Deep Reinforcement Learning

4A – IABD – Intro

Spoiler

- Livre au top :
  - http://incompleteideas.net/book/RLbook2018.pdf
- Un cours sur Coursera qui suit le livre (très bonne qualité):
  - https://www.coursera.org/specializations/reinforcement-learning
- Cours dispensé à Stanford :
  - https://www.youtube.com/watch?v=FgzM3zpZ55o&list=PLoROMvodv4rOSOPzutgyCTapiGlY2Nd8u
- Intro dispensée par DeepMind :
  - https://www.youtube.com/watch?v=iOh7QUZGyiU&list=PLqYmG7hTraZDNJre23vqCGIVpfZ_K2RZs

Markov Decision Process

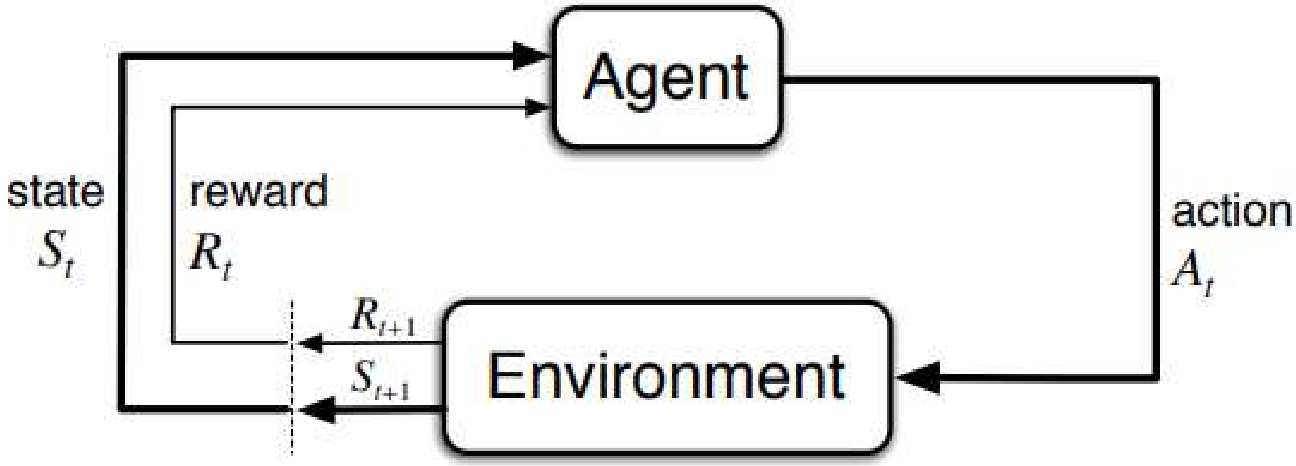

Figure 3.1: The agent–environment interaction in a Markov decision process.

Markov Decision Process

- Ensemble d'Action : A
- Ensemble d'Etats : S
- Ensemble de Récompenses immédiates : R

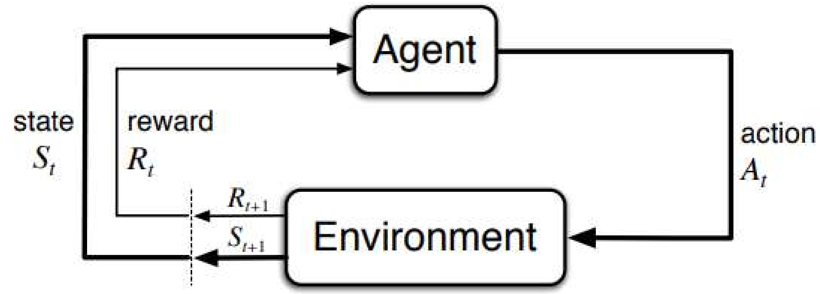

Figure 3.1: The agent–environment interaction in a Markov decision process.

Markov Decision Process

- Ensemble d'Action : A
- Ensemble d'Etats : S
- Ensemble de Récompenses immédiates : R
- Hypothèse : S_{t+1} et R_{t+1} ne dependent que de S_t et de A_t

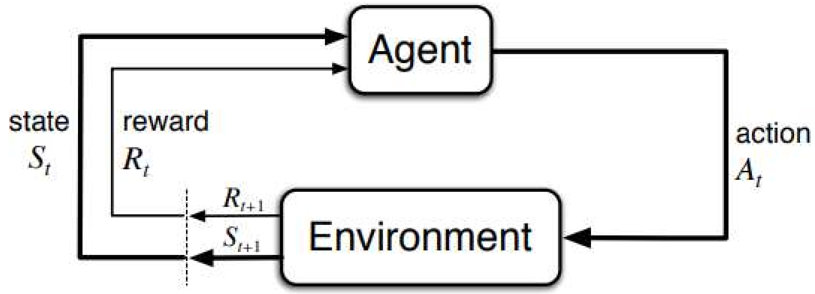

Figure 3.1: The agent–environment interaction in a Markov decision process.

Markov Decision Process

- Ensemble d'Action : A
- Ensemble d'Etats : S
- Ensemble de Récompenses immédiates : R
- Hypothèse : S_{t+1} et R_{t+1} ne dependent que de S_t et de A_t
- On parle d'hypothèse Markovienne

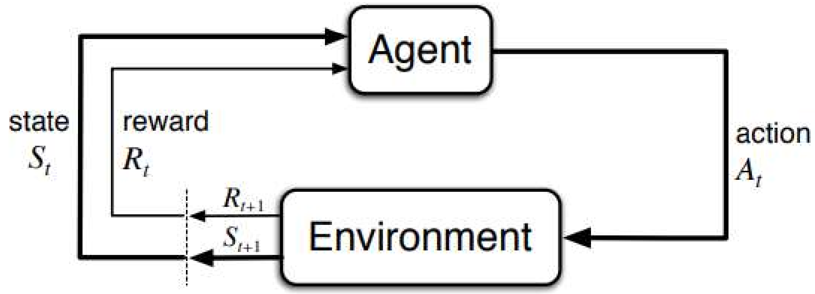

Figure 3.1: The agent–environment interaction in a Markov decision process.

Markov Decision Process

- Ensemble d'Action : A
- Ensemble d'Etats : S
- Ensemble de Récompenses immédiates : R
- Hypothèse : S_{t+1} et R_{t+1} ne dependent que de S_t et de A_t
- On parle d'hypothèse Markovienne
- On suppose l'existence de :
  - $p(s', r | s, a)$
  - (Environment dynamics)

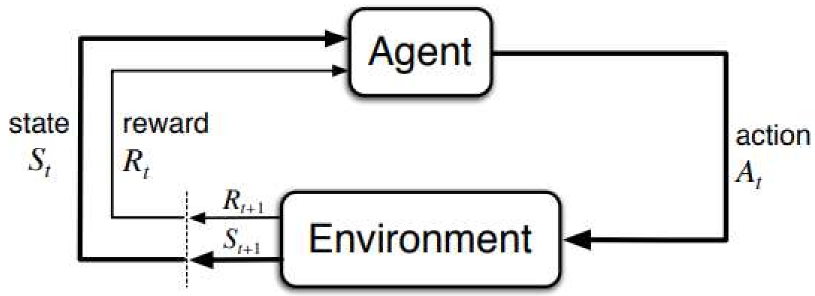

Figure 3.1: The agent–environment interaction in a Markov decision process.

Markov Decision Process

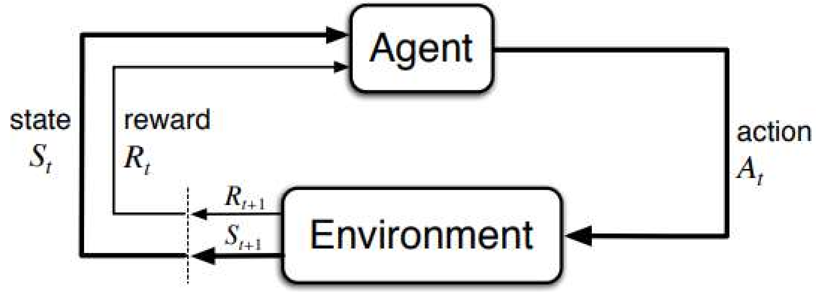

Figure 3.1: The agent–environment interaction in a Markov decision process.

- Ensemble d’Action : A
- Ensemble d’Etats : S
- Ensemble de Récompenses immédiates : R
- Hypothèse : S_{t+1} et R_{t+1} ne dependent que de S_t et de A_t
- On parle d’hypothèse Markovienne
- On suppose l’existence de :
  - $p(s', r|s, a)$
  - (Environment dynamics)
  - Il s’agit de la <u>probabilité</u> d’obtenir l’état s’ et la recompense r à partir de l’état s en effectuant l’action a

Markov Decision Process

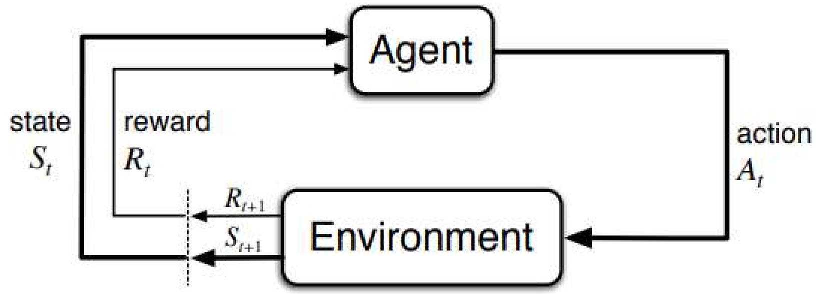

Figure 3.1: The agent–environment interaction in a Markov decision process.

- On suppose l'existence de :
  - $p(s', r|s, a)$
  - (Environment dynamics)
  - Il s'agit de la <u>probabilité</u> d'obtenir l'état s' et la recompense r à partir de l'état s en effectuant l'action a
- Ainsi :

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \,|\, s, a) = 1, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s). \tag{3.3}$$

Quel est le but de l'Agent dans ce contexte ?

- Deux cas de figures :
  - Tâches épisodique
  - Tâches continues

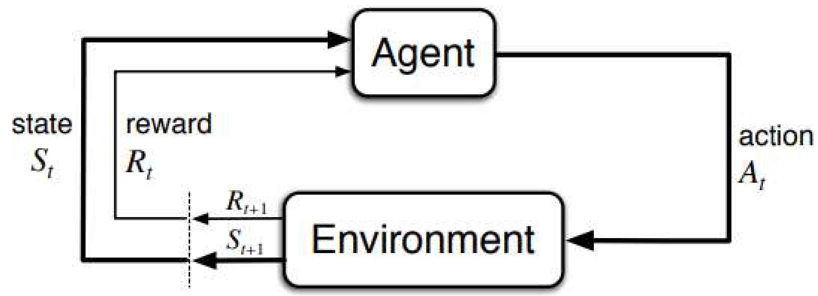

Figure 3.1: The agent–environment interaction in a Markov decision process.

Quel est le but de l'Agent dans ce contexte ?

- Deux cas de figures :
  - Tâches épisodique
    - L'agent est certain de se retrouver dans une situation terminale tôt ou tard
  - Tâches continues
    - Il n'y a pas de situation terminale

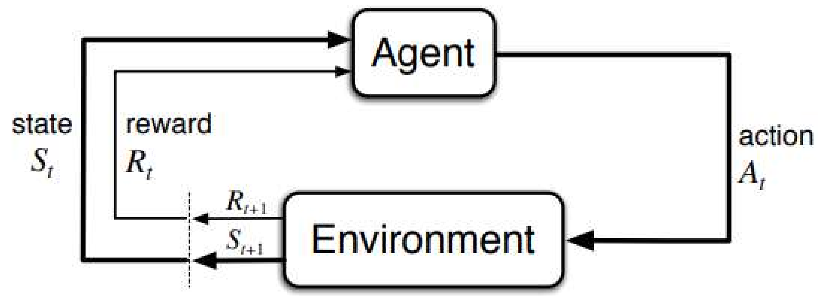

Figure 3.1: The agent–environment interaction in a Markov decision process.

Quel est le but de l'Agent dans ce contexte ?

- Deux cas de figures :
  - Tâches épisodique
    - L'agent est certain de se retrouver dans une situation terminale tôt ou tard
  - Tâches continues
    - Il n'y a pas de situation terminale

- Le but (Goal) de l'agent est de maximiser sa récompense cumulée long terme
  - $G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T, \quad (3.7)$
  - $G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \quad (3.8)$

Quel est le but de l’Agent dans ce contexte ?γ

- Deux cas de figures :
  - Tâches épisodique
    - L’agent est certain de se retrouver dans une situation terminale tôt ou tard
  - Tâches continues
    - Il n’y a pas de situation terminale

- Le but (Goal) de l’agent est de maximiser sa récompense cumulée long terme
  - $$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T, \qquad (3.7)$$
  - $$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \qquad (3.8)$$

Quel est le but de l'Agent dans ce contexte ?

- Le but (Goal) de l'agent est de maximiser sa récompense cumulée long terme :
  - $$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T, \qquad (3.7)$$
  - $$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \qquad (3.8)$$

- On remarque :
  - $$G_t = R_{t+1} + \gamma G_{t+1}$$
  - $$G_t \doteq \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

Comment y parvenir ?

- Notations :
  - « policy »
    - $\pi(a|s)$
    - Renvoyant la probabilité de réaliser l'action a dans un état s
  - « value function »
    - $v_\pi(s)$
    - Renvoyant la moyenne (espérance) de la récompense long terme cumulée à partir de l'état s en suivant la stratégie π
  - « action-value function »
    - $q_\pi(s,a)$
    - Renvoyant la moyenne (espérance) de la récompense long terme cumulée à partir de l'état s en effectuant l'action a puis en suivant la stratégie π à partir de l'état suivant

Comment y parvenir ?

- Notations :
  - « policy »
    - $\pi(a|s)$
    - Renvoyant la probabilité de réaliser l'action a dans un état s
  - « value function »
    - $v_\pi(s)$
    - Renvoyant la moyenne (espérance) de la récompense long terme cumulée à partir de l'état s en suivant la stratégie π

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right], \text{ for all } s \in \mathcal{S}, \qquad (3.12)$$

  - « action-value function »
    - $q_\pi(s, a)$
    - Renvoyant la moyenne (espérance) de la récompense long terme cumulée à partir de l'état s en effectuant l'action a puis en suivant la stratégie π à partir de l'état suivant

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s, A_t = a\right]. \qquad (3.13)$$

Comment y parvenir ?

- Nous pouvons réécrire $v_\pi(s)$:

$$
\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r \mid s, a) \Big[r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s'] \Big] \\
&= \sum_a \pi(a|s) \sum_{s', r} p(s', r \mid s, a) \Big[r + \gamma v_\pi(s') \Big], \quad \text{for all } s \in \mathcal{S},
\end{aligned}
$$

- Il s’agit d’une des équations de Bellman !

Comment y parvenir ?

- Nous pouvons réécrire $v_\pi(s)$:

$$\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r \mid s, a)\Big[r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s']\Big] \\
&= \sum_a \pi(a|s) \sum_{s', r} p(s', r \mid s, a)\Big[r + \gamma v_\pi(s')\Big], \quad \text{for all } s \in \mathcal{S},
\end{aligned}$$

- Il s'agit d'une des équations de Bellman !
- Tentons d'évaluer la « value function » d'une « policy » uniformément aléatoire sur un exemple jouet => « Policy Evaluation »

Comment y parvenir ?

- Tentons d'évaluer la « value function » d'une « policy » uniformément aléatoire sur un exemple jouet
  - « Policy Evaluation »

- On parle de tâche de « Prediction »

- Répéter (pour $k = 0..$) jusqu'à convergence et pour tous les s :

$$
\begin{aligned}
v_{k+1}(s) &\doteq \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\Big[r + \gamma v_k(s')\Big], \qquad (4.5)
\end{aligned}
$$

Comment y parvenir ?

- Pseudo code pour évaluer une stratégie a.k.a « Policy Evaluation » :

Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input π, the policy to be evaluated
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
  $\Delta \leftarrow 0$
  Loop for each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s', r \,|\, s, a)\big[r + \gamma V(s')\big]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Comment y parvenir ?

- Ne perdons pas de vue l'objectif !
  - Trouver la meilleure des « policies » ! (a.k.a. celle qui maximise en moyenne le reward long terme cumulé)

Comment y parvenir ?

- Ne perdons pas de vue l'objectif !
  - Trouver la meilleure des « policies » ! (a.k.a. celle qui maximise en moyenne le reward long terme cumulé)
  - Notons ces ou cette stratégie π_*
  - Il peut y en avoir plusieurs !
    - Exemple

Comment y parvenir ?

- Ne perdons pas de vue l'objectif !
  - Trouver la meilleure des « policies » ! (a.k.a. celle qui maximise en moyenne le reward long terme cumulé)
  - On parle de tâche de « Control »
  - Notons ces ou cette stratégies π_*
  - Il peut y en avoir plusieurs !
    - Exemple
  - Cependant, elles ont toutes la même « value function » optimale associée
  - Notons cette dernière v_*

Comment y parvenir ?

- Ne perdons pas de vue l'objectif !
  - Trouver la meilleure des « policies » ! (a.k.a. celle qui maximise en moyenne le reward long terme cumulé)
  - On parle de tâche de « Control »
  - Notons ces ou cette stratégies π_*
  - Il peut y en avoir plusieurs !
    - Exemple
  - Cependant, elles ont toutes la même « value function » optimale associée
  - Notons cette dernière v_*
- En effet :
  - $v_*(s) \doteq \max_\pi v_\pi(s), \qquad (3.15)$

Comment y parvenir ?

- Cependant, elles ont toutes la même « value function » optimale associée
  - Notons cette dernière v_*
- En effet :
  - $$v_*(s) \doteq \max_\pi v_\pi(s), \tag{3.15}$$
- Il en va de même pour l' « action-value function » optimale
  - Notons cette dernière q_*
- En effet :
  - $$q_*(s,a) \doteq \max_\pi q_\pi(s,a), \tag{3.16}$$

Comment y parvenir ?

- Si l'on trouve v_* et que l'on connait $p(s', r|s, a)$ alors nous pouvons en déduire une des π_* !

- Si l'on trouve q_* alors nous pouvons en déduire une des π_* !

- Comment trouver v_* ou q_* ?

Comment y parvenir ?

- Partons de deux des équations d'optimalité de Bellman :

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\
&= \max_a \mathbb{E}_{\pi_*}[G_t \mid S_t = s, A_t = a] \\
&= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \\
&= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \\
&= \max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma v_*(s')\big].
\end{aligned}
$$

$$
\begin{aligned}
q_*(s,a) &= \mathbb{E}\Big[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \;\Big|\; S_t = s, A_t = a\Big] \\
&= \sum_{s',r} p(s', r \mid s, a)\Big[r + \gamma \max_{a'} q_*(s', a')\Big].
\end{aligned}
$$

Comment y parvenir ?

- Pseudo code pour évaluer puis améliorer en boucle un stratégie a.k.a. « Policy Iteration »:

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
       $\Delta \leftarrow 0$
       Loop for each $s \in \mathcal{S}$:
           $v \leftarrow V(s)$
           $V(s) \leftarrow \sum_{s',r} p(s',r|s,\pi(s))\big[r + \gamma V(s')\big]$
           $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
       old-action $\leftarrow \pi(s)$
       $\pi(s) \leftarrow \operatorname{argmax}_a \sum_{s',r} p(s',r|s,a)\big[r + \gamma V(s')\big]$
       If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Comment y parvenir ?

- Nous pouvons être plus rapide en itérant directement sur v a.k.a. « Value Iteration »:

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $\quad v \leftarrow V(s)$
| $\quad V(s) \leftarrow \max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
| $\quad \Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \operatorname{argmax}_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$