Apprentissage automatique

apprendre par renforcement

Charles Prud'homme

Charles.Prudhomme@imt-atlantique.fr

TASC (CNRS/IMT Atlantique)

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Apprentissage de réflexes par renforcement

Mise en situation

Vous jouez pour la première fois à un jeu dont vous ne connaissez pas les règles. Après une centaine de coups, votre adversaire annonce : "**Tu as perdu**".

Définition

Un algorithme d'apprentissage par renforcement s'attache

- à apprendre les actions à prendre
- à partir d'expériences
- de façon à optimiser une récompense quantitative
- au cours du temps.

Apprentissage de réflexes par renforcement

Mise en situation

Vous jouez pour la première fois à un jeu dont vous ne connaissez pas les règles. Après une centaine de coups, votre adversaire annonce : "**Tu as perdu**".

Définition

Un algorithme d'apprentissage par renforcement s'attache

- à apprendre les actions à prendre
- à partir d'expériences
- de façon à optimiser une récompense quantitative
- au cours du temps.

Suppositions

1. L'agent ne connaît pas ou mal son environnement :
   - Ne connait pas *a priori* quels sont les renforcements associés à chaque état ou transition
   - Ne connait pas la topologie de l'espace
2. L'agent ne connaît pas ou mal l'effet de ses actions dans un état donné

Suppositions

1. L'agent ne connaît pas ou mal son environnement :
   - Ne connait pas *a priori* quels sont les renforcements associés à chaque état ou transition
   - Ne connait pas la topologie de l'espace
2. L'agent ne connaît pas ou mal l'effet de ses actions dans un état donné

Défi

- Connaissance faible du monde et de ses réactions
- Mesures sur les états peuvent être imparfaites
- Renforcements pauvre en information, parfois tardif
- $\Rightarrow$ Nécessite énormément d'interactions
- Relativement inefficace mais très adaptable

Suppositions

1. L'agent ne connaît pas ou mal son environnement :
   - Ne connait pas *a priori* quels sont les renforcements associés à chaque état ou transition
   - Ne connait pas la topologie de l'espace
2. L'agent ne connaît pas ou mal l'effet de ses actions dans un état donné

Seules hypothèses valables

Le monde est

- stochastique : les actions peuvent avoir des effets non déterministes,
- stationnaire : les probabilités de transition et les renforcements restent stables

Modélisation

L'agent communique avec son environnement par 3 canaux :

1. Un **canal perceptif** : $s(t)$, mesure l'état dans lequel il se trouve dans l'environnement
2. Un canal spécifique **aux signaux de renforcement** : $r(t)$, renseigne sur la qualité de l'état courant,
3. Un canal **d'action** qui transmet l'action de l'agent, $a(t)$, à l'environnement.

Notations

Modélisation

L'agent communique avec son environnement par 3 canaux :

1. Un **canal perceptif** : $s(t)$, mesure l'état dans lequel il se trouve dans l'environnement
2. Un canal spécifique **aux signaux de renforcement** : $r(t)$, renseigne sur la qualité de l'état courant,
3. Un canal **d'action** qui transmet l'action de l'agent, $a(t)$, à l'environnement.

Notations

À l'instant t

- $s_t \in \mathcal{E}$, l'espace des états
- $r_t \in \mathcal{R}$, l'espace des signaux, $r(t) \in [-a, +b],\ a, b \in \mathbb{R}^+$
- $a_t \in \mathcal{A}$, l'espace des actions

Modélisation

L'agent communique avec son environnement par 3 canaux :

1. Un **canal perceptif** : $s(t)$, mesure l'état dans lequel il se trouve dans l'environnement
2. Un canal spécifique **aux signaux de renforcement** : $r(t)$, renseigne sur la qualité de l'état courant,
3. Un canal **d'action** qui transmet l'action de l'agent, $a(t)$, à l'environnement.

Notations

- L' **agent** est une fonction $s_t \mapsto a_t$
- Cette fonction de comportement est appelée *politique*, π_t
- $\pi(s, a)$: la probabilité de choisir l'action a dans l'état s.

Modélisation

L'agent communique avec son environnement par 3 canaux :

1. Un **canal perceptif** : $s(t)$, mesure l'état dans lequel il se trouve dans l'environnement
2. Un canal spécifique **aux signaux de renforcement** : $r(t)$, renseigne sur la qualité de l'état courant,
3. Un canal **d'action** qui transmet l'action de l'agent, $a(t)$, à l'environnement.

Notations

- L'**environnement** est une fonction $(s_t, a_t) \mapsto (s_{t+1}, r_t)$
- En pratique, elle est décomposée :
  - la fonction de transition entre états, $T : (s_t, a_t) \mapsto s_{t+1}$
  - la fonction de renforcement **immédiat**, $R : (s_t, a_t) \mapsto r_t$

Fonctionnement

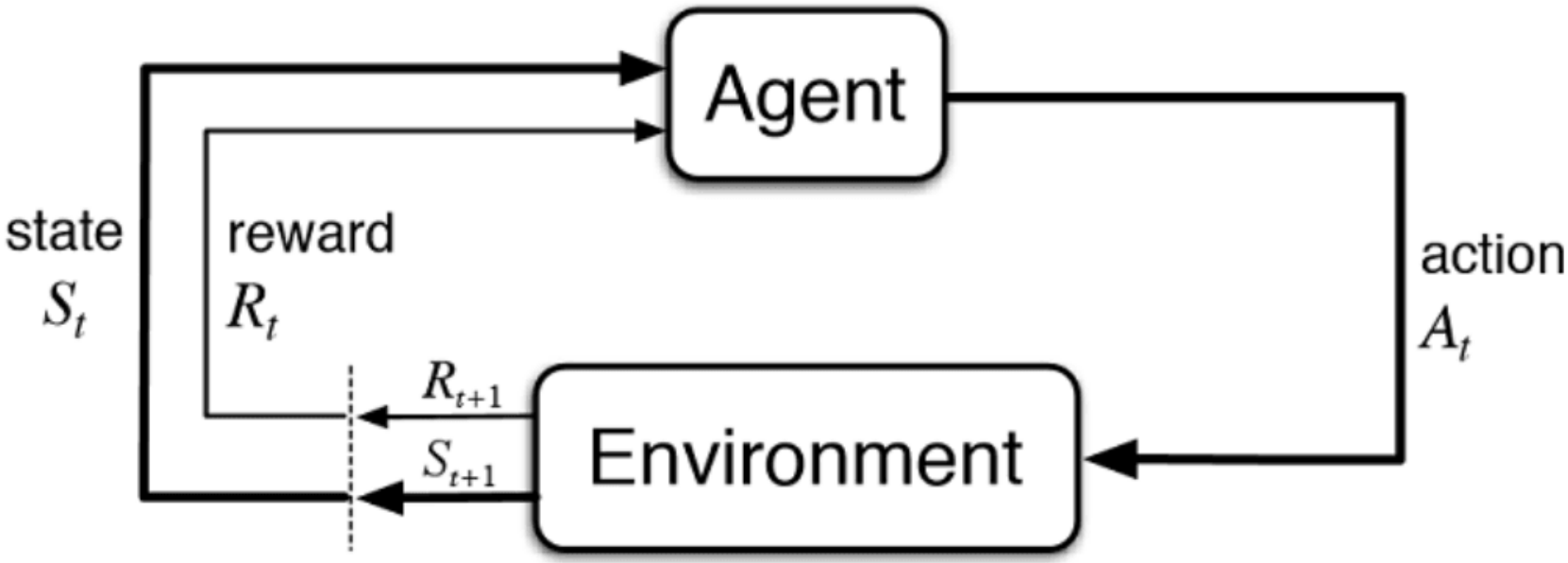

FIGURE – Boucle d'actions.
Source : Sutton, R. S. and Barto, A. G. Introduction to Reinforcement Learning

Les mesures de gain

Objectif

Un apprenant est plongé dans un environnement et doit essayer, par ses actions, de **maximiser une mesure de gain** dépendant des signaux qu'il reçoit pendant son existence.

Les mesures de gain

- Pas de mesure de gain universelle
- Doit être spécifiée par problème
- En général, on choisit de cumuler le gain dans le temps
- Choix de la mesure est **déterminant**

Les mesures de gain

- **Gain cumulé avec horizon fini**

$$R_T = \sum_{t=0}^{T-1} r_t | s_0$$

- **Gain cumulé avec intérêt et horizon infini**

$$R = \sum_{t=0}^{\infty} \gamma^t r_t | s_0, \ 0 \leq \gamma \leq 1$$

- **Gain en moyenne**

$$R_T = \frac{1}{T-1} \sum_{t=0}^{T-1} r_t | s_0$$

Les processus décisionnels de Markov

Apprentissage par renforcement
$\Rightarrow$ problème de *décision séquentielle dans l'incertain*.

Chaque décision de l'agent a un effet sur les décisions à suivre
$\Rightarrow$ les entrées ne peuvent pas être considérées comme *i.i.d.*.

Formalisation par Processus Décisionnel de Markov

Généralisation de la recherche de plus court chemin dans un environnement stochastique.

La solution d'un problème de décision markovien

- n'est pas une séquence de décisions
- est une politique/stratégie qui spécifie l'action à prendre en chacun des états rencontrés pour maximiser l'espérance de gain.

Les processus décisionnels de Markov

Apprentissage par renforcement
$\Rightarrow$ problème de *décision séquentielle dans l'incertain*.

Chaque décision de l'agent a un effet sur les décisions à suivre
$\Rightarrow$ les entrées ne peuvent pas être considérées comme *i.i.d.*.

Formalisation par Processus Décisionnel de Markov

Généralisation de la recherche de plus court chemin dans un environnement stochastique.

La solution d'un problème de décision markovien

- n'est pas une séquence de décisions
- est une politique/stratégie qui spécifie l'action à prendre en chacun des états rencontrés pour maximiser l'espérance de gain.

Les processus décisionnels de Markov

Apprentissage par renforcement
$\Rightarrow$ problème de *décision séquentielle dans l'incertain*.

Chaque décision de l'agent a un effet sur les décisions à suivre
$\Rightarrow$ les entrées ne peuvent pas être considérées comme *i.i.d.*.

Formalisation par Processus Décisionnel de Markov

Généralisation de la recherche de plus court chemin dans un environnement stochastique.

La solution d'un problème de décision markovien

- n'est pas une séquence de décisions
- est une politique/stratégie qui spécifie l'action à prendre en chacun des états rencontrés pour maximiser l'espérance de gain.

Les processus décisionnels de Markov

Apprentissage par renforcement
$\Rightarrow$ problème de *décision séquentielle dans l'incertain*.

Chaque décision de l'agent a un effet sur les décisions à suivre
$\Rightarrow$ les entrées ne peuvent pas être considérées comme *i.i.d.*.

Formalisation par Processus Décisionnel de Markov

Généralisation de la recherche de plus court chemin dans un environnement stochastique.

La solution d'un problème de décision markovien

- **n'est pas** une séquence de décisions
- est une **politique**/stratégie qui spécifie l'action à prendre en chacun des états rencontrés pour maximiser l'espérance de gain.

Les processus décisionnels de Markov

Apprentissage par renforcement
$\Rightarrow$ problème de *décision séquentielle dans l'incertain*.

Chaque décision de l'agent a un effet sur les décisions à suivre
$\Rightarrow$ les entrées ne peuvent pas être considérées comme *i.i.d.*.

Formalisation par Processus Décisionnel de Markov

Généralisation de la recherche de plus court chemin dans un environnement stochastique.

La solution d'un problème de décision markovien

- **n'est pas** une séquence de décisions
- est une **politique**/stratégie qui spécifie l'action à prendre en chacun des états rencontrés pour maximiser l'espérance de gain.

Les approches

Plusieurs approches pour résoudre ce problème :

Model-Based Reinforcement Learning

- Apprendre directement un modèle de l'environnement (*i.e.*, fonction de renforcement + fonction de transition)

\+ $\approx$ Apprentissage supervisé

\- Ignore les interactions entre états

Les approches

Plusieurs approches pour résoudre ce problème :

Value-Based RL

- Introduit les **fonctions d'utilité** au Model-based RL
- Utilité : considére les intéractions
- Agir sur le monde et calculer sur le long terme la qualité des états ou des couples état-action

soit $V(s)$: espérance de gain à partir d'un **état**

soit $Q(s, a)$: espérance de gain à partir d'un **couple état-action**

Fonctions d'utilité

Volontés

- Optimiser la conduite sur le long terme
- . . . sur la base de décisions locales
- . . . ne nécessitant pas de recherche en avant

$\Rightarrow$ l'information locale **doit refléter** l'espérance de gain à long terme !

$$V^{\pi}(s) = \mathbb{E}_{\pi}\{R_t | s_t = s\}$$

FIGURE – Espérance de gain à partir de l'étape s en suivant la politique π.

Fonctions d'utilité

Volontés

- Optimiser la conduite sur le long terme
- . . . sur la base de décisions locales
- . . . ne nécessitant pas de recherche en avant

$\Rightarrow$ l'information locale **doit refléter** l'espérance de gain à long terme !

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\{R_t | s_t = s, a_t = a\}$$

FIGURE – Espérance de gain à partir de l'étape s, en effectuant l'action a, puis en suivant la politique π.

Préliminaires

On suppose ici connus :

- Les probabilités de transition
- Les renforcements associés
- L'agent sait ce qu'il peut atteindre dans l'environnement

Mais ne connait pas les fonctions d'utilités

Donc il ne connait pas l'impact de ses décisions sur le long terme

Deux problèmes à résoudre

1. Comment les apprendre pour une politique donnée ?
2. Comment approcher une politique optimale ?

Préliminaires

On suppose ici connus :

- Les probabilités de transition
- Les renforcements associés
- L'agent sait ce qu'il peut atteindre dans l'environnement

Mais ne connait pas les fonctions d'utilités

Donc il ne connait pas l'impact de ses décisions sur le long terme

Deux problèmes à résoudre

1. Comment les apprendre pour une politique donnée ?
2. Comment approcher une politique optimale ?

Évaluer une politique
par propagation locale d'information

Approche simple

- Tester tous les états s
- En suivant la politique π (au moins une fois)
- Et calculer la moyenne des gains cumulés

Notation

$\pi(s, a)$: probabilité de choisir a dans l'état s alors qu'on applique la politique π

Évaluer une politique
par propagation locale d'information

Approche simple

- Tester tous les états s
- En suivant la politique π (au moins une fois)
- Et calculer la moyenne des gains cumulés

Notation

$\pi(s, a)$: probabilité de choisir a dans l'état s alors qu'on applique la politique π

Évaluation

P-ex., gain cumulé avec intérêts et horizon infini :

$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}_{\pi}\{R_t | s_t = s\} \\
&= \mathbb{E}_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s \right\} \\
&= \mathbb{E}_{\pi}\left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \middle| s_t = s \right\} \\
&= \mathbb{E}_{\pi}\left\{ r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s' \right\}
\end{aligned}
$$

L'espérance de gain d'un état dépend de

- son propre renforcement
- et des espérances de gain des états qu'il peut atteindre

Évaluation

P-ex., gain cumulé avec intérêts et horizon infini :

$$V^\pi(s) = \mathbb{E}_\pi\{R_t | s_t = s\}$$

$$= \mathbb{E}_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s\right\}$$

$$= \mathbb{E}_\pi\left\{r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \middle| s_t = s\right\}$$

$$= \mathbb{E}_\pi\left\{r_{t+1} + \gamma V^\pi(s_{t+1}) | s_t = s'\right\}$$

L'espérance de gain d'un état dépend de

- son propre renforcement
- et des espérances de gain des états qu'il peut atteindre

Évaluation

P-ex., gain cumulé avec intérêts et horizon infini :

$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}_{\pi}\{R_t | s_t = s\} \\
&= \mathbb{E}_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s \right\} \\
&= \mathbb{E}_{\pi}\left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \middle| s_t = s \right\} \\
&= \mathbb{E}_{\pi}\left\{ r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s' \right\}
\end{aligned}
$$

L'espérance de gain d'un état dépend de

- son propre renforcement
- et des espérances de gain des états qu'il peut atteindre

Évaluation

P-ex., gain cumulé avec intérêts et horizon infini :

$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}_{\pi}\{R_t | s_t = s\} \\
&= \mathbb{E}_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s \right\} \\
&= \mathbb{E}_{\pi}\left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \middle| s_t = s \right\} \\
&= \mathbb{E}_{\pi}\left\{ r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s' \right\}
\end{aligned}
$$

L'espérance de gain d'un état dépend de

- son propre renforcement
- et des espérances de gain des états qu'il peut atteindre

Évaluation

P-ex., gain cumulé avec intérêts et horizon infini :

$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}_{\pi}\{R_t | s_t = s\} \\
&= \mathbb{E}_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s \right\} \\
&= \mathbb{E}_{\pi}\left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \middle| s_t = s \right\} \\
&= \mathbb{E}_{\pi}\left\{ r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s' \right\}
\end{aligned}
$$

L'espérance de gain d'un état dépend de

- son propre renforcement
- et des espérances de gain des états qu'il peut atteindre

Évaluation

P-ex., gain cumulé avec intérêts et horizon infini :

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_\pi\{R_t | s_t = s, a_t = a\} \\ &= \mathbb{E}_\pi\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s, a_t = a \right\} \\ &= \gamma \sum_{s'} p^\pi(s'|s_t) V^\pi(s') \end{aligned}$$

notons que : $V^\pi(s) = \sum_a \pi(s,a) Q^\pi(s,a)$

L'espérance de gain d'un état dépend de

- son propre renforcement
- et des espérances de gain des états qu'il peut atteindre

Évaluation

P-ex., gain cumulé avec intérêts et horizon infini :

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}_\pi\{R_t | s_t = s, a_t = a\} \\ &= \mathbb{E}_\pi\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s, a_t = a \right\} \\ &= \gamma \sum_{s'} p^\pi(s'|s_t) V^\pi(s') \end{aligned}$$

notons que : $V^\pi(s) = \sum_a \pi(s,a) Q^\pi(s,a)$

L'espérance de gain d'un état dépend de

- son propre renforcement
- et des espérances de gain des états qu'il peut atteindre

Politique optimale

Ordre sur les politiques

$$\pi \geq \pi' \iff V^{\pi}(s) \geq V^{\pi'}(s), \forall s \in \mathcal{E}$$

Politique optimale π^*

Si une politique est supérieure à toutes les autres, elle est optimale et notée π^*.

Conduite optimale

Politique optimale

Ordre sur les politiques

$$\pi \geq \pi' \iff V^{\pi}(s) \geq V^{\pi'}(s), \forall s \in \mathcal{E}$$

Politique optimale π^*

Si une politique est supérieure à toutes les autres, elle est optimale et notée π^*.

Conduite optimale

Politique optimale

Ordre sur les politiques

$$\pi \geq \pi' \iff V^{\pi}(s) \geq V^{\pi'}(s), \forall s \in \mathcal{E}$$

Politique optimale π^*

Si une politique est supérieure à toutes les autres, elle est optimale et notée π^*.

Conduite optimale

Politique optimale

Ordre sur les politiques

$$\pi \geq \pi' \iff V^{\pi}(s) \geq V^{\pi'}(s), \forall s \in \mathcal{E}$$

Politique optimale π^*

Si une politique est supérieure à toutes les autres, elle est optimale et notée π^*.

Conduite optimale

Si l'agent dispose des $V^*(s) = \max_{\pi} V^{\pi}(s)$

Alors $\forall s \in \mathcal{E}$

- pour chaque action a, faire un pas en avant
- choisir l'action avec la meilleure espérance de gain

Politique optimale

Ordre sur les politiques

$$\pi \geq \pi' \iff V^{\pi}(s) \geq V^{\pi'}(s), \forall s \in \mathcal{E}$$

Politique optimale π^*

Si une politique est supérieure à toutes les autres, elle est optimale et notée π^*.

Conduite optimale

Si l'agent dispose des $Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$

Alors $\forall s \in \mathcal{E}$ et $\forall a \in \mathcal{A}$

- choisir l'action avec la meilleure espérance de gain

Processus itératif

$$\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{A} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{A} \pi_2 \xrightarrow{E} \cdots \xrightarrow{A} \pi^* \xrightarrow{E} V^*$$

*E : évaluation, A : amélioration.

Bilan

+ Convergence en un nombre fini d'itérations
  - Phase d'évaluation de politique est **très coûteuse**
+ On peut la limiter à un passage par état
  - En pratique, tous les états ne peuvent pas être visités...

Sans modèle du monde : la **méthode de Monte-Carlo**
→ Estimer les valeurs de probabilités de transitions et de renforcement par un échantillonage.

Processus itératif

$$\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{A} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{A} \pi_2 \xrightarrow{E} \cdots \xrightarrow{A} \pi^* \xrightarrow{E} V^*$$

*E : évaluation, A : amélioration.

Bilan

+ Convergence en un nombre fini d'itérations
  - Phase d'évaluation de politique est **très coûteuse**
+ On peut la limiter à un passage par état
  - En pratique, tous les états ne peuvent pas être visités...

Sans modèle du monde : la **méthode de Monte-Carlo**
→ Estimer les valeurs de probabilités de transitions et de renforcement par un échantillonage.

Processus itératif

$$\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{A} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{A} \pi_2 \xrightarrow{E} \cdots \xrightarrow{A} \pi^* \xrightarrow{E} V^*$$

*E : évaluation, A : amélioration.

Bilan

+ Convergence en un nombre fini d'itérations
- Phase d'évaluation de politique est **très coûteuse**
+ On peut la limiter à un passage par état
- En pratique, tous les états ne peuvent pas être visités...

Sans modèle du monde : la **méthode de Monte-Carlo**
$\rightarrow$ Estimer les valeurs de probabilités de transitions et de renforcement par un échantillonage.

Processus itératif

$$\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{A} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{A} \pi_2 \xrightarrow{E} \cdots \xrightarrow{A} \pi^* \xrightarrow{E} V^*$$

*E : évaluation, A : amélioration.

Bilan

+ Convergence en un nombre fini d'itérations
- Phase d'évaluation de politique est **très coûteuse**
+ On peut la limiter à un passage par état
- En pratique, tous les états ne peuvent pas être visités...

Sans modèle du monde : la **méthode de Monte-Carlo**
$\rightarrow$ Estimer les valeurs de probabilités de transitions et de renforcement par un échantillonage.

Méthode des différences temporelles

Principes

On approxime la formule : $V^\pi(s) = \mathbb{E}_\pi\{R_t | s_t = s\}$ par :

$$V(s) \leftarrow V(s) + \alpha[R_t - V(s)]$$

où R_t mesure le gain après l'instant t en partant de s,
$R_t = r + V(s')$.

- $R_t - V(s)$ calcul l'erreur sur l'estimation courante = direction
- Pas besoin de connaissances *a priori* sur l'environnement
- Nécessite de ne mémoriser que $V(s)$ + calcul simple
- α est constant ou décroissant lentement

Méthode des différences temporelles

1: **procedure** TemporalDifference
2: Initialiser $V(s)$ arbitrairement et π à la politique à évaluer.
3: **repeat**
4: **for all** *épisode* **do**
5: Partir de s
6: **repeat**
7: **for all** *étape de l'épisode* **do**
8: $a \leftarrow$ l'action donnée par π pour l'état s
9: Executer a, recevoir r et s'
10: $V^\pi(s) \leftarrow V^\pi(s) + \alpha[r + \gamma V^\pi(s') - V^\pi(s)]$
11: $s \leftarrow s'$
12: **until** s est terminal
13: **until** critère d'arrêt

Amélioration de politique

SARSA : méthode "sur politique"

1. Choisir l'action a selon une politique suivie *presque* tout le temps (procédure ε-*gloutonne*)
2. Après observation de s' et r, mettre à jour la valeur d'utilité

$$Q^\pi(s,a) \leftarrow Q^\pi(s,a) + \alpha[r + \gamma Q^\pi(s',a') - Q^\pi(s,a)]$$

Amélioration de politique

Q-learning : méthode "hors politique"

1. Choisir l'action a avec une procédure ε-*gloutonne*
2. Après observation de s' et r, mettre à jour la valeur d'utilité

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a' \in \mathcal{A}} Q(s',a') - Q(s,a)]$$

Exploration -vs- Exploitation

Situation

Je déménage dans une ville inconnue. J'aime manger au restaurant, mais je n'en connais aucun. Je les essaye tous une fois :

1. je n'en favorise aucun, et continue de choisir à l'aveugle : **Exploration pure**
2. je les note tous un par un, puis je ne vais plus qu'au meilleur : **Exploitation pure**

Il vaut mieux trouver un compromis entre exploration et exploitation.

Mais comment ? $\Rightarrow$ résoudre un problème d'optimisation

Exploration -vs- Exploitation

Situation

Je déménage dans une ville inconnue. J'aime manger au restaurant, mais je n'en connais aucun. Je les essaye tous une fois :

1. je n'en favorise aucun, et continue de choisir à l'aveugle : **Exploration pure**
2. je les note tous un par un, puis je ne vais plus qu'au meilleur : **Exploitation pure**

Il vaut mieux trouver un compromis entre exploration et exploitation.

Mais comment ? $\Rightarrow$ résoudre un problème d'optimisation

Exploration -vs- Exploitation

Situation

Je déménage dans une ville inconnue. J'aime manger au restaurant, mais je n'en connais aucun. Je les essaye tous une fois :

1. je n'en favorise aucun, et continue de choisir à l'aveugle : **Exploration pure**
2. je les note tous un par un, puis je ne vais plus qu'au meilleur : **Exploitation pure**

Il vaut mieux trouver un compromis entre exploration et exploitation.

Mais comment ? $\Rightarrow$ résoudre un problème d'optimisation

Exploration -vs- Exploitation

Situation

Je déménage dans une ville inconnue. J'aime manger au restaurant, mais je n'en connais aucun. Je les essaye tous une fois :

1. je n'en favorise aucun, et continue de choisir à l'aveugle : **Exploration pure**
2. je les note tous un par un, puis je ne vais plus qu'au meilleur : **Exploitation pure**

Il vaut mieux trouver un compromis entre exploration et exploitation.
Mais comment ? $\Rightarrow$ résoudre un problème d'optimisation

Problème des bandits à bras mutliples

Définition

- Il existe un ensemble de K bras, chacun défini par une distribution de récompense ν_k (dans $[0, 1]$) de loi inconnue
- À chaque pas de temps t, l'agent doit choisir un bras k_t. Il reçoit une récompense $r_t \overset{i.i.d}{\sim} \nu_{k_t}$
- **But** : trouver une politique de sélection des bras de manière à maximiser la somme des récompenses sur une durée donnée

Exemple de méthodes de résolution

- Méthode ε-greedy / non dirigée (*i.e.*,évalue les actions)
- Méthode basée sur la récence / dirigée (*i.e.*,+ mémoire)
- **Upper Confidence Bound**

Problème des bandits à bras mutliples

Définition

- Il existe un ensemble de K bras, chacun défini par une distribution de récompense ν_k (dans $[0, 1]$) de loi inconnue
- À chaque pas de temps t, l'agent doit choisir un bras k_t. Il reçoit une récompense $r_t \overset{i.i.d}{\sim} \nu_{k_t}$
- **But** : trouver une politique de sélection des bras de manière à maximiser la somme des récompenses sur une durée donnée

Exemple de méthodes de résolution

- Méthode ε-greedy / non dirigée (*i.e.*,évalue les actions)
- Méthode basée sur la récence / dirigée (*i.e.*,+ mémoire)
- **Upper Confidence Bound**

L'algorithme UCB

procedure UCB

Initialisation : Jouer chaque bras une fois

repeat

Jouer le bras j qui maximise $\overline{x}_j + \sqrt{\frac{2 \ln n}{T_j(n)}}$

until fin du jeu

Où $\overline{x}_j$ est le renforcement moyen obtenu en jouant le bras j, $T_j(n)$ le nombre de fois où le bras j a été joué et n le nombre total de tirage jusque là.

Apprentissage automatique

apprendre par renforcement

Charles Prud'homme

Charles.Prudhomme@imt-atlantique.fr

TASC (CNRS/IMT Atlantique)

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom