

Apprentissage par renforcement

Stéphane Airiau

Université Paris Dauphine

Pour les deux algorithmes vus précédemment ("iteration sur les valeurs" et "iteration sur les politiques"), on devait connaître :

- le modèle de transition $T_{ss'}^a$
- le modèle de récompense R_s

Aujourd'hui, on va voir des méthodes qui **ne** nécessitent **pas** la connaissance des ces modèles.

- ⇒ seule l'expérience va guider le choix
- ⇒ véritablement de l'apprentissage

Environnement épisodique

On va se placer seulement dans des PDMs épisodique :

- chaque épisode doit se terminer
- on va apprendre d'un épisode en entier
- un épisode : une partie de black jack

1. Méthodes Monte Carlo

- Evaluation d'une politique
- Estimation de la valeur des actions
- Contrôle par Monte Carlo ("on policy" et "off policy")

2. Temporal differences

- Evaluation d'une politique
- Estimation de la valeur des actions
- Sarsa ("on policy") et Q-learning ("off policy")

Méthode Monte Carlo : Evaluation d'une politique

- apprendre v_π à partir des épisodes en suivant une politique π
- On veut apprendre la valeur *à long terme*
Pour un épisode qui se termine à l'itération k , on a

$$G_t = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{k-t} R_{k+t}$$

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

- on va utiliser l'expérience de l'agent pour estimer cette valeur pour chaque état
- estimer la valeur d'être passé par cet état lors des précédents épisodes

Algorithme "première visite"

```
1 |  $v \in \mathbb{R}^n$ 
2 |  $n \in \mathbb{N}^n$ 
3 |  $R \in \mathbb{R}^n$ 
4 | initialise  $v(s) = 0$  pour chaque état  $s \in S$ 
6 | initialise  $n(s) = 0$  pour chaque état  $s \in S$ 
7 | initialise  $n(s) = 0$  pour chaque état  $s \in S$ 
8 |
9 | Répète éternellement
10 |   Simule un épisode en suivant la politique  $\pi$ 
11 |   Pour chaque état  $s$  qui apparaît dans l'épisode
12 |     Pour la première itération  $t$  où  $s$  est visité dans l'épisode
13 |        $R(s) \leftarrow R + G_t$ 
14 |        $n(s) \leftarrow n(s) + 1$ 
15 |        $v(s) \leftarrow \frac{R(s)}{n(s)}$ 
```

chaque valeur de G_t est un échantillon tiré de manière indépendante et identiquement distribué.

↪ avec la loi des grands nombres, on a

$$\lim_{n(s) \rightarrow \infty} v(s) = v_{\pi}(s)$$

Algorithme "chaque visite"

```
1 |  $v \in \mathbb{R}^n$ 
2 |  $n \in \mathbb{N}^n$ 
3 |  $R \in \mathbb{R}^n$ 
4 | initialise  $v(s) = 0$  pour chaque état  $s \in S$ 
6 | initialise  $n(s) = 0$  pour chaque état  $s \in S$ 
7 | initialise  $R(s) = 0$  pour chaque état  $s \in S$ 
8 |
9 | Répète éternellement
10 |   Simule un épisode en suivant la politique  $\pi$ 
11 |   Pour chaque itération  $t$  qui visite l'état  $s$ 
12 |      $R(s) \leftarrow R + G_t$ 
13 |      $n(s) \leftarrow n(s) + 1$ 
14 |      $v(s) \leftarrow \frac{R(s)}{n(s)}$ 
```

Ici, chacun des échantillons n'est pas forcément indépendant des autres.
Mais on a quand même convergence vers $v_\pi(s)$.

Très différent de l'utilisation de la programmation dynamique

- toutes les transitions possibles / seulement les transitions de l'épisode
- une seule transition / toutes les transitions de l'épisode
- l'estimation de chaque état est fait de manière indépendante / l'estimation d'un état dépend de l'estimation des autres états
- ➡ l'estimation est indépendante du nombre d'états $|S|$.
- ➡ on peut partir d'un état et faire des simulations pour apprendre ces états sans se soucier des autres

Evaluation de la valeur des actions

- Si on n'a pas le modèle $T_{ss'}^a$, on a beau avoir $v_\pi(s)$, on ne peut pas déduire l'action optimale!
- Rappel $q_\pi(s,a)$ estime la valeur à long terme de prendre l'action a puis de suivre la politique π .
- même stratégie "première visite" et "chaque visite" possible

Evaluation de la valeur des actions

- Si on n'a pas le modèle $T_{ss'}^a$, on a beau avoir $v_\pi(s)$, on ne peut pas déduire l'action optimale!
- Rappel $q_\pi(s,a)$ estime la valeur à long terme de prendre l'action a puis de suivre la politique π .
- même stratégie "première visite" et "chaque visite" possible
- petit problème : si π est déterministe, on n'a pas la valeur de toutes les paires (état, action)
Même si π est stochastique, on introduirait peut-être un biais!

Evaluation de la valeur des actions

- Si on n'a pas le modèle $T_{ss'}^a$, on a beau avoir $v_\pi(s)$, on ne peut pas déduire l'action optimale!
- Rappel $q_\pi(s,a)$ estime la valeur à long terme de prendre l'action a puis de suivre la politique π .
- même stratégie "première visite" et "chaque visite" possible
- petit problème : si π est déterministe, on n'a pas la valeur de toutes les paires (état, action)
Même si π est stochastique, on introduirait peut-être un biais!
- une stratégie : "exploring starts"
on tire au hasard une paire (état, action) pour l'état initial et on utilise "première visite"

- Si on n'a pas le modèle $T_{ss'}^a$, on a beau avoir $v_\pi(s)$, on ne peut pas déduire l'action optimale!
- Rappel $q_\pi(s,a)$ estime la valeur à long terme de prendre l'action a puis de suivre la politique π .
- même stratégie "première visite" et "chaque visite" possible
- petit problème : si π est déterministe, on n'a pas la valeur de toutes les paires (état, action)
Même si π est stochastique, on introduirait peut-être un biais!
- une stratégie : "exploring starts"
on tire au hasard une paire (état, action) pour l'état initial et on utilise "première visite"
- Evidemment, ceci est problématique pour des interactions avec un environnement réel (on ne peut pas forcément choisir l'état initial !!)

- on conserve d'avoir une approximation de la fonction de valeurs et de la politique
 - la fonction de valeur approxime de mieux en mieux la politique courante
 - mais on améliore la politique courante
 - à la limite, le processus va converger vers l'optimal

- wishful thinking :
 - utiliser une infinité d'épisode pour estimer $q_{\pi}(s,a)$ avec "exploring starts"
 - améliore de façon gloutonne la politique π comme dans itération des politiques
 - même garantie de convergence que pour itération des politiques

- on conserve d'avoir une approximation de la fonction de valeurs et de la politique
 - la fonction de valeur approxime de mieux en mieux la politique courante
 - mais on améliore la politique courante
 - à la limite, le processus va converger vers l'optimal
- wishful thinking :
 - utiliser une infinité d'épisode pour estimer $q_{\pi}(s,a)$ avec "exploring starts"
 - améliore de façon gloutonne la politique π comme dans itération des politiques
 - même garantie de convergence que pour itération des politiques
- utiliser une convergence à ϵ près pour estimer $q_{\pi}(s,a)$

- on conserve d'avoir une approximation de la fonction de valeurs et de la politique
 - la fonction de valeur approxime de mieux en mieux la politique courante
 - mais on améliore la politique courante
 - à la limite, le processus va converger vers l'optimal
- wishful thinking :
 - utiliser une infinité d'épisode pour estimer $q_{\pi}(s,a)$ avec "exploring starts"
 - améliore de façon gloutonne la politique π comme dans itération des politiques
 - même garantie de convergence que pour itération des politiques
- utiliser une convergence à ϵ près pour estimer $q_{\pi}(s,a)$
- plus extrême : utiliser seulement un épisode avant de faire une amélioration.

Evaluation de la politique optimale : "Monte Carlo Exploring Starts"

```
1 |  $q \in \mathbb{R}^n$ 
2 |  $n \in \mathbb{N}^n$ 
3 |  $R \in \mathbb{R}^n$ 
4 | initialise  $v(s) = 0$  pour chaque état  $s \in S$ 
5 | initialise  $n(s) = 0$  pour chaque état  $s \in S$ 
6 | initialise  $R(s) = 0$  pour chaque état  $s \in S$ 
7 |
8 | Répète éternellement
9 |   Tire aléatoirement une paire  $(s_0, a_0) \in S \times A$ 
10 |   Simule un épisode en suivant la politique  $\pi$  en partant de  $(s_0, a_0)$ 
11 |   Pour chaque paire  $(s, a)$  qui est visitée dans l'épisode
12 |     Si la première occurrence de  $(s, a)$  est à l'instant  $t$ 
12 |        $R(s, a) \leftarrow R(s, a) + G_t$ 
13 |        $n(s, a) \leftarrow n(s, a) + 1$ 
14 |        $q(s, a) \leftarrow \frac{R(s, a)}{n(s, a)}$ 
15 |     Pour chaque état  $s$  dans l'épisode
16 |        $\pi(s) \leftarrow \arg \max_{a \in A} q(s, a)$ 
```

pas encore de démonstration que la convergence soit garantie!!!
mais empiriquement, ça marche!

"on-policy" Monte Carlo

- estime et améliorer la politique tout en l'utilisant
- utiliser une politique stochastique avec des probabilités strictement positives
 $\pi(s,a) > 0$ "politique soft"
- ➡ graduellement mettre à jour cette politique vers une politique déterministique (et optimale !)
- ajouter de l'**exploration** aléatoire

ex : ϵ -greedy {

- utiliser $\arg \max_{a \in A} q(s,a)$ avec une probabilité $1 - \epsilon$
- tirer une action avec une probabilité uniforme avec une probabilité ϵ

Evaluation de la politique optimale : "on policy Monte Carlo"

1 $q : S \times A \rightarrow \mathbb{R}$

2 $n : S \times A \rightarrow \mathbb{N}$

3 $R : S \times A \rightarrow \mathbb{R}$

3 $\pi : S \times A \rightarrow \Delta(A)$

5 initialise $n(s,a) = 0$ pour chaque état $s \in S$

5 initialise $q(s,a) = 0$ pour chaque état $s \in S$

6 initialise $R(s,a) = 0$ pour chaque état $s \in S$

6 initialise $\pi(s,a) > 0$ pour chaque état $s \in S$ et action $a \in A$

7

8 Répète éternellement

10 Simule un épisode en suivant la politique π

11 Pour **chaque paire** (s,a) qui est visitée dans l'épisode

12 Si la première occurrence de (s,a) est à l'instant t

12 $R(s,a) \leftarrow R(s,a) + G_t$

13 $n(s,a) \leftarrow n(s,a) + 1$

14 $q(s,a) \leftarrow \frac{R(s,a)}{n(s,a)}$

15 Pour chaque état s dans l'épisode

16 $a^* \leftarrow \arg \max_{a \in A} q(s,a)$

15 Pour chaque action a

16
$$\pi(s,a) \leftarrow \begin{cases} 1 - \epsilon + \frac{\epsilon}{|A|} & \text{if } a = a^* \\ \frac{\epsilon}{|A|} & \text{if } a \neq a^* \end{cases}$$

Vérification de l'amélioration

On nomme π' la politique ϵ -greedy.

Pour n'importe quelle politique π , on a $\sum_{a \in A} \left(\pi(s,a) - \frac{\epsilon}{|A|} \right) = 1 - \epsilon$

$$\begin{aligned} q(s, \pi'(s)) &= \sum_{a \in A} \pi'(s,a) q(s,a) \\ &= \frac{\epsilon}{|A|} \sum_{a \in A} q(s,a) + (1 - \epsilon) \max_{a \in A} q(s,a) \\ &= \frac{\epsilon}{|A|} \sum_{a \in A} q(s,a) + (1 - \epsilon) \sum_{a \in A} \frac{\pi(s,a) - \frac{\epsilon}{|A|}}{1 - \epsilon} \max_{a \in A} q(s,a) \\ &\geq \frac{\epsilon}{|A|} \sum_{a \in A} q(s,a) + (1 - \epsilon) \sum_{a \in A} \frac{\pi(s,a) - \frac{\epsilon}{|A|}}{1 - \epsilon} q(s,a) \\ &\geq \frac{\epsilon}{|A|} \sum_{a \in A} q(s,a) + \sum_{a \in A} \pi(s,a) q(s,a) - \frac{\epsilon}{|A|} \sum_{a \in A} q(s,a) \\ &\geq \sum_{a \in A} \pi(s,a) q(s,a) = q_{\pi}(s,a) \text{ On a bien une amélioration! } \square \end{aligned}$$

Evaluer une politique tout en en suivant une autre

- Supposons qu'on suive une politique π' .
- Peut-on calculer v_π pour une autre politique π ?

Evaluer une politique tout en en suivant une autre

- Supposons qu'on suive une politique π' .
- Peut-on calculer v_π pour une autre politique π ?
- oui si $\pi(s,a) > 0 \Rightarrow \pi'(s,a) > 0$

Evaluer une politique tout en en suivant une autre

- Supposons qu'on suive une politique π' .
- Peut-on calculer v_π pour une autre politique π ?
- oui si $\pi(s,a) > 0 \Rightarrow \pi'(s,a) > 0$
- Soit t le moment où on visite pour la première fois l'état s
- Soit R_t la récompense à long terme observée
- Soit $p_t(s)$ et $p'_t(s)$ les probabilités que cette séquence soit effectivement la séquence d'états visités à partir de s en suivant π et π' .
- Soit n_s le nombre d'observation pour l'état s (i.e. le nombre d'épisode où s a été visité)

$$v_\pi(s) = \frac{\sum_{i=1}^{n_s} \frac{p_i(s)}{p'_i(s)} R_i(s)}{\sum_{i=1}^{n_s} \frac{p_i(s)}{p'_i(s)}}$$

Evaluer une politique tout en en suivant une autre

A priori, il faut connaître $p_i(s)$ et $p'_i(s)$.

En fait, il suffit de connaître les deux politiques π et π' .

En effet :

Soit $T_i(s)$ l'itération de fin du $i^{\text{ème}}$ épisode où s est visité.

$$\frac{p_i(s)}{p'_i(s)} = \frac{\prod_{t=1}^{T_i(s)-1} \pi(s_t, a_t) T_{s_t s_{t+1}}^{a_t}}{\prod_{t=1}^{T_i(s)-1} \pi'(s_t, a_t) T_{s_t s_{t+1}}^{a_t}} = \prod_{t=1}^{T_i(s)-1} \frac{\pi(s_t, a_t)}{\pi'(s_t, a_t)}$$

Evaluation de la politique optimale : "off policy Monte Carlo"

On sépare ici

- la politique du comportement courant
- la politique que l'on cherche à optimiser : la politique estimée
- On utilise la technique précédente pour améliorer la politique estimée
- le choix de la politique courante va rendre la convergence plus ou moins rapide



Richard Sutton and Andrew Barto
Reinforcement learning : An introduction.
Cambridge : MIT press, 1998.