\texttt{wait}}$: la récompense espérée quand le robot attend
- $r_{\texttt{recharge}}$: récompense (ou pénalité) encourue lors d'une recharge manuelle

Contraintes

- $r_{\texttt{search}} > r_{\texttt{wait}} > 0$: plus intéressant de chercher que d'attendre
- $r_{\texttt{recharge}} < 0$: pénalisant de recharger la batterie manuellement

Conséquences

- $1 - \alpha$: probabilité de passer de `high` vers `low`
- $1 - \beta$: probabilité que la batterie se décharge

Graphe

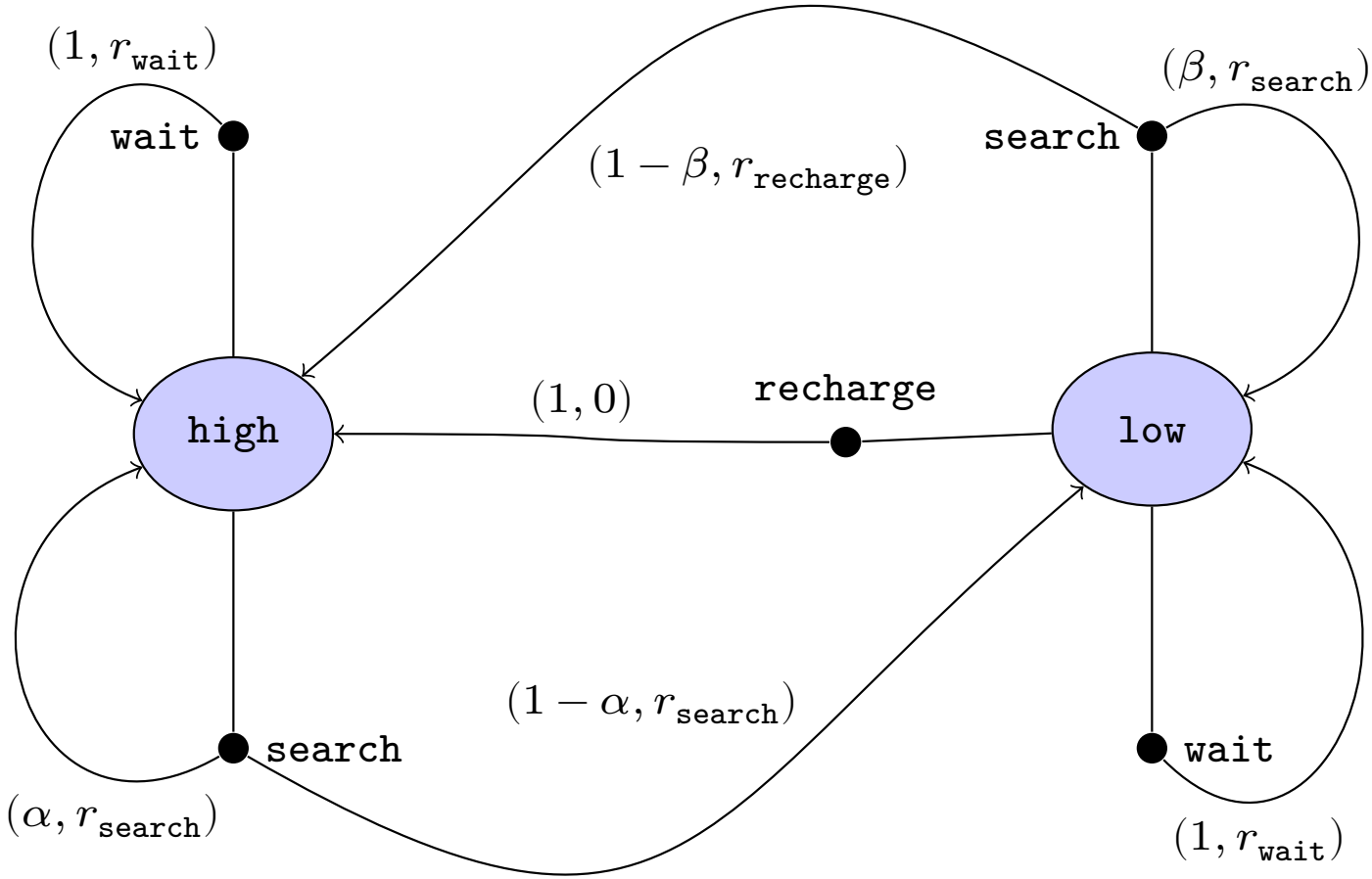


Image adaptée du livre de Sutton et Barto (chapitre 3)

Tableau de transition (complet)

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	$1-\beta$	$r_{\texttt{recharge}}$
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	$r_{\texttt{wait}}$
high	recharge	high	0	—
high	recharge	low	0	—
low	recharge	high	1	0
low	recharge	low	0	—

Tableau de transition (compact)

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	$1 - \beta$	$r_{\texttt{recharge}}$
low	search	low	β	$r_{\texttt{search}}$
high	wait	high	1	$r_{\texttt{wait}}$
low	wait	low	1	$r_{\texttt{wait}}$
low	recharge	high	1	0

Exemple 2: déplacement sur une grille

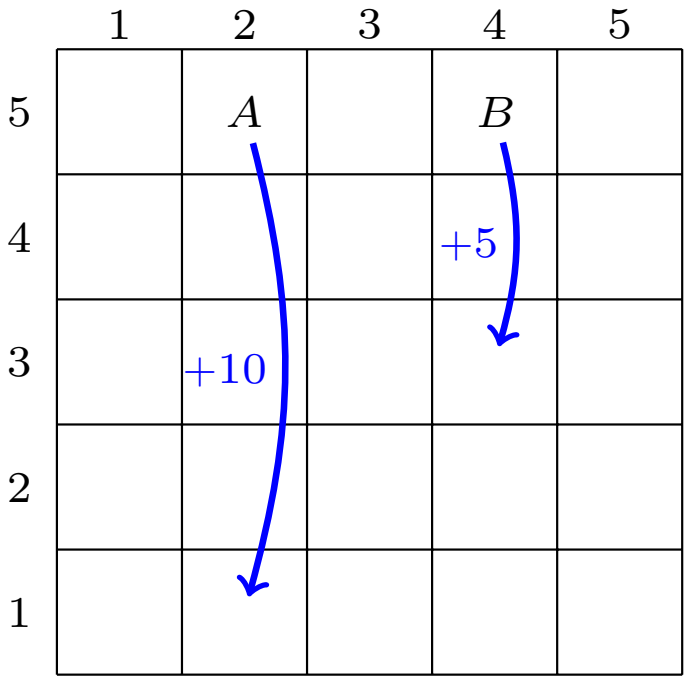


Image adaptée du livre de Sutton et Barto (chapitre 3)

États: chacune des 25 cases

Actions: `est`, `nord`, `ouest` et `sud`

- *Cas général*: change selon l'action
- *Hors grille*: ne change pas d'état
- *État spécial* A: au bout de la flèche
- *État spécial* B: au bout de la flèche

Récompenses:

- *Hors grille*: -1
- *État spécial* A: $+10$
- *État spécial* B: $+5$
- *Autrement*: 0

Retour espéré

Définition

Rappel

Suite de **récompenses**: R_1, R_2, ..., $R_{t_{\text{max}}}$

Retour (ou gain) espéré

Pour $t = 0, 1, \ldots, t_{\text{max}} - 1$, on définit le **retour espéré** par

$$G_t = R_{t+1} + R_{t+2} + \ldots + R_{t_{\text{max}}}$$

Si le processus est **continu** (épisode infini), c'est plutôt

$$G_t = R_{t+1} + R_{t+2} + \ldots$$

Interprétation

Récompense **totale** qu'on espère avoir à partir de maintenant

Processus continu

- Quand $t_{\text{max}} \rightarrow \infty$
- Le **retour espéré** $G_t = R_{t+1} + R_{t+2} + ...$ peut tendre vers ∞

Facteur d'actualisation

- On souhaite **éviter** cette complication
- On introduit un **paramètre** γ, avec $0 \leq \gamma \leq 1$
- Appelé **facteur d'actualisation** (en anglais, *discount*)
- **Analogie**: inflation
- Et on définit plutôt

$$\begin{aligned} G_t &= \sum_{k=0}^{t_{\text{max}}} \gamma^k R_{t+k+1} \\ &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + ... \end{aligned}$$

Facteur d'actualisation

- $\gamma = 0$: l'agent est « myope », ne se préoccupe que des récompenses immédiates
- $\gamma \to 1$: vision à long terme
- Pour assurer la **convergence** de

$$G_t = \sum_{k=0}^{t_{\text{max}}} \gamma^k R_{t+k+1}$$

on impose que $t_{\text{max}} \neq \infty$ **ou** $\gamma < 1$

Notation uniforme

- Le paramètre t_{max} peut être **fini** ou **infini**
- Pour uniformiser, dans le cas fini, on pose $R_t = 0$ pour $t \geq t_{\text{max}}$
- Et donc on peut poser, pour tout $t \geq 0$,

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

Expression récursive

- On remarque que

$$\begin{aligned} G_t &= \sum_{k \geq 0} \gamma^k R_{t+k+1} \\ &= R_{t+1} + \sum_{k \geq 1} \gamma^k R_{t+k+1} \\ &= R_{t+1} + \gamma \sum_{k \geq 0} \gamma^k R_{t+k+2} \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

- Il suffit donc de calculer le retour à partir de la **fin**

Exercices

(Tirés du livre de Sutton et Barto, chapitre 3)

Exercice 1

- Soit $\gamma = 0.5$ et supposons qu'on reçoive les récompenses suivantes, dans l'ordre:

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

  avec $t_{\text{max}} = 5$.
- Que vaut G_t pour $t = 0, 1, \ldots, 5$?

Exercice 2

- Soit $\gamma = 0.9$ et supposons que $R_1 = 2$, puis $R_t = 7$ pour tout $t \geq 2$
- Que valent G_0 et G_1?

Politique et valuation

Politique

- Une **politique** est une fonction

$$\pi : \mathcal{A} \times \mathcal{S} \to [0, 1]$$

définie par

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

pour tout t

- Noter que π ne **dépend pas** de t
- Autrement dit, π est **fixe** pour toute la durée de **épisode**
- L'apprentissage par renforcement vise à **améliorer** π

Valuation d'un état

- Soit π une **politique**
- On définit la **valuation** (en anglais, *state-value function*) d'un état par

$$v_\pi : \mathcal{S} \to \mathbb{R}$$

en posant

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

pour tout t, où $\mathbb{E}_\pi$ est l'espérance mathématique en supposant que l'agent **applique** la politique π

- Ainsi, pour tout $s \in \mathcal{S}$,

$$v_\pi(s) = \mathbb{E}_\pi \left[\sum_{k \geq 0} \gamma^k R_{t+k+1} \,\middle|\, S_t = s \right]$$

Valuation d'une paire action-état

- Soit π une **politique**
- De la même façon, on définit la **valuation** (en anglais, *action-value function*) d'une paire action-état par

$$q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

en posant

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

pour tout t

- On a donc, pour tout $s \in \mathcal{S}$,

$$q_\pi(s, a) = \mathbb{E}_\pi \left[\sum_{k \geq 0} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a \right]$$

Équations de Bellman (1/4)

- La **valuation** est définie par

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

- Aussi, on a vu que G_t satisfait la **récurrence**

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

- On aimerait donc trouver une expression **récursive** pour $v_\pi(s)$ en fonction des états **successeurs** de s
- Permettra de **calculer** récursivement v_π
- À partir des états **finaux** vers les états **précédents**

Équations de Bellman (2/4)

- On trouve

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[G_t \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} \mid S_t = s] + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_t = s] \end{aligned}$$

  par la **linéarité** de l'espérance mathématique
- On montre ensuite que

$$\mathbb{E}_\pi[R_{t+1} \mid S_t = s] = \sum_{r,a,s'} r \; \pi(a \mid s) \; p(s', r \mid s, a)$$

  et

$$\mathbb{E}_\pi[G_{t+1} \mid S_t = s] = \sum_{r,a,s'} \pi(a \mid s) \; p(s', r \mid s, a) \; \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s]$$

Équations de Bellman (3/4)

D'une part,

$$\begin{aligned}
\mathbb{E}_\pi[R_{t+1} \mid S_t = s] &= \sum_r r \, \mathbb{P}_\pi(R_{t+1} = r \mid S_t = s) \\
&= \sum_r r \sum_{a,s'} \pi(a \mid s) \; p(s', r \mid s, a) \\
&= \sum_{r,a,s'} r \; \pi(a \mid s) \; p(s', r \mid s, a)
\end{aligned}$$

D'autre part, on peut démontrer avec des notions plus avancées que

$$\mathbb{E}_\pi[G_{t+1} \mid S_t = s] = \sum_{r,a,s'} \mathbb{E}_\pi[R_{t+1} \mid S_{t+1} = s]$$

Équations de Bellman (4/4)

- Ainsi,

$$\begin{aligned} v_\pi(s) &= \sum_{r,a,s'} \pi(a \mid s)\; p(s', r \mid s, a)\,(r + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s']) \\ &= \sum_{r,a,s'} \pi(a \mid s)\; p(s', r \mid s, a)\,(r + \gamma v_\pi(s')) \end{aligned}$$

- De façon similaire, on trouve

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

- Les équations ci-haut sont **linéaires**

$\rightarrow$ une variable par état pour v_π

$\rightarrow$ une variable par paire état-action pour q_π

- Et elles admettent chacune une solution **unique**

Relations entre v_π et q_π

- Il est possible de calculer v_π en fonction de q_π

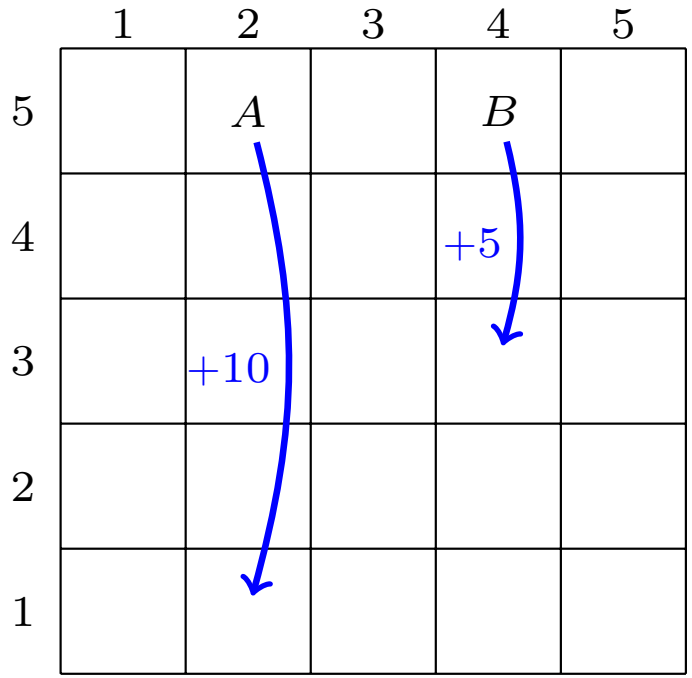
$$v_\pi(s) = \sum_a \pi(a \mid s)\ q_\pi(s, a)$$

- Et, inversement, q_π en fonction de v_π

$$q_\pi(s, a) = \sum_{r, s'} p(s', r \mid s, a)(r + \gamma v_\pi(s'))$$

Exemple de la grille

Environnement



Fonction de valuation v_π

	1	2	3	4	5
5	3.3	8.8	4.4	5.3	1.5
4	1.5	3.0	2.3	1.9	0.5
3	0.1	0.7	0.7	0.4	−0.4
2	−1.0	−0.4	−0.4	−0.6	−1.2
1	−1.9	−1.3	−1.2	−1.4	−2.0

Avec $\gamma = 0.9$ et quatre actions **équiprobables**

Valeurs tronquées à la position des **dixièmes**

(Images adaptées du livre de Sutton et Barto, chapitre 3)

Politiques et valuations optimales

Politique optimale

- Considérons un processus de décision **markovien**
- Avec la même notation ($\mathcal{S}$, $\mathcal{A}$, p, etc.)
- Soit Π l'ensemble de **toutes** les politiques pour ce processus
- On définit une relation **d'ordre partiel** sur Π par

$$\pi \geq \pi'$$

si et seulement si

$$v_\pi(s) \geq v_{\pi'}(s), \quad \text{pour tout } s \in \mathcal{S}$$

- Toute politique π qui est maximale pour la relation $\geq$ est dite **optimale**
- On écrit alors

$$v_*(s) = \max_{\pi \in \Pi} v_\pi(s) \quad \text{et} \quad q_*(s,a) = \max_{\pi \in \Pi} q_\pi(s,a)$$

Équations d'optimalité de Bellman (1/2)

- Soit π_* une politique **optimale**
- Alors

$$
\begin{aligned}
v_*(s) &= \max_a q_{\pi_*}(s,a) \\
&= \max_a E_{\pi_*}\left[G_t \mid S_t = s, A_t = a\right] \\
&= \max_a E_{\pi_*}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right] \\
&= \max_a E\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right] \\
&= \max_a \sum_{r,s'} p(s', r \mid s, a)[r + \gamma\ v_*(s')]
\end{aligned}
$$

Équations d'optimalité de Bellman (1/2)

- De façon similaire, on trouve

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

- Les équations ne sont **pas linéaires**
- Car elles utilisent la fonction $\max$
- Il faut utiliser des approches **numériques**
- En revanche, la solution est **unique** pour v_π et q_π
- On construit facilement une **politique optimale** à partir de là

$\rightarrow$ pour **chaque état**

$\rightarrow$ on vérifie quelles sont les **actions optimales**

$\rightarrow$ on assigne une probabilité **nulle** à chaque action non optimale

$\rightarrow$ on répartit le poids 1 entre les **actions optimales**

Retour sur le robot recycleur (1/2)

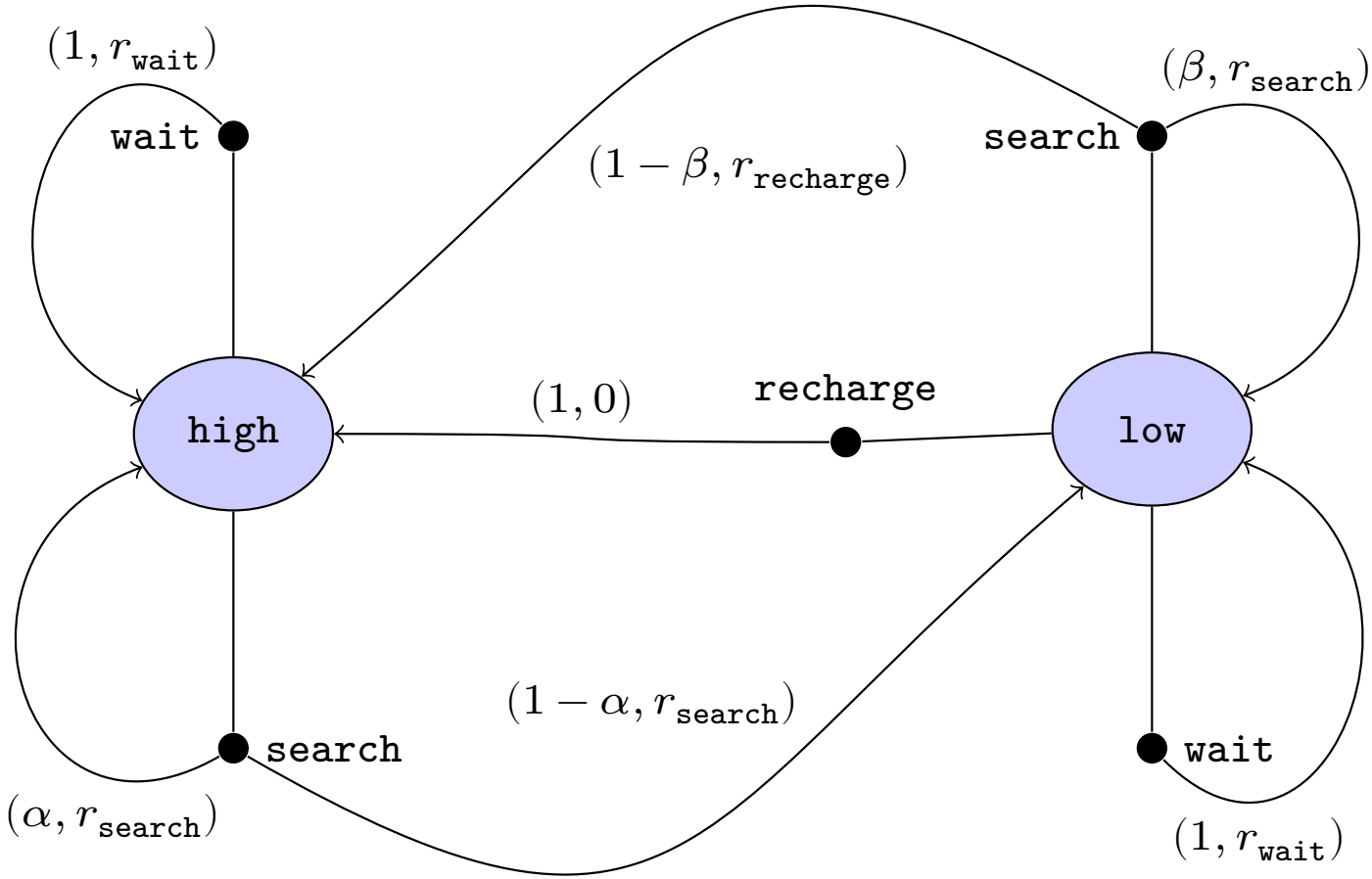


Image adaptée du livre de Sutton et Barto (chapitre 3)

Retour sur le robot recycleur (2/2)

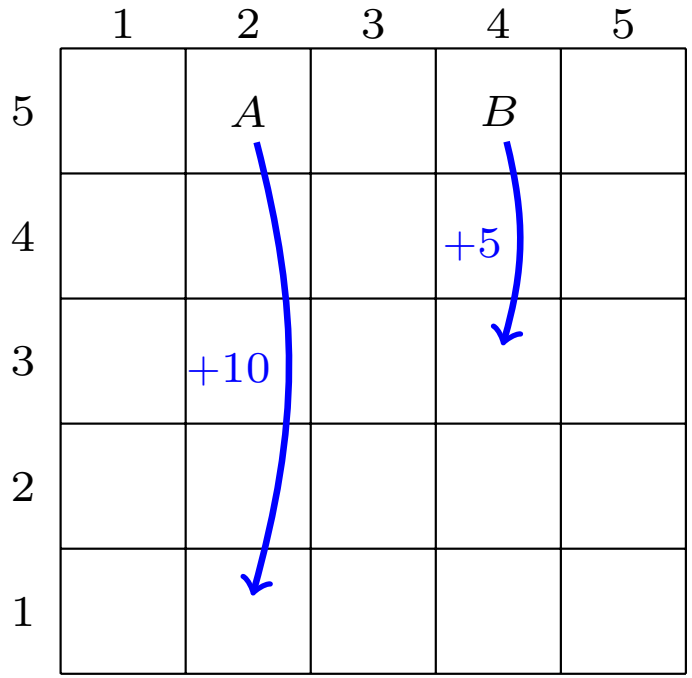
- Les équations à résoudre sont

$$
\begin{aligned}
v_*(\texttt{high}) &= \max\{r_{\texttt{search}} + \gamma[\alpha v_*(\texttt{high}) + (1-\alpha)v_*(\texttt{low})], \\
&\qquad r_{\texttt{wait}} + \gamma v_*(\texttt{high})\} \\
v_*(\texttt{low}) &= \max\{\beta r_{\texttt{search}} + r_{\texttt{recharge}}(1-\beta) \\
&\qquad + \gamma[(1-\beta)v_*(\texttt{high}) + \beta v_*(\texttt{low})], \\
&\qquad r_{\texttt{wait}} + \gamma v_*(\texttt{low}), \\
&\qquad \gamma v_*(\texttt{high})\}
\end{aligned}
$$

- La solution est **unique** lorsque les paramètres sont fixés

Retour sur la grille (1/2)

Environnement



Fonction de valuation v_π

	1	2	3	4	5
5	3.3	8.8	4.4	5.3	1.5
4	1.5	3.0	2.3	1.9	0.5
3	0.1	0.7	0.7	0.4	−0.4
2	−1.0	−0.4	−0.4	−0.6	−1.2
1	−1.9	−1.3	−1.2	−1.4	−2.0

Avec $\gamma = 0.9$ et quatre actions **équiprobables**

Valeurs tronquées à la position des **dixièmes**

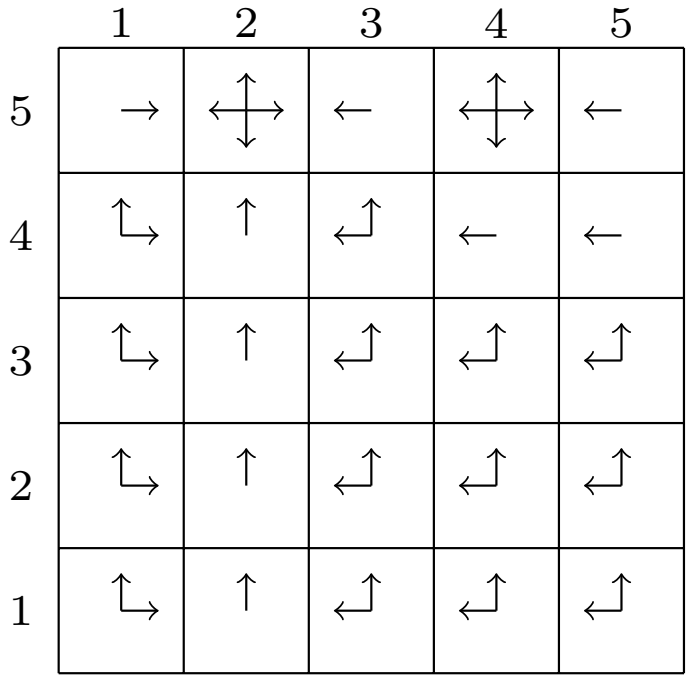
(Images adaptées du livre de Sutton et Barto, chapitre 3)

Retour sur la grille (2/2)

Valuation optimale v_*

	1	2	3	4	5
5	22.0	24.4	22.0	19.4	17.5
4	19.8	22.0	19.8	17.8	16.0
3	17.8	19.8	17.8	16.0	14.4
2	16.0	17.8	16.0	14.4	13.0
1	14.4	16.0	14.4	13.0	11.7

Politique optimale π_*



Valeurs tronquées à la position des **dixièmes**

(Images adaptées du livre de Sutton et Barto, chapitre 3)

En pratique

Coûteux

- On résoud **rarement** les équations d'optimalité de Bellman
- Trop coûteux en **temps de calcul**
- Et en **espace mémoire**

Approximations

- Ignorer les états **peu probables**
- Méthode de **Monte-Carlo**
- Méthode des **différences temporelles**

La suite dans les **prochaines parties**!