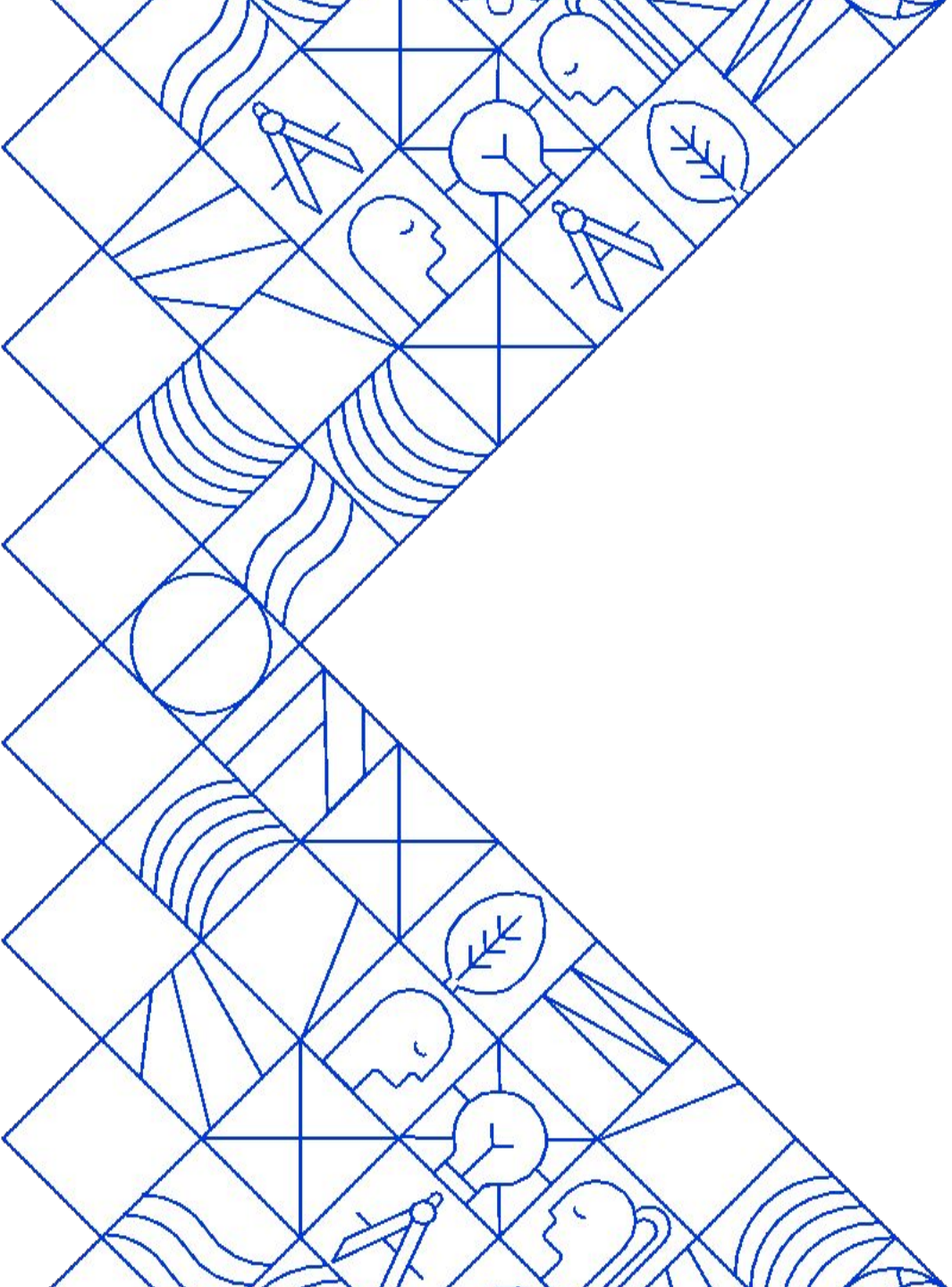


ENSTA

INSTITUT POLYTECHNIQUE DE PARIS

Introduction à l'apprentissage par renforcement

Cours 3 : Algorithmes basés sur la politique

Zhi YAN
28 janvier 2025

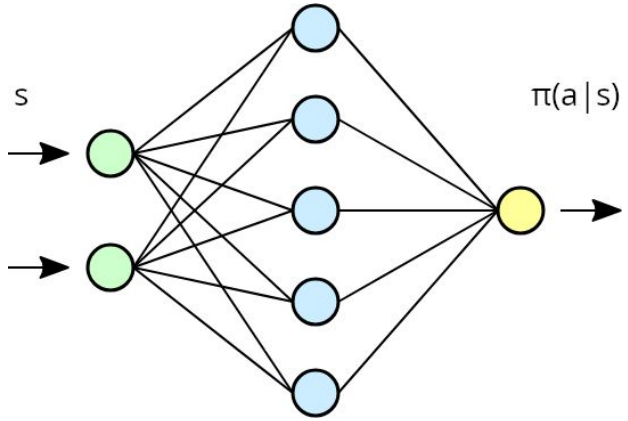
Policy-based Algorithms

- Apprendre directement la politique “π(a|s)” (la probabilité d’entreprendre l’action “a” étant donné l’état “s”).
- Pour chaque état “s”, la politique “π(a|s)” renvoie une distribution de probabilité sur les actions.

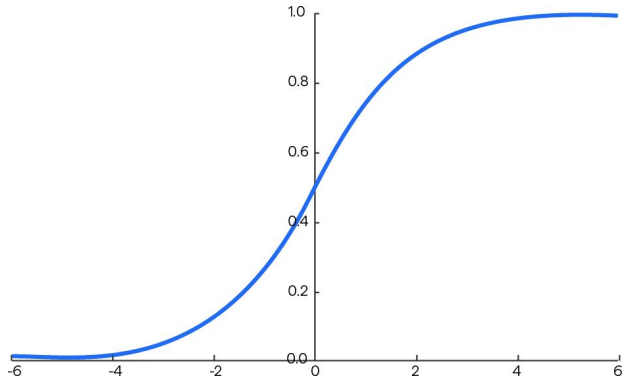
Policy-based Algorithms

θ est le paramètre de la politique

- Pour représenter, apprendre et optimiser une politique (dans un ordinateur), elle doit être paramétrée, notée $\pi_\theta(a|s)$.



- Par exemple :
  - Réseau neuronal (θ correspond aux poids)
  - Fonction linéaire, par exemple $\pi(a|s)$ $= \text{softmax}(\theta^T \varphi(s, a))$, où θ est le paramètre et φ(s, a) est le vecteur de caractéristiques état-action.



Policy-based Algorithms

- **L'objectif : trouver une politique optimale π* qui maximise le rendement cumulé.**
- **Les fonctions objectives couramment utilisées :**
  - Valeur de l'état de départ

$$J(\theta) = V^{\pi_\theta}(s_0) = E_{\pi_\theta}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0]$$

  - Rendement moyen

$$J(\theta) = \lim_{T \to \infty} (1/T) E_{\pi_\theta}[\sum_{t=0}^{T} r_t]$$

  - Retour moyen par étape

$$J(\theta) = E_{s \sim \rho\pi_\theta, a \sim \pi_\theta}[r(s,a)]$$

où $\rho\pi_\theta(s)$ est la distribution d'état sous la politique π_θ

ENSTA IP PARIS

Policy-based Algorithms

Caractéristique	Value-based algorithms	Policy-based algorithms
Objectif d'apprentissage	Fonction de valeur (V(s) ou Q(s, a))	Politique (π(s))
Type de sortie	La valeur d'un état ou d'une action	La distribution de probabilité sur les actions possibles dans un état donné
Problème applicable	Espace d'action discret, petit espace d'état	Espace d'action continu, espace d'état à haute dimension
Avantages	Rendement élevé et bonne convergence	Efficaces dans les espaces d'actions continus, stratégies directes
Inconvénients	Inefficaces dans les espaces d'actions continus, stratégies indirectes	Variance élevée et lente convergence
Exemples	Q-Learning, SARSA	REINFORCE, Actor-Critic

ENSTA IP PARIS

Policy Gradient

- Un terme général désignant une famille de méthodes.
- “Gradient” (de manière abstraite) : la direction dans l’espace des paramètres de politique où la fonction change le plus rapidement.
- “Gradient” (plus précisément) : le vecteur de dérivées partielles de la fonction objective (généralement la récompense cumulative attendue) par rapport aux paramètres de la politique.

Policy Gradient

- La fonction objective :

$$J(\theta) = E_{\tau \sim \pi_\theta}[R(\tau)]$$

$E_{\tau \sim \pi\theta}$ représente la valeur attendue de toutes les trajectoires possibles sous la politique π_θ.

R(τ) représente la récompense cumulée de la trajectoire τ.

$$R(\tau) = \sum_{t=0}^{T} r_{t+1}$$

Le “Gradient” doit prendre en compte une séquence d’états, d’actions et de récompenses de l’interaction de l’agent avec l’environnement, appelée “trajectory”, notée τ, $\tau = (s_0, a_0, r_1, s_1, a_1, r_2, s_2, \ldots, s_T)$.

ENSTA IP PARIS

Policy Gradient

- “Gradient” : le gradient de J(θ) par rapport à θ, noté :

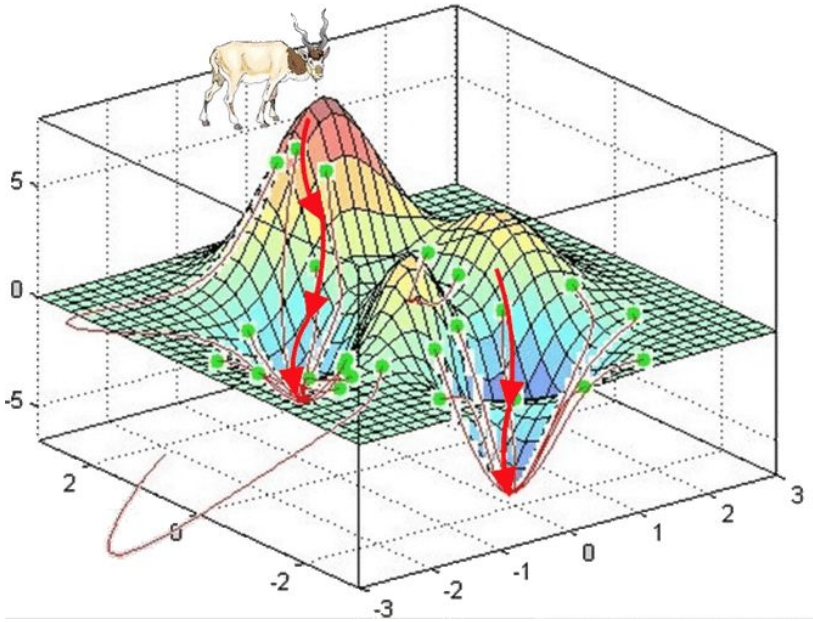
$$\nabla_\theta J(\theta)$$

Dans quelle direction dans l’espace des paramètres de politique l’ajustement de θ augmentera le plus rapidement la récompense cumulative attendue J(θ).

Policy Gradient

- Policy Gradient : utiliser l'ascension du gradient pour mettre à jour les paramètres de la politique et ainsi améliorer les performances de la politique :

$$\nabla_\theta J(\theta) = E_{\tau \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(\tau) R(\tau)]$$



ENSTA
IP PARIS

Policy Gradient

1: Initialiser les paramètres de la politique θ

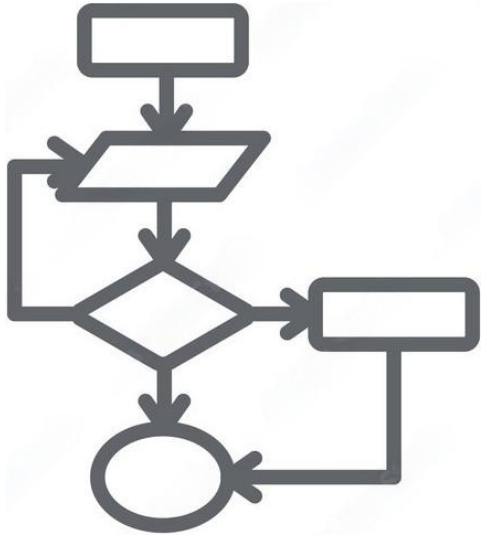
2: **Répéter** (pour chaque épisode) :

3: Utiliser la politique actuelle π_θ pour interagir avec l'environnement et collecter des données de trajectoire

4: Sur la base des données collectées, calculez le gradient $\nabla_\theta J(\theta)$ de la fonction objective $J(\theta)$

5: $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$ // α est le taux d'apprentissage

6: **Jusqu'à ce que** la condition de résiliation soit remplie



REINFORCE

- Un algorithme spécifique de “Policy Gradient”.
- Ronald J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning.” *Machine Learning*, 8:229-256, 1992.
- Utiliser une méthode de Monte Carlo pour estimer le récompense “R(τ)”, c-à-d le récompense utilisant la trajectoire complète.

REINFORCE

- Pourquoi “trajectoire complète” ?
- La méthode de Monte Carlo nécessite des échantillons complets pour une estimation efficace.
- La trajectoire complète d’un épisode doit être connue pour calculer la récompense totale de l’épisode.
- La performance de la politique dans l’épisode peut être correctement évaluée.

REINFORCE

La signification de “REINFORCE” :

Renforcer la politique en fonction des résultats de l’épisode afin qu’elle puisse mieux accomplir la tâche à l’avenir.

REINFORCE

- La règle de mise à jour de REINFORCE :

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\theta)$$

optimisation de la politique avec ascension du gradient

where

$$\nabla_\theta J(\theta) = E_{\tau \sim \pi_\theta}[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) R(\tau)]$$

gradient de politique de base

and

$$\nabla_\theta J(\theta) \approx (1/N) \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^i|s_t^i) G_t^i$$

estimation du gradient de politique de base

ENSTA IP PARIS

REINFORCE

- La règle de mise à jour de REINFORCE :

$$\nabla_\theta J(\theta) \approx (1/N) \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t^i | s_t^i) G_t^i$$

estimation du gradient de politique de base

  - N : le nombre d'épisodes
  - T : le nombre d'étape dans chaque épisode
  - G_t^i : la récompense à partir de l'étape t dans le i-ème épisode :

$$G_t^i = \sum_{k=t}^{T} \gamma^{k-t} r_k^i$$

REINFORCE

- Un problème de l'utilisation des données complètes de l'épisode : une variance plus élevée dans l'estimation du gradient.

- Solution 1 : Impliquer une ligne de base :

$$\nabla_\theta J(\theta) = E_{\tau \sim \pi_\theta}[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)(R(\tau) - b(s_t)]$$

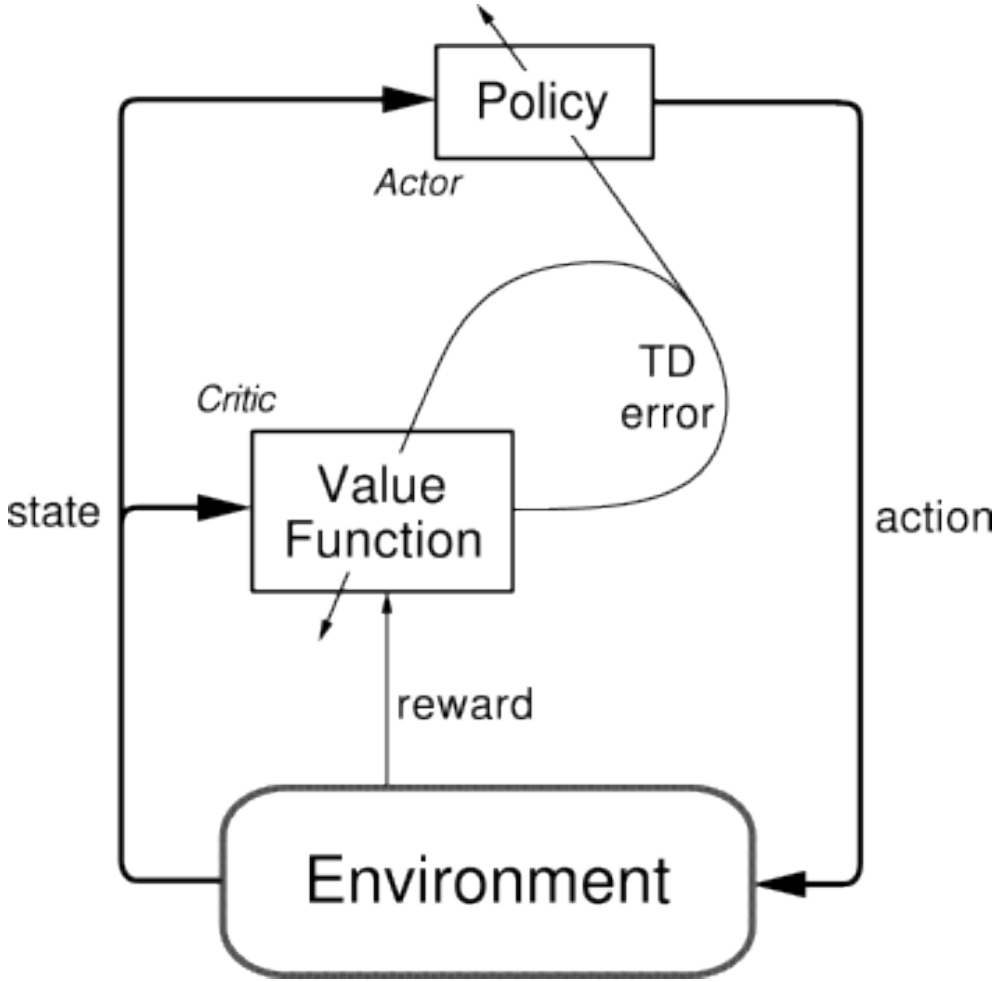
p. ex. la fonction état-valeur $V(s_t)$

- Solution 2 : L'algorithme Actor-Critic.

Actor-Critic

- Ne nécessite pas de données d'épisode complètes.
- Combiner les méthodes basées sur la politique et sur la valeur : surmonter le problème de forte variance dans les méthodes de gradient de politique.
- Pouvoir mettre à jour la politique en ligne.

Actor-Critic



Actor-Critic

Actor

Apprendre une politique $\pi_\theta(a|s)$ qui détermine quelle action doit être prise dans un état donné, considérant “les commentaires” fournis par le critique

Critic

Evaluer la performance de l’acteur et apprendre une fonction de valeur V(s) ou Q(s, a)

Actor-Critic

- Comment l'acteur peut-il intégrer le feedback fourni par le critique dans la mise à jour de sa politique ?
- Solution courante : "TD error" :

$$\delta = r + \gamma V(s') - V(s)$$

or

$$\delta = r + \gamma Q(s', a') - Q(s, a)$$

Actor-Critic

- La règle de mise à jour pour l'acteur est similaire à la méthode du gradient de politique, mais utilise l'erreur TD δ au lieu de la récompense totale R(τ) :

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta \log \pi_\theta(a_t|s_t)\delta$$

Fin