

Apprentissage par renforcement appliqué

Notions fondamentales

[revision 3.4]

Brahim Chaib-draa

Brahim.Chaib-Draa@ift.ulaval.ca



Université Laval

2020-07-01

- 1** Vue d'ensemble
- 2** Formalisme et outil de l'apprentissage par renforcement
- 3** L'univers du point de vue de l'agent
- 4** Pour aller plus loin

Vue d'ensemble

1 Vue d'ensemble

- Le paradigme d'apprentissage par renforcement
- Environnement : l'univers de l'agent

2 Formalisme et outil de l'apprentissage par renforcement

3 L'univers du point de vue de l'agent

4 Pour aller plus loin

Vue d'ensemble

Le paradigme d'apprentissage par renforcement

Qu'est-ce que l'apprentissage par renforcement ?

De façon informelle :

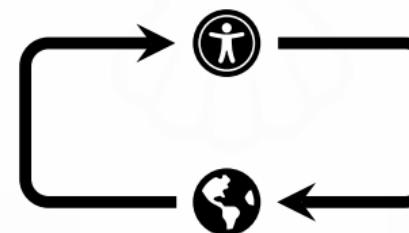
- ❶ – Un **agent** : une entité qui **observe, prend des décisions, agit et apprend de ses actions**
- ❷ – Un **environnement** : l'univers où l'agent existe. Pour chaque action prise par l'agent, l'univers **renvoie une rétroaction** 🎉, c.-à-d. un signal de récompense 👍 ou 👎
- ❸ – Un **but** (pour l'agent) : l'état final à atteindre



Qu'est-ce que l'apprentissage par renforcement ?

De fa on informelle :

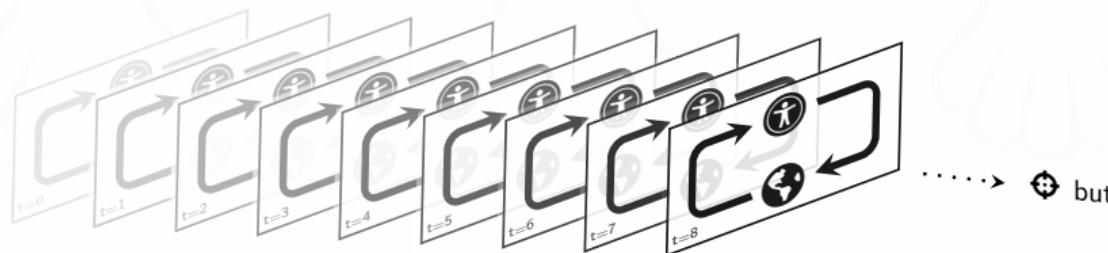
- ⦿ – Un **agent** : une entit  qui **observe, prend des d cisions, agit et apprend de ses actions**
- ⦿ – Un **environnement** : l'univers o  l'agent existe. Pour chaque action prise par l'agent, l'univers **renvoie une r troaction** 🎉, c.- -d. un signal de **r compense** 🎊 ou 🚫
- ⦿ – Un **but** (pour l'agent) : l' tat final  atteindre



Qu'est-ce que l'apprentissage par renforcement ?

De fa on informelle :

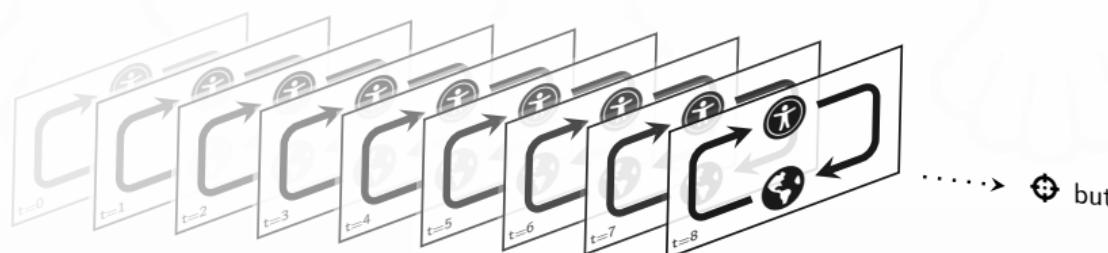
- Un **agent** : une entit  qui **observe, prend des d cisions, agit et apprend de ses actions**
- Un **environnement** : l'univers o  l'agent existe. Pour chaque action prise par l'agent, l'univers **renvoie une r troaction** 🎉, c.- -d. un signal de **récompense** 🎊 ou 🙅
- Un **but** (pour l'agent) : l' tat final   atteindre



Qu'est-ce que l'apprentissage par renforcement ?

De fa on informelle :

- Un **agent** : une entit  qui **observe, prend des d cisions, agit et apprend de ses actions**
- Un **environnement** : l'univers o  l'agent existe. Pour chaque action prise par l'agent, l'univers **renvoie une r troaction** 🎉, c.- -d. un signal de **r compense** 🎊 ou 🎉
- Un **but** (pour l'agent) : l' tat final   atteindre



★ – L'objectif (informel) de l'apprentissage par renforcement :

Que l'**agent** trouve par lui-m me la **meilleure strat gie** (une **politique optimale**) pour atteindre ce **but** en se basant uniquement sur les **r compenses**.

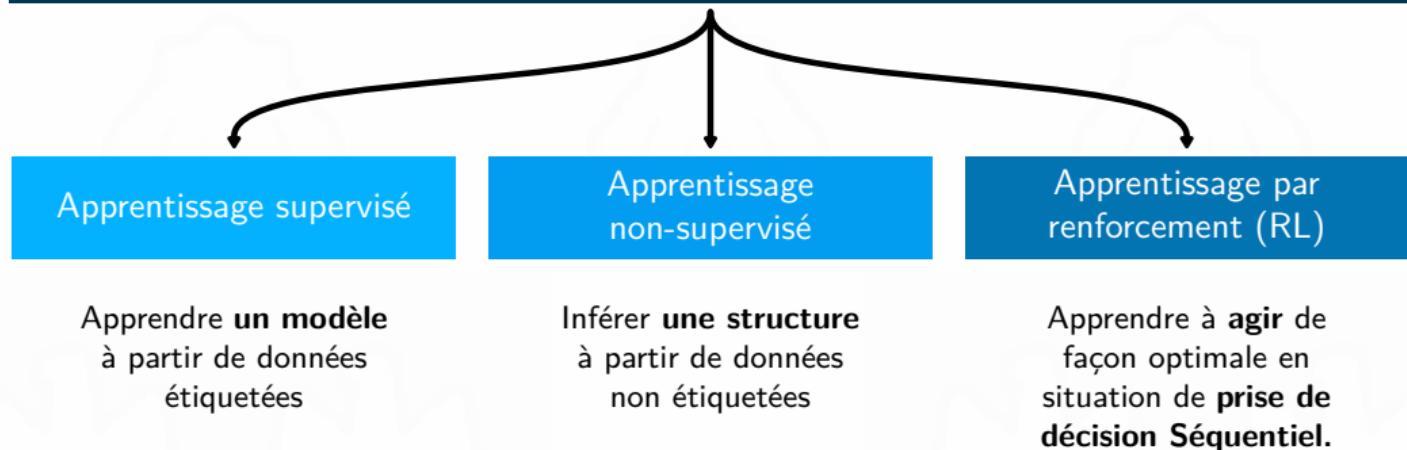
Apprentissage machine
(ML)

Probl me de D cision S quentiel
(PDS)

Apprentissage par renforcement (RL)

Situer l'apprentissage par renforcement par rapport aux autres paradigmes d'apprentissage machine

L'apprentissage machine (ML)



En quoi l'apprentissage par renforcement est différent des autres paradigmes ?

- **Pas d'étiquette**, juste des **observations** et un signal de **récompense**
- Les données sont **séquentielles** et **ordonnées** (leur ordre fait une différence),
 - 1 Sauter hors de l'avion
 - 2 compter jusqu'à 10
 - 3 😱 ... ensuite mettre son parachute !
- Les récompenses sont **dépendantes des actions** exécutées,
ex : Arroser des semences de plan de tomate → des belles tomates mûres en santé
- ⚠ Les conséquences d'une action ne sont pas nécessairement immédiates,
ex : J'ensemence des plants de tomate → Yééé !! Les tomates sont mûres ... on maaaange !!
- Problème :** Comment savoir si une récompense r est causée par l'action a ?

En quoi l'apprentissage par renforcement est différent des autres paradigmes ?

- **Pas d'étiquette**, juste des **observations** et un signal de **récompense**
 - Les données sont **séquentielles** et **ordonnées** (leur ordre fait une différence),
 - 1 Sauter hors de l'avion
 - 2 compter jusqu'à 10
 - 3 😱 ... ensuite mettre son parachute !
 - Les récompenses sont **dépendantes des actions** exécutées,
ex : Arroser des semences de plan de tomate → des belles tomates mûres en santé
 - ⚠ Les conséquences d'une action ne sont pas nécessairement immédiates,
ex : J'ensemence des plants de tomate → Yééé !! Les tomates sont mûres ... on maaaange !!
- Problème :** Comment savoir si une récompense r est causée par l'action a ?

En quoi l'apprentissage par renforcement est différent des autres paradigmes ?

- **Pas d'étiquette**, juste des **observations** et un signal de **récompense**
 - Les données sont **séquentielles** et **ordonnées** (leur ordre fait une différence),
 - 1 Sauter hors de l'avion
 - 2 compter jusqu'à 10
 - 3 ... ensuite mettre son parachute !
 - Les récompenses sont **dépendantes des actions** exécutées,
ex : Arroser des semences de plan de tomate → des belles tomates mûre en santé
 - ⚠ Les conséquences d'une action ne sont pas nécessairement immédiates,
ex : J'ensemence des plants de tomate → Yééé !! Les tomates sont mûres ... on maaaanne !
- Problème :** Comment savoir si une récompense r est causée par l'action a ?

En quoi l'apprentissage par renforcement est différent des autres paradigmes ?

- **Pas d'étiquette**, juste des **observations** et un signal de **récompense**
 - Les données sont **séquentielles** et **ordonnées** (leur ordre fait une différence),
 - 1 Sauter hors de l'avion
 - 2 compter jusqu'à 10
 - 3 ... ensuite mettre son parachute !
 - Les récompenses sont **dépendantes des actions** exécutées,
ex : Arroser des semences de plan de tomate → des belles tomates mûre en santé
 - ⚠ Les conséquences d'une action ne sont pas nécessairement immédiates,
ex : J'ensemence des plants de tomate → Yééé !! Les tomates sont mûres ... on maaaanne !
- Problème :** Comment savoir si une récompense r est causée par l'action a ?

En quoi l'apprentissage par renforcement est différent des autres paradigmes ?

- **Pas d'étiquette**, juste des **observations** et un signal de **récompense**
 - Les données sont **séquentielles** et **ordonnées** (leur ordre fait une différence),
 - 1 Sauter hors de l'avion
 - 2 compter jusqu'à 10
 - 3 ... ensuite mettre son parachute !
 - Les récompenses sont **dépendantes des actions** exécutées,
ex : Arroser des semences de plan de tomate → des belles tomates mûre en santé
 - ▲ Les conséquences d'une action ne sont pas nécessairement immédiates,
ex : J'ensemence des plants de tomate → Yééé !! Les tomates sont mûres ... on maaaannge !!

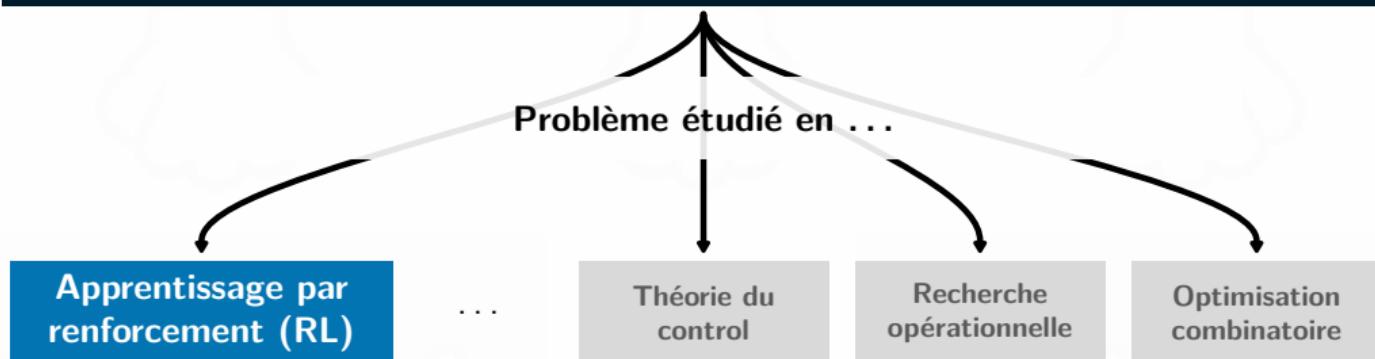
Idée clé à retenir :

En RL, les données **ne sont pas indépendantes et identiquement distribuées**

(contrairement à l'assumption qu'on fait en apprentissage profond)

Situer le RL par rapport aux autres approches de r solution de Probl me de D cision S quentiel

Probl me de D cision S quentiel (PDS)



En quoi l'apprentissage par renforcement se distingue-t-il des autres approches ?

Cas o  l'approche RL se distingue :

- 1 Aucun **mod le** de l'environnement n'existe et la seule fa on de **recueillir de l'information** sur l'environnement est **par interaction** avec celui-ci ;
- 2 Un **mod le** de l'environnement est disponible, mais il est **incomplet et/ou inexact** ;
- 3 Un **mod le** de l'environnement est disponible, mais il n'existe **pas de solution analytique** ;

Composante cl  du RL

- Echantillonnage et Optimization ;
- Utilisation d'approximateur de fonction ;

 – Un **mod le de l'environnement** : C'est la repr sentation abstraite que l'agent poss de du comportement de l'univers
Ex. : les r gles de la physique , les r gles d'un jeu  

Situer le RL par rapport aux autres approches de r solution de Probl me de D cision S quentiel

Probl me de D cision S quentiel (PDS)

Pas de mod le de l'environnement
(ou mod le imparfait)
et/ou
pas de solution analytique.

Un mod le exact
de l'environnement est disponible
et
une solution analytique existe.

Apprentissage par
renforcement (RL)

Th orie du
contrle

Recherche
op rationnelle

Optimisation
combinatoire

...

Vue d'ensemble

Environnement : l'univers de l'agent

- **Simulateur** : Simulateur de physique, jeux vidéo, ...
 - **Robotique** : Pilotage d'aéronef, manipulation d'un bras robot, ...
 - **Ressource r elle** : Calculateur en grappe (*cluster*), syst me de contr le industriel ...
- ⋮

Exemple d'environnement : simulateur

Simulateur de physique pour l'étude de la locomotion :

TerrainRL Simulator [1]

▶ Publication

▶ Projet

▶ GitHub

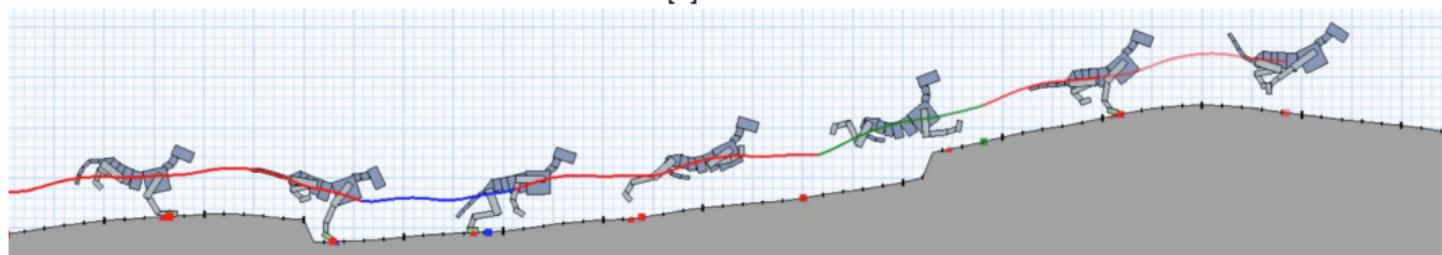


Image : University of British Columbia & UC Berkley

▶ Vid o

- *Terrain-Adaptive Locomotion Skills Using Deep Reinforcement Learning (MACE, 2016) [2]*

▶ Publication

Exemple d'environnement : simulateur

L'environnement de recherche bas e la suite de jeux Atari de OpenAI Gym

▶ OpenAI - Atari



Image : OpenAI

- *Playing Atari with Deep Reinforcement Learning (DQN, 2013) [3]* ▶ Vid o
- *Human-level control through deep reinforcement learning (Publication dans la revue Nature) [4]* ▶ GitHub

Exemple d'environnement : simulateur

L'environnement de recherche PySC2 par Deepmind bas e le jeu StarCraft II de *Blizzard Entertainment*.

[PySC2 GitHub](#)

[Vid o: PySC2 feature layer API](#)

[StarCraft II](#)

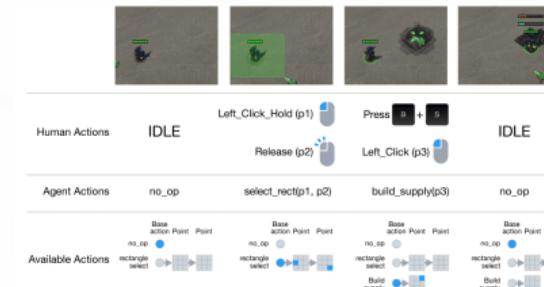


Image : DeepMind

- Deepmind and blizzard open starcraft II ai research environment [Publication](#)
- StarCraft II : A New Challenge for Reinforcement Learning [5] [Publication](#)
- Mastering the Real-Time Strategy Game StarCraft II (AlphaStar, 2019) [Publication](#)
- DeepMind AlphaStar : The inside story [Vid o \(5 min\)](#)

Exemple d'environnement : robotique

Pilotage autonome d'a閞onef :

Stanford University autonomous helicopter

▶ Projet



Image : Stanford Autonomous Helicopter

▶ Vid o

- *An Application of Reinforcement Learning to Aerobatic Helicopter Flight (LQR, 2006) [6]*

Exemple d'environnement : robotique

Automatisation de chaine logistique : flexible et r siliente face aux changements

Covariant

► Covariant.ai / Use cases

► Covariant.ai / Research



► Vid o (3:38 min)



► Vid o (9 min)

Image : covariant.ai

- *AI Helps Warehouse Robots Pick Up New Tricks (WIRED)* ► Magazine
- *AI-powered robot warehouse pickers are now ready to go to work (MIT Technology Review)* ► Magazine
- *Covariant, ABB partner to integrate AI and robots for warehouses (The Robot Report)* ► Magazine

Exemple d'environnement : ressources industrielles

Probl me de gestion des ressources :

- Gestion des tâches dans un calculateur en grappe (*computer cluster*) :

Resource Management with Deep Reinforcement Learning (DeepRM, 2016) [7] [▶ Publication](#)

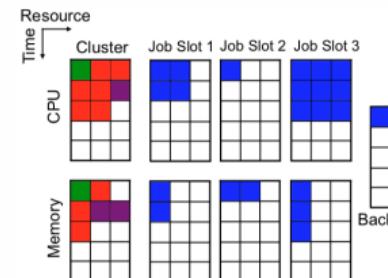


Image : Publication *Resource Management with Deep Reinforcement Learning*, MIT & Microsoft Research

- Gestion de la temp rature d'un m gacentre de donn es :

DeepMind AI Reduces Google Data Centre Cooling Bill by 40% [▶ Publication](#)

Particularit  des diff rents types d'environnements

Probl mes et enjeux

Simulateur :

- oblige une planification   long terme (d lais d'observation de cause   effet) ;
- oblige une strat gie efficace d'exploration/exploitation ;
- espace d' tats/d'observation est en haute dimension ;
- temps r el (Atari, Starcraft, Dota 2) ;
- information est imparfaite (Atari, Starcraft, Dota 2) ;
- espace d'action est en haute dimension (Starcraft, Dota 2) ;
- coop ration multi agent (Dota 2) ;

Robotique :

- l'espace d'action et d'observation est continu ;
- temps r el ;
- information est imparfaite(l'univers  tant observ  au moyen de senseurs qui sont limit s/bruit ) ;

Formalisme et outil de l'apprentissage par renforcement

1 Vue d'ensemble

2 Formalisme et outil de l'apprentissage par renforcement

- Le langage des Processus de Décision de Markov (MDP)
- Terminologie et notation en RL
- Interaction agent-environnement

3 L'univers du point de vue de l'agent

4 Pour aller plus loin

Formalisme et outil de l'apprentissage par renforcement

Le langage des Processus de Décision de Markov (MDP)

Pourquoi utiliser les MDP ?

- C'est un **formalisme math matique** pour l' tude de Probl me de D cision S quentiel **en contexte d'incertitude**.
- C'est l'outil dont on se sert en RL pour d finir le probl me
- R gle g n rale, on va chercher  r soudre **explicitelement ou implicitement** un MDP.
- *Markovien ?*

Pourquoi utiliser les MDP ?

- C'est un **formalisme math matique** pour l' tude de Probl me de D cision S quentiel **en contexte d'incertitude**.
- C'est l'outil dont on se sert en RL pour d finir le probl me
- R gle g n rale, on va chercher  r soudre **explicitelement ou implicitement** un MDP.
- ***Markovien ?***

Pourquoi utiliser les MDP ?

- C'est un **formalisme math matique** pour l' tude de Probl me de D cision S quentiel **en contexte d'incertitude**.
- C'est l'outil dont on se sert en RL pour d finir le probl me
- R gle g n rale, on va chercher  r soudre **explicitelement ou implicitement** un MDP.
- **Markovien** ? Pour un processus de d cision, cela signifie que **l'environnement est enti rement caract ris  par l' tat pr sent** \implies l'historique n'apporte aucune information suppl mentaire.

Propri t  de Markov

« *Le futur est ind pendant du pass   tant donn  le pr sent* »

$$p(s_{t+1}|s_t, a_t) = p(s_{t+1}|s_1, a_1, \dots, s_t, a_t)$$

Un MDP est d fini par un tuple, $\langle \mathcal{S}, \mathcal{A}, r, p, s_0 \rangle$ tel que :

- \mathcal{S} est l'ensemble des **『tats』** valides
- \mathcal{A} est l'ensemble des **actions** valides
- $r(s, a, s')$ sp cifie la fonction de **récompense**
- $p(s'|s, a)$ sp cifie le **mod le** du syst me
- s_0 est l'**『t at initiale』** du syst me

$$s \in \mathcal{S} \quad \text{Notation explicite : } s \in \mathcal{S}_t$$

$$a \in \mathcal{A} \quad \text{Notation explicite : } a \in \mathcal{A}(s) \text{ ou } a \in \mathcal{A}_t$$

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

$$p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{P}$$



Formalisme et outil de l'apprentissage par renforcement

Terminologie et notation en RL

Notation

Convention dans le cadre du cours :

Terme **notation** [Terme anglophone¹ , Terme alternatif² , ...]

exemple : Trajectoire τ [Trajectory, Episode, Rollout]

Note au lecteur : Soyez avis  que la litt rature en *apprentissage par renforcement* est riche, diverse et influenc e par de multiples domaines d'expertise : math matiques, th orie du contr le, recherche op rationnelle, statistique, science informatique, robotique, neuroscience ...

Il n'existe pas de notations unifi es en *RL*   l'heure actuelle et **la terminologie** utilis e pour nommer un m me concept **peut varier d'un auteur   l'autre**.

Recommandation :

- v rifier comment l'auteur d finit la notation pour l'article que vous consultez ;
- portez attention au contexte d'utilisation ;

1. La grande majorit  de la litt rature en *RL* est anglophone

2. Une pr cision sera apport e lorsque la nuance entre deux termes est un d tail important  tant donn  le contexte d'utilisation.

Notation

Convention relativement   la typographie

c est un scalaire

C est un vecteur al atoires (est souvent omise dans la litt rature afin de simplifier la formulation)

\mathbf{c} est la r alisation d'un vecteur al atoires

\mathcal{C} est un espace de probabilit  (l'ensemble des valeurs possibles d'une variable al atoires)

- Temps t [*Time step*] et Horizon T [*Time horizon*]
-  tat s_t [*State*] vs Observation \mathbf{o}_t [*Observation*]
- Action a_t [*Action*] et Politique $\pi(s_t|a_t)$ [*Policy*]
- Trajectoire τ [*Trajectory, Episode, Rollout*]
- Mod le $p(s'|s, a)$ [*Model, Transition function, System dynamic, Dynamic*]
- R compense r_t [*Reward*] et Retour G_t [*Return*]
- L'objectif π^* en RL [*RL Objective*]

Temps et horizon

Temps t [Time step]

- Peut- tre discret ou continue¹
- N'est pas n cessairement temporel. Ex. : une machine t at

Horizon T [Time horizon]

- Le dernier temps t d'une s quence,
Par exemple pour une s quence d'action a_t

$$a_1, a_2, a_3, \dots, a_T$$

- Peut- tre fini ou infini²

1. Voir note de bas de page 2, page 48 de Reinforcement Learning : An introduction [8]

2. Pour le cas horizon-infini $T = \infty$, voir la section : [Pour aller plus loin : Le Retour G sur un horizon infini](#)

Observation vs tat

Observation¹ $\mathbf{o}_t \in \mathcal{O}$ [Observation]

- L'espace d'tat \mathcal{O} est l'ensemble des tats observables \mathbf{o} possibles d'un environnement.
- \mathbf{o} repr sente une partie de l'information disponible sur l'environnement.

$$\text{Eye icon} \subseteq \text{Globe icon} \mapsto \mathbf{o}_t$$

tat $\mathbf{s}_t \in \mathcal{S}$ [State]

- L'espace d'tat \mathcal{S} est l'ensemble des tats \mathbf{s} possibles d'un environnement.
- \mathbf{s} repr sente la totalit  des informations disponibles sur l'environnement.

$$\text{Eye icon} = \text{Globe icon} \mapsto \mathbf{s}_t$$

Si $\mathbf{s}_t = \mathbf{o}_t$ alors on dit que l'environnement est compl tement observ .

Si $\mathbf{s}_t \supset \mathbf{o}_t$ alors on dit que l'environnement est partiellement observ .

1. Mise en garde sur la notation : \mathbf{s}_t est souvent employ  dans la litt rature pour noter un tat partiellement observ  au lieu de \mathbf{o}_t alors qu'il serait techniquement plus juste d'utiliser \mathbf{s}_t . Porter une attention particuli re au contexte d'utilisation. 

Action

Action $a_t \in \mathcal{A}$ [Action]

- L'espace d'action \mathcal{A} est l'ensemble des actions a valides d'un environnement.
- L'espace d'action peut-être :

discret ex. : Les actions possibles dans le jeu Pac-Man ($\uparrow, \downarrow, \leftarrow, \rightarrow$)

continu ex. : Les commandes de direction et de profondeur d'un avion 

Politique

Politique π [Policy]

- C'est une strat gie, une fa on d'agir qui indique quelle action a_t prendre en fonction du contexte
Ex. « le client a toujours raison », « la meilleure d fensive c'est l'offensive », ...
- La politique peut  tre soit de type **d terministe** : $a_t = \pi(s_t)$

Exemple :

$$\begin{aligned}\pi_{greedy} &: \mathcal{S} \longrightarrow \mathcal{A} \\ s &\longmapsto \pi_{greedy}(s_t) \doteq \arg \max_a Q^\pi(s_t, a)\end{aligned}$$

soit de type **stochastique** : $a_t \sim \pi(\cdot | s_t)$

Exemple :

$$\begin{aligned}\pi_{gaussian} &: \mathcal{A} \times \mathcal{S} \longrightarrow [0, 1] \\ (a, s) &\longmapsto \pi_{gaussian}(a_t | s_t) \doteq \mathbb{P}_{A \sim \mathcal{N}(\mu(s), \Sigma)} [A_t = a | S_t = s]\end{aligned}$$

Trajectoire

Trajectoire τ [Trajectory, Episode, Rollout]

- C'est une s quence compl te d' tats s et d'actions a jusqu'  terminaison, peu importe l'issue.

$$\tau = s_1, a_1, r_2, s_2, a_2, r_3, \dots, s_T, a_T, r_{T+1}$$

Mod le

Mod le $p(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ [Model p , Transition function T , System dynamic, Dynamic]

- Ce sont les r gles qui gouvernent l'environnement.
Ex. : les r gles de la physique , les r gles d'un jeu  , ...
- Dans le contexte du **RL sans-mod le**, puisque le mod le est inconnu, alors celui-ci est **mod lis  comme une distribution de probabilit ** qui satisfait la propri t  de Markov

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

$$(\mathbf{s}', r, \mathbf{s}, \mathbf{a}) \longmapsto p(\mathbf{s}', r | \mathbf{s}, \mathbf{a}) \doteq \mathbb{P}[S_{t+1} = \mathbf{s}', R_{t+1} = r | S_t = \mathbf{s}, A_t = \mathbf{a}]$$

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

$$(\mathbf{s}', \mathbf{s}, \mathbf{a}) \longmapsto p(\mathbf{s}' | \mathbf{s}, \mathbf{a}) \doteq \mathbb{P}[S_{t+1} = \mathbf{s}' | S_t = \mathbf{s}, A_t = \mathbf{a}]$$

$$= \sum_{r \in \mathcal{R}} \mathbb{P}[S_{t+1} = \mathbf{s}', R_{t+1} = r | S_t = \mathbf{s}, A_t = \mathbf{a}]$$

Remarque : $1 = \sum_{\mathbf{s}' \in \mathcal{S}} p(\mathbf{s}' | \mathbf{s}, \mathbf{a})$ pour tous $\mathbf{s} \in \mathcal{S}$ et $\mathbf{a} \in \mathcal{A}$

Une intuition   propos du mod le p

▲ Probl me : En RL, le mod le de l'environnement est g n ralement inconnu (ou inexact)

Solution : Mod liser l'environnement comme une variable al atoire

Un bon moyen pour visualiser cette approche et comprendre le r le que joue p dans les formules math m tiques en RL est de regarder le comportement des *Processus de Markov*. [► Exemple interactif](#)

Intuition : Imaginez-vous sur un bateau   la d rive avec

s = vos coordonn es g ographiques

et p = les courants marins, l'influence des mar es et du vent ...



Image : Setosa blog

Processus de Markov $\langle \mathcal{S}, p, s_0 \rangle$

- C'est comme un MDP, mais sans action a ni r compense r
- Le passage d'un  tat s   un autre est strictement bas  sur le mod le $p(s'|s)$.

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

R  compense, retour et objectif

Id  e cl   : formaliser la notion de but 

R  compense $r_t \in \mathcal{R} \subset \mathbb{R}$ [Reward]

   C'est le « feedback » imm  diat de l'environnement en cons  quence d'une action.

■ Fonction de r  compense (version   3 arguments, forme la plus courante)

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

$$(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) \longmapsto r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$$

■ Fonction de r  compense (version   2 arguments)

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

$$\begin{aligned} (\mathbf{s}_t, \mathbf{a}_t) \longmapsto r(\mathbf{s}_t, \mathbf{a}_t) &\doteq \mathbb{E}_{(\mathbf{s}', r) \sim p(\cdot | \mathbf{s}, \mathbf{a})} [R_{t+1} = r | S_t = \mathbf{s}, A_t = \mathbf{a}] \\ &= \sum_{r \in \mathcal{R}} \sum_{\mathbf{s}' \in \mathcal{S}} p(\mathbf{s}', r | \mathbf{s}, \mathbf{a}) r(\mathbf{s}, \mathbf{a}, \mathbf{s}') \end{aligned}$$

Notation abr  g  e   : $r_t \doteq r(\mathbf{s}_{t-1}, \mathbf{a}_{t-1}, \mathbf{s}_t)$

★ Observation : L'agent re  oit la r  compense **apr  s** avoir ex  cut   l'action

R  compense, retour et objectif

Id  e cl   : formaliser la notion de but 

Retour G_t [Return]

- La **somme des r  compenses future**   partir du temps t jusqu'   l'horizon T .
- Cas : $T = \text{horizon fini}^1$ (Ce sera l'assumption pour le reste du cours)

$$G_t(\tau) \doteq \sum_{t'=t}^T r(s_{t'}, a_{t'}, s_{t'+1})$$

Remarque : On emploie souvent $G(\tau)$ lorsque qu'on calcule le retour sur une trajectoire compl  te $t = 1, 2, \dots, T$

★ | L'objectif (formel) de l'apprentissage par renforcement [RL objective]

 **Maximiser** le total des r  compenses r accumul   au cours d'une trajectoire τ

$$\begin{aligned}\pi^* &= \arg \max_{\pi} \mathbb{E}_{\tau \sim \pi} [G(\tau)] \\ &= \arg \max_{\pi} \sum_{t=1}^T \mathbb{E}_{(s_t, a_t, s_{t+1}) \sim \pi(\tau)} [r(s_t, a_t, s_{t+1})]\end{aligned}$$

1. Pour le cas horizon-infini $T = \infty$, voir la section : [Pour aller plus loin : Le Retour \$G\$ sur un horizon infini](#)

Retour et *Discount factor*

Id e cl  : les r compenses re ues imm diatement ont plus de valeur que celle re ue plus tard

Retour G_t avec *Discount factor* [Discounted Return]

- La **somme des r compenses future (avec *Discount factor* γ)**  tart du temps t jusqu'  l'horizon T avec $T = \text{horizon fini}^1$

$$\begin{aligned} G_t(\tau) &\doteq \sum_{t'=t}^T \gamma^{t'-t} r(s_{t'}, a_{t'}, s_{t'+1}) \quad \text{avec } 0 \leq \gamma \leq 1 \\ &= r(s_t, a_t, s_{t+1}) + \gamma^1 r(s_{t+1}, a_{t+1}, s_{t+2}) + \gamma^2 r(s_{t+2}, a_{t+2}, s_{t+3}) + \dots \end{aligned}$$

$$\gamma \rightarrow 0 \implies G(\tau) = 1 \cdot r(s_t, a_t, s_{t+1}) \quad \text{« L'agent est  troit d'esprit et ne consid re pas le futur »}$$

$$\gamma \rightarrow 0.5 \implies G(\tau) = 1 \cdot r(s_t, a_t, s_{t+1}) + 0.5 \cdot r(s_{t+1}, a_{t+1}, s_{t+2}) + 0.25 \cdot r(s_{t+2}, a_{t+2}, s_{t+3}) + 0.125 \cdot r(s_{t+3}, a_{t+3}, s_{t+4}) + \dots$$

$$\gamma \rightarrow 1 \implies G(\tau) = 1 \cdot r(s_t, a_t, s_{t+1}) + 1 \cdot r(s_{t+1}, a_{t+1}, s_{t+2}) + 1 \cdot r(s_{t+2}, a_{t+2}, s_{t+3}) + \dots$$

« L'agent ne fait pas de diff rence entre les r compenses pr sent es et futures »

1. Pour le cas horizon-infini $T = \infty$, voir la section : [Pour aller plus loin : Le Retour G sur un horizon infini](#)

Distribution de probabilit  d'une trajectoire

Notation $\tau \sim \pi(\tau)$ | Distribution de probabilit  de la trajectoire τ induite par la politique π

■ Cas politique stochastique :

$$\tau \sim \pi(\tau) \implies s_1 \sim p(\cdot) \quad a_t \sim \pi(\cdot | s_t) \quad s_{t+1} \sim p(\cdot | s_t, a_t)$$

$$\pi(\tau) = \pi(s_1, a_1, \dots, s_T, a_T) = p(s_1) \prod_{t=1}^T p(s_{t+1} | s_t, a_t) \pi(a_t | s_t)$$

■ Cas politique d閞ministe :

$$\tau \sim \pi(\tau) \implies s_1 \sim p(\cdot) \quad a_t = \pi(s_t) \quad s_{t+1} \sim p(\cdot | s_t, a_t)$$

$$\pi(\tau) = \pi(s_1, a_1, \dots, s_T, a_T) = p(s_1) \prod_{t=1}^T p(s_{t+1} | s_t, \pi(s_t))$$

Distribution de probabilit  d'une trajectoire

Notation $\tau \sim \pi(\tau)$ | Distribution de probabilit  de la trajectoire τ induite par la politique π

■ Cas politique stochastique :

$$\tau \sim \pi(\tau) \implies s_1 \sim p(\cdot) \quad a_t \sim \pi(\cdot | s_t) \quad s_{t+1} \sim p(\cdot | s_t, a_t)$$

$$\pi(\tau) = \pi(s_1, a_1, \dots, s_T, a_T) = p(s_1) \prod_{t=1}^T p(s_{t+1} | s_t, a_t) \pi(a_t | s_t)$$

■ Cas politique d閞iministe :

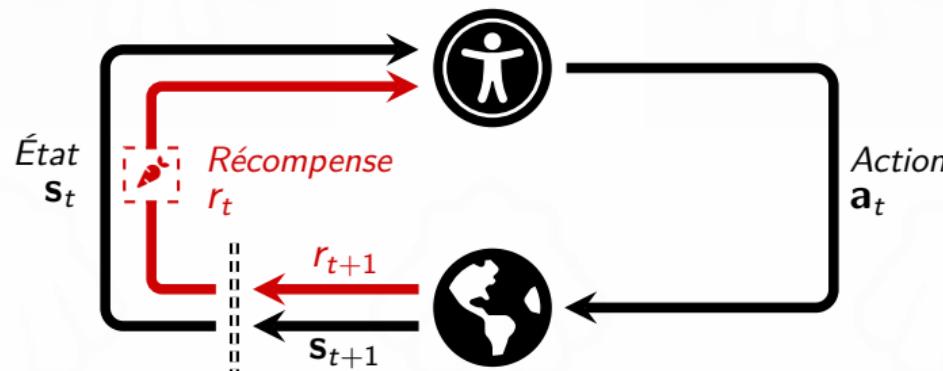
$$\tau \sim \pi(\tau) \implies s_1 \sim p(\cdot) \quad a_t = \pi(s_t) \quad s_{t+1} \sim p(\cdot | s_t, a_t)$$

$$\pi(\tau) = \pi(s_1, a_1, \dots, s_T, a_T) = p(s_1) \prod_{t=1}^T p(s_{t+1} | s_t, \pi(s_t))$$

Formalisme et outil de l'apprentissage par renforcement

Interaction agent-environnement

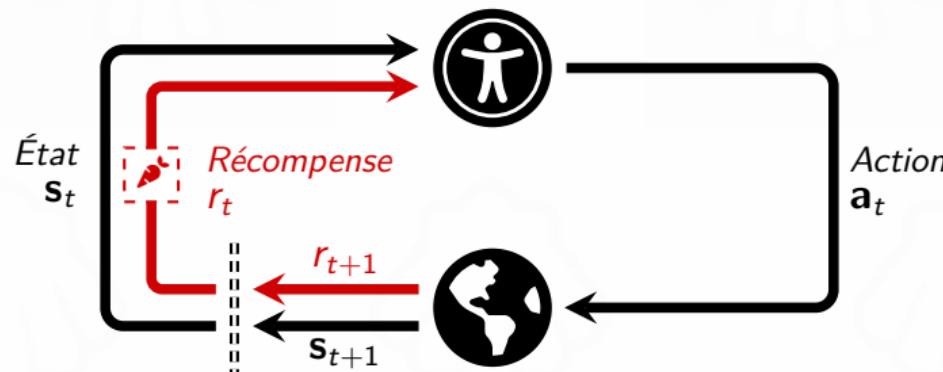
Le cycle d'interaction



Pour tout temps $t = 0, 1, 2, \dots, T$

- 1 L'agent observe l'environnement : $s_t \in \mathcal{S}$
- 2 Bas  sur cette observation, l'agent choisit et ex cute une action : $a_t \in \mathcal{A}(s_t)$
- 3 L'environnement renvoie un signal de r ompense (une r troaction) : $r_{t+1} \in \mathcal{R}$
- 4 et renvoie le nouvel  tat r sultant de cette action : $s_{t+1} \in \mathcal{S}$

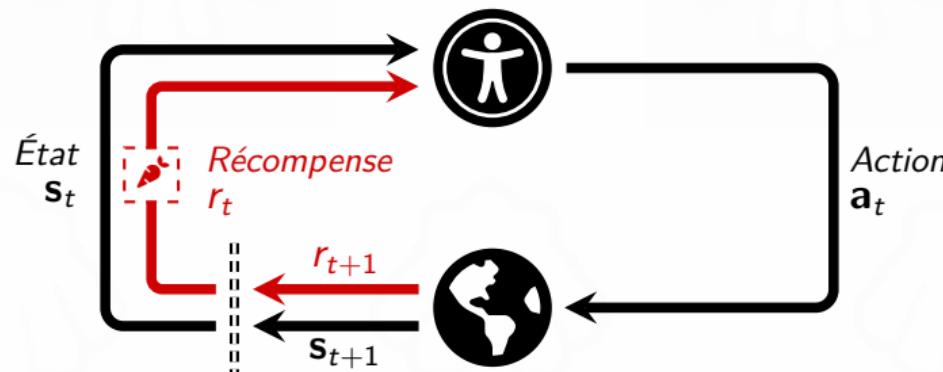
Le cycle d'interaction



Pour tout temps $t = 0, 1, 2, \dots, T$

- 1 L'agent **observe l'environnement** : $s_t \in \mathcal{S}$
- 2 Bas  sur cette observation, l'agent **choisit et ex cute une action** : $a_t \in \mathcal{A}(s_t)$
- 3 L'environnement renvoie un **signal de r  ompense** (une r  troaction) : $r_{t+1} \in \mathcal{R}$
- 4 et renvoie le nouvel ** tat r  sultant** de cette action : $s_{t+1} \in \mathcal{S}$

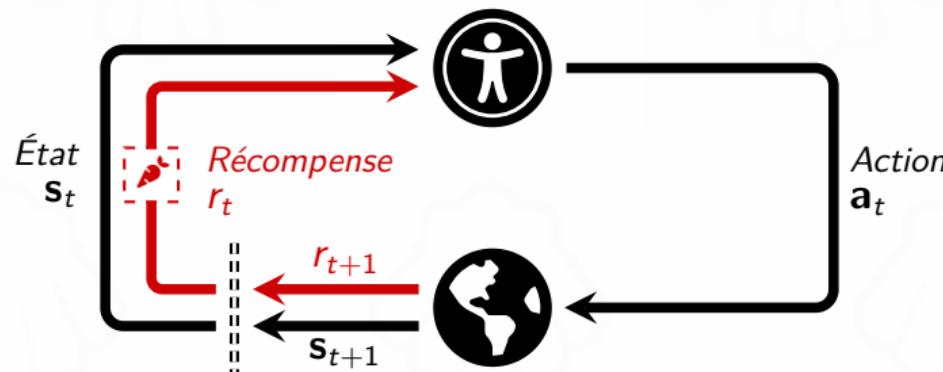
Le cycle d'interaction



Pour tout temps $t = 0, 1, 2, \dots, T$

- 1 L'agent **observe l'environnement** : $s_t \in \mathcal{S}$
- 2 Bas  sur cette observation, l'agent **choisit et ex cute une action** : $a_t \in \mathcal{A}(s_t)$
- 3 L'environnement renvoie un **signal de r  ompense** (une r  troaction) : $r_{t+1} \in \mathcal{R}$
- 4 et renvoie le nouvel  tat r  sultant de cette action : $s_{t+1} \in \mathcal{S}$

Le cycle d'interaction



Pour tout temps $t = 0, 1, 2, \dots, T$

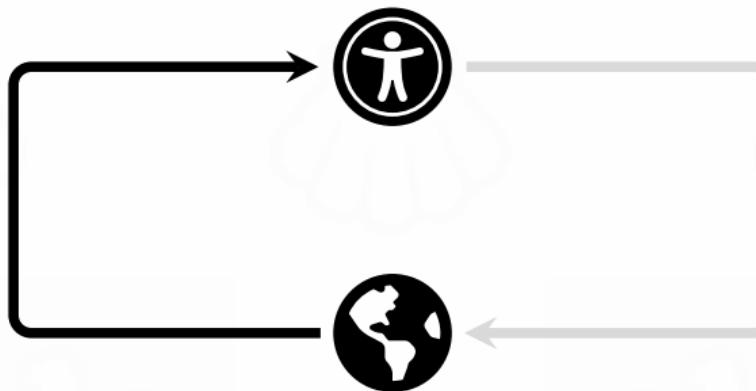
- 1 L'agent **observe l'environnement** : $s_t \in \mathcal{S}$
- 2 Bas  sur cette observation, l'agent **choisit et ex cute une action** : $a_t \in \mathcal{A}(s_t)$
- 3 L'environnement renvoie un **signal de r compense** (une r troaction) : $r_{t+1} \in \mathcal{R}$
- 4 et renvoie le nouvel ** tat r sultant** de cette action : $s_{t+1} \in \mathcal{S}$

Le cycle d'interaction

Temps $t = 1$

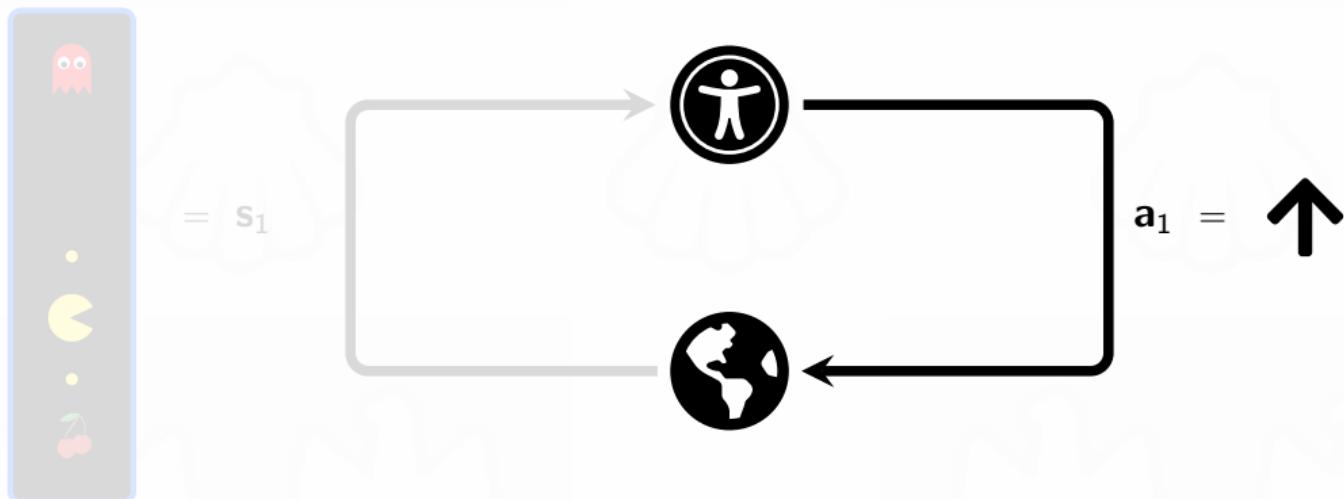


= s_1



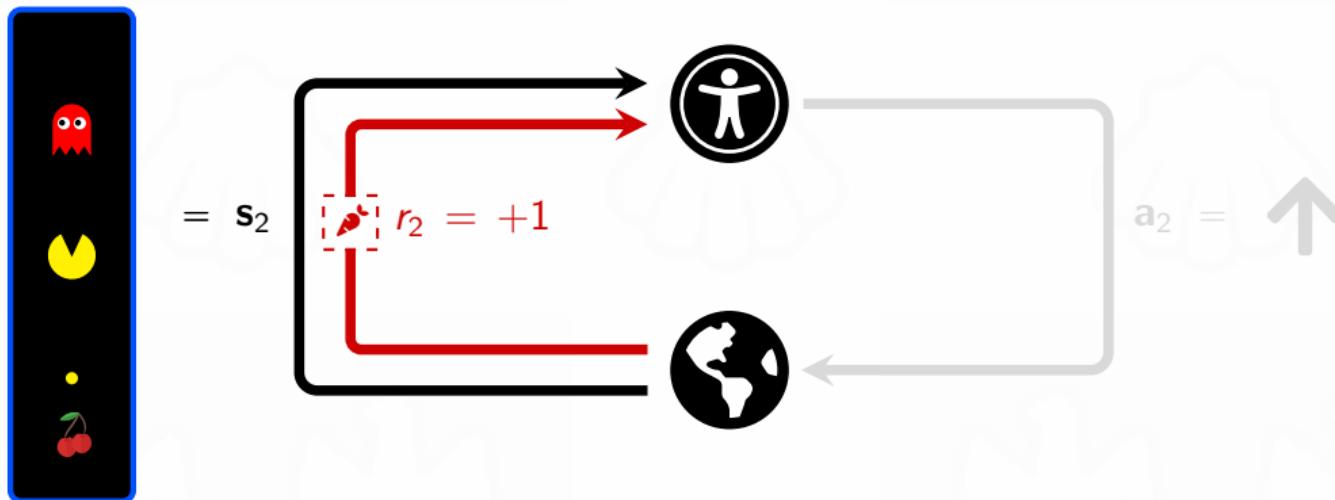
Le cycle d'interaction

Temps $t = 1$



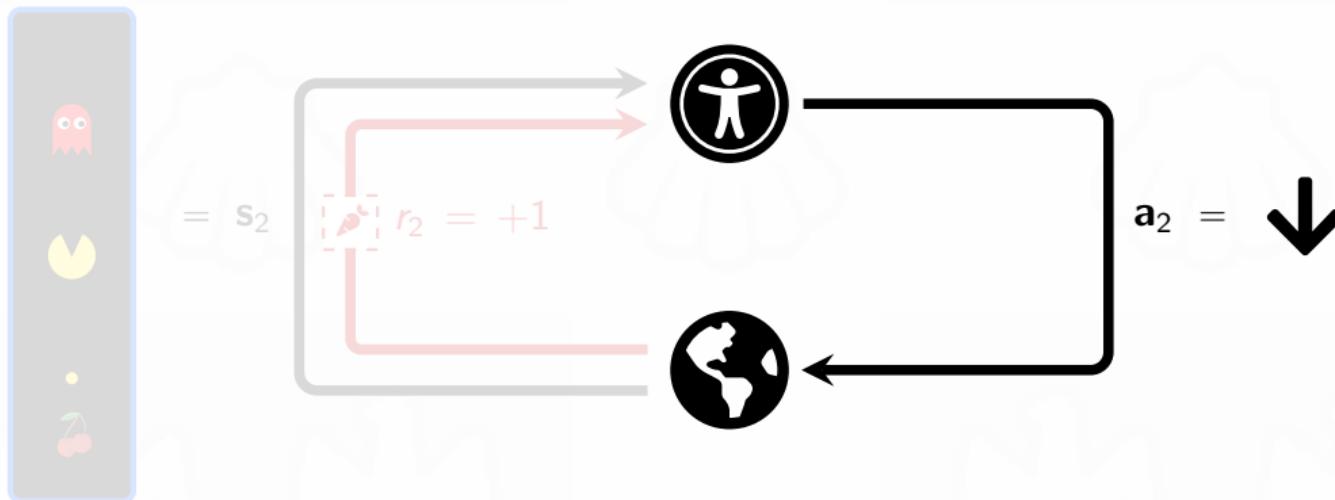
Le cycle d'interaction

Temps $t = 2$



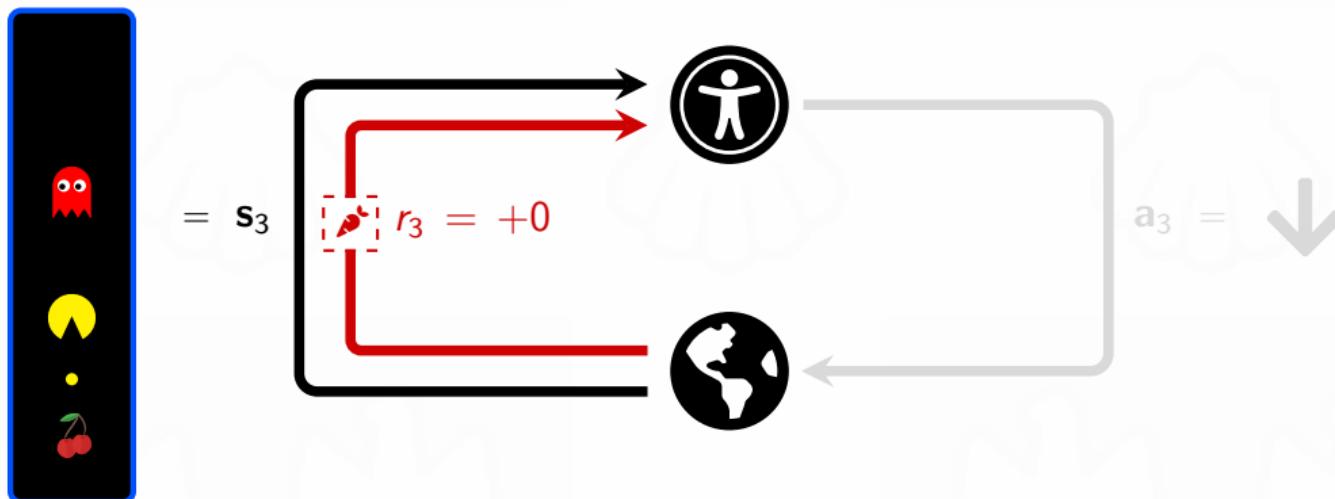
Le cycle d'interaction

Temps $t = 2$



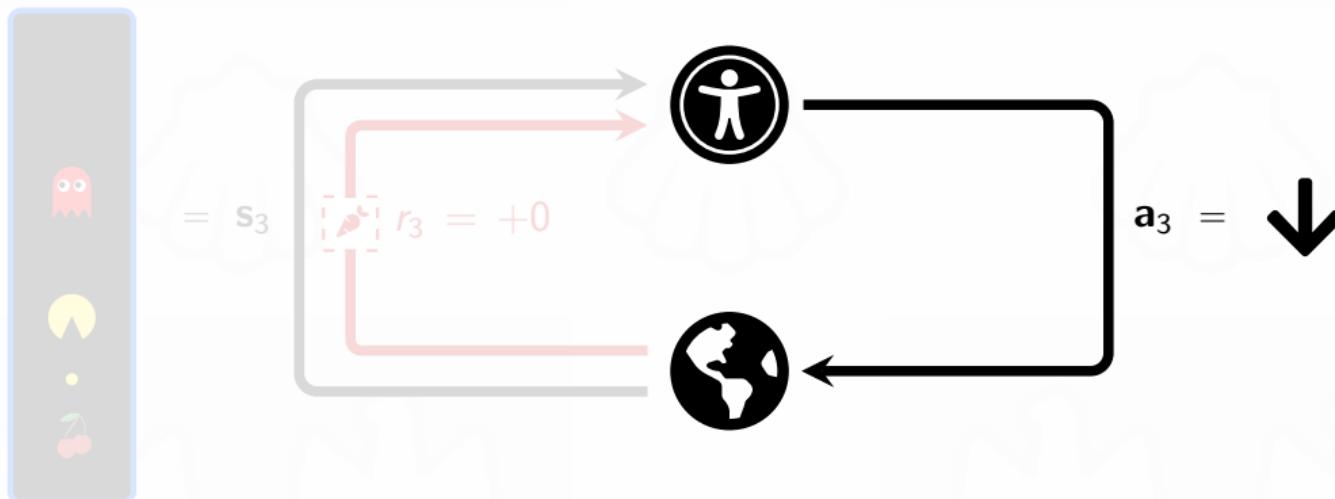
Le cycle d'interaction

Temps $t = 3$



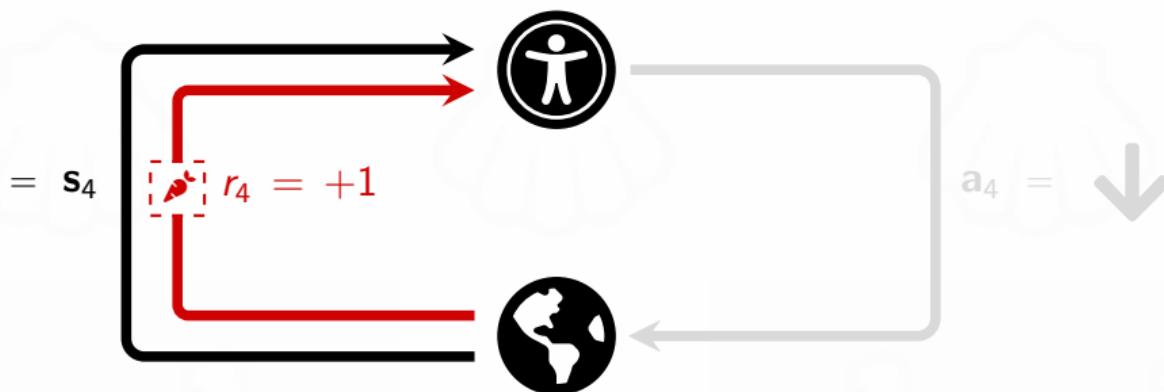
Le cycle d'interaction

Temps $t = 3$



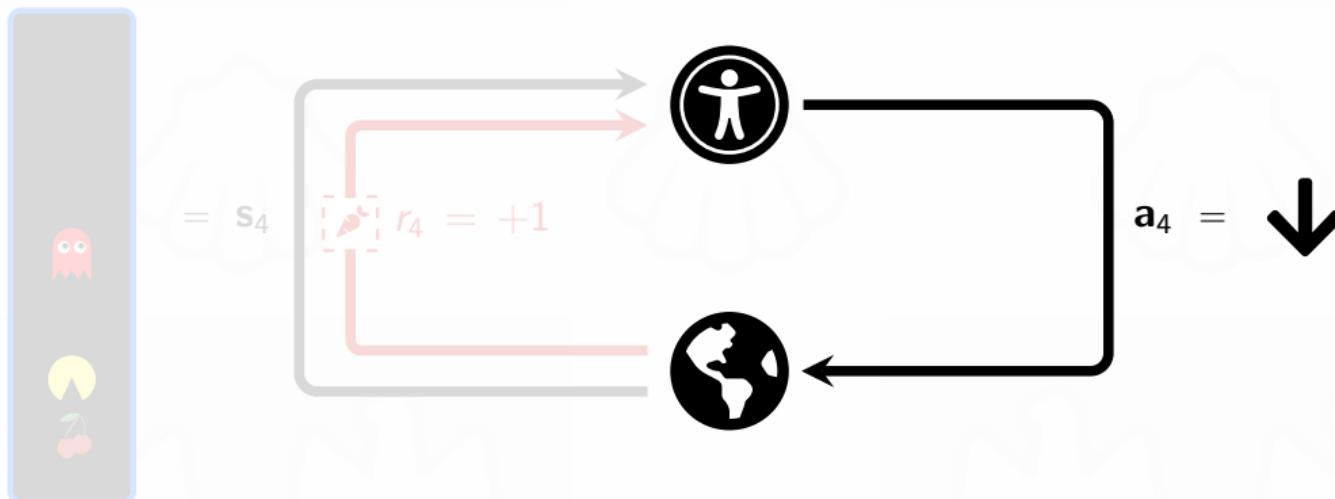
Le cycle d'interaction

Temps $t = 4$



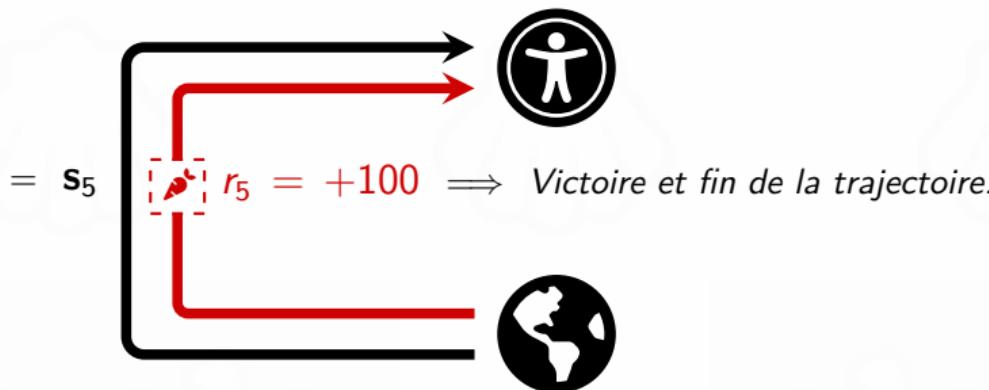
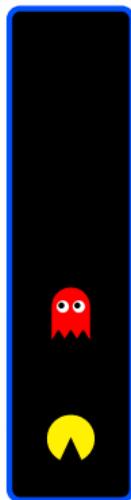
Le cycle d'interaction

Temps $t = 4$



Le cycle d'interaction

Temps $t = 5$



L'univers du point de vue de l'agent

1 Vue d'ensemble

2 Formalisme et outil de l'apprentissage par renforcement

3 L'univers du point de vue de l'agent

- Ce qu'on cherche à faire
- Quelle information l'agent peut-il utiliser pour apprendre ?
- Les fonctions de valeur Q^π et V^π
- Équation d'optimalité de Bellman

4 Pour aller plus loin

L'univers du point de vue de l'agent

Ce qu'on cherche à faire

Le probl me d'apprentissage par renforcement (Rappel)

★ – (Informellement) L'objectif de l'apprentissage par renforcement :

Que l'**agent** trouve par lui-m me **la meilleure strat gie** (une **politique optimale**) pour atteindre un **but** en se basant uniquement sur les **r compenses** retourn es par l'environnement.

★ – (Formellement) L'objectif de l'apprentissage par renforcement :

 **Maximiser** le total des r compenses r accumul  au cours d'une trajectoire τ

$$\begin{aligned}\pi^* &= \arg \max_{\pi} \mathbb{E}_{\tau \sim \pi} [G(\tau)] \\ &= \arg \max_{\pi} \sum_{t=1}^T \mathbb{E}_{(s_t, a_t, s_{t+1}) \sim \pi} [r(s_t, a_t, s_{t+1})]\end{aligned}$$

avec l'assumption que $\gamma = 1$

OK . . . par o  on commence ?

L'univers du point de vue de l'agent

Quelle information l'agent peut-il utiliser pour apprendre ?

L'agent ne sait pas :

- comment l'environnement fonctionne ;
- comment la fonction de r compense est d finie ;
- les conditions pour qu'une trajectoire termine ;

L'agent sait :

- qu'une bonne action est r compens e et qu'une mauvaise action est punis ;
- que plus il re oit de r compenses, plus la vie est belle ;

Question : Et si l'agent construisait un *mapping* action-r compense $S \times R$?

Quel est le probl me avec cette approche ?

L'agent ne sait pas :

- comment l'environnement fonctionne ;
- comment la fonction de r compense est d finie ;
- les conditions pour qu'une trajectoire termine ;

L'agent sait :

- qu'une bonne action est r compens e et qu'une mauvaise action est punis ;
- que plus il re oit de r compenses, plus la vie est belle ;

Question : Et si l'agent construisait un *mapping* action-r compense $\mathcal{S} \times \mathcal{R}$?

Quel est le probl me avec cette approche ?

L'agent ne sait pas :

- comment l'environnement fonctionne ;
- comment la fonction de r compense est d finie ;
- les conditions pour qu'une trajectoire termine ;

L'agent sait :

- qu'une bonne action est r compens e et qu'une mauvaise action est punis ;
- que plus il re oit de r compenses, plus la vie est belle ;

Question : Et si l'agent construisait un *mapping* action-r compense $\mathcal{S} \times \mathcal{R}$?

Quel est le probl me avec cette approche ?

L'agent ne sait pas :

- comment l'environnement fonctionne ;
- comment la fonction de r compense est d finie ;
- les conditions pour qu'une trajectoire termine ;

L'agent sait :

- qu'une bonne action est r compens e et qu'une mauvaise action est punis ;
- que plus il re oit de r compenses, plus la vie est belle ;

Question : Et si l'agent construisait un *mapping* action-r compense $\mathcal{S} \times \mathcal{R}$?

Quel est le probl me avec cette approche ?

 valu  les cons quences d'une action : voir le long terme

Probl me : L'agent n'a aucun moyen direct de savoir si la r compense qu'il re oit maintenant est due   l'action qu'il vient juste d'ex cuter !

Pris seul, une r compense ne dit pas grand-chose

Elle peut  tre due   une action qui est arriv e longtemps auparavant

Ex. : mettre son parachute avant de sauter en bas d'un avion

Donc la seule fa on pour l'agent d' valuer les cons quences des actions qu'il prend est **en gardant une vue d'ensemble** de la trajectoire \Rightarrow la seule quantit  qui lui permet de d duire qu'il a fait quelque chose de bien durant une trajectoire, c'est **le retour**

$$G_1(\tau) = r_2 + r_3 + r_4 + \dots + r_{\tau-2} + r_{\tau-1} + r_{\tau}$$

avec l'assumption que $\gamma = 1$

 valuer le ou les cause(s) d'une r action ? (Est-ce possible ?)

On sait maintenant que l'agent **doit  valuer les cons quences** de ses actes **en se basant sur le retour** G_1 des trajectoires τ .

Pr sentement la seule information que l'agent peut d duire du retour G_1 est qu'une trajectoire τ est meilleure qu'une autre τ' .

Si $\tau = s_1, a_1, r_2, s_2, a_2, r_3, \dots, s_T, a_T, r_{T+1} \mapsto \text{ }$

et $\tau' = s'_1, a'_1, r'_2, s'_2, a'_2, r'_3, \dots, s'_T, a'_T, r'_{T+1} \mapsto \text{ }$

alors $\tau > \tau'$

 valu  le ou les cause(s) d'une r action ? (Est-ce possible ?)

Question : Comment **identifier le ou les cause(s)** qui font que cette trajectoire est meilleure que les autres ? Comment l'agent peut-il faire pour valuer la causalit  d'un vnement ?

« Est-ce que cette personne est toujours vivante aujourd'hui
parce qu'elle a arr t  de fumer le mois dernier

ou

  cause qu'elle a mis un parachute avant de sauter d'un avion hier ? »

L'agent a besoin d'un **moyen pour comparer les segments** d'une trajectoire entre eux tout **en respectant la contrainte de vue d'ensemble**.

 valu  le ou les cause(s) d'une r action ? (Est-ce possible ?)

Question : Comment identifier le ou les cause(s) qui font que cette trajectoire est meilleure que les autres ? Comment l'agent peut-il faire pour valuer la causalit  d'un vnement ?

« Est-ce que cette personne est toujours vivante aujourd'hui
parce qu'elle a arr t  de fumer le mois dernier

ou

  cause qu'elle a mis un parachute avant de sauter d'un avion hier ? »

L'agent a besoin d'un moyen pour comparer les segments d'une trajectoire entre eux tout en respectant la contrainte de vue d'ensemble.

 valu  le ou les cause(s) d'une r action ? (Est-ce possible ?)

Question : Comment **identifier le ou les cause(s)** qui font que cette trajectoire est meilleure que les autres ? Comment l'agent peut-il faire pour valuer la causalit  d'un vnement ?

« Est-ce que cette personne est toujours vivante aujourd'hui
parce qu'elle a arr t  de fumer le mois dernier

ou

  cause qu'elle a mis un parachute avant de sauter d'un avion hier ? »

L'agent a besoin d'un **moyen pour comparer les segments** d'une trajectoire entre eux tout **en respectant la contrainte de vue d'ensemble**.

Choisir le bon point de vue

Question : Sur quelle base l'agent pourrait-il comparer des segments de trajectoire ?

$$\tau = s_1, a_1, r_2, s_2, a_2, r_3, s_3, a_3, r_4, s_4, a_4, r_5, \dots, s_T, a_T, r_{T+1}$$

Exemple : Par rapport   l'utilit  de ...

- (s) = d' tre assis dans un cours important et [prendre des notes ou regarder son cellulaire ou ...]
- (s, a) = de prendre des notes et d' tre assis dans un cours important
- $\pi(s)$ = de suivre la bonne habitude de toujours noter ce qui est important en cours

(s) ----- (s, a)

$\pi(s)$

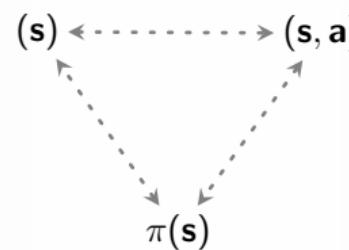
Choisir le bon point de vue

Question : Sur quelle base l'agent pourrait-il comparer des segments de trajectoire ?

$$\tau = s_1, a_1, r_2, s_2, a_2, r_3, s_3, a_3, r_4, s_4, a_4, r_5, \dots, s_T, a_T, r_{T+1}$$

Exemple : Par rapport   l'utilit  de ...

- (s) = d' tre assis dans un cours important et [prendre des notes ou regarder son cellulaire ou ...]
- (s, a) = de prendre des notes et d' tre assis dans un cours important
- $\pi(s)$ = de suivre la bonne habitude de toujours noter ce qui est important en cours



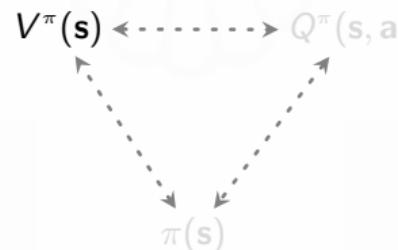
Choisir le bon point de vue

Perspective  propos des trajectoires (plus formellement)

$V^\pi(s)$ est la valeur d'tre dans une tat s
et de continu d'agir en suivant la politique π

$Q^\pi(s, a)$ est la valeur d'excuter une action a  partir d'un tat s
et de continu d'agir en suivant la politique π

$\pi(s)$ est l'utilit d'agir selon une politique particulire



Ce sont trois **ides fondamentales** qui reviennent constamment
dans la formulation des problmes en *RL*.

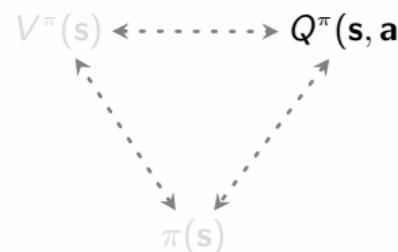
Choisir le bon point de vue

Perspective   propos des trajectoires (plus formellement)

$V^\pi(s)$ est la valeur d' tre dans une  tat s
et de continu  d'agir en suivant la politique π

$Q^\pi(s, a)$ est la valeur d'ex cuter une action a   partir d'un  tat s
et de continu  d'agir en suivant la politique π

$\pi(s)$ est l'utilit  d'agir selon une politique particuli re



Ce sont trois **id es fondamentales** qui reviennent constamment
dans la formulation des probl mes en *RL*.

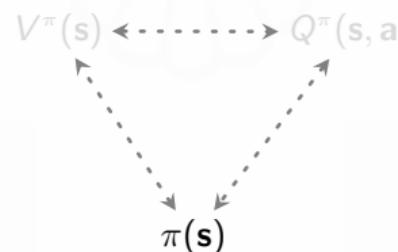
Choisir le bon point de vue

Perspective   propos des trajectoires (plus formellement)

$V^\pi(s)$ est la valeur d' tre dans une  tat s
et de continu  d'agir en suivant la politique π

$Q^\pi(s, a)$ est la valeur d'ex閐uter une action a   partir d'un  tat s
et de continu  d'agir en suivant la politique π

$\pi(s)$ est l'utilit  d'agir selon une politique particuli re



Ce sont trois **id es fondamentales** qui reviennent constamment dans la formulation des probl mes en *RL*.

Choisir le bon point de vue

Perspective  propos des trajectoires (plus formellement)

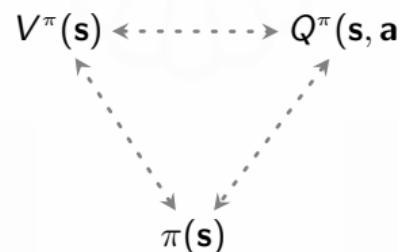
$V^\pi(s)$ est la valeur d'tre dans une tat s

et de continu d'agir en suivant la politique π

$Q^\pi(s, a)$ est la valeur d'excuter une action a  partir d'un tat s

et de continu d'agir en suivant la politique π

$\pi(s)$ est l'utilit d'agir selon une politique particulire



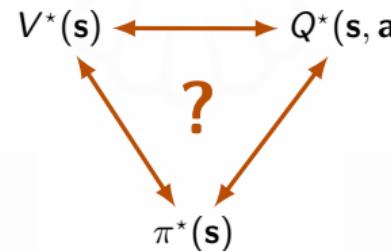
Ce sont trois **ides fondamentales** qui reviennent constamment dans la formulation des problmes en *RL*.

Apprendre quoi ? π , V^π ou Q^π

Ce trio est aussi au coeur d'un **dilemme critique à la résolution** de problème de RL :

Qu'est-ce que l'agent doit apprendre pour atteindre l'objectif ?

On doit optimiser quoi concrètement ?



On va tenter d'éclaircir la question pendant le reste du cours

L'univers du point de vue de l'agent

Les fonctions de valeur Q^π et V^π

La valeur d'une action   partir d'un   tat [Q-function, action-value function]

Rappel : $r_{t+1} = r(s_t, a_t, s_{t+1})$

Remarque : $G_t(\tau) = \sum_{k=t}^{\tau} \gamma^{k-t} r(s_k, a_k, s_{k+1}) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = r_{t+1} + \gamma G_{t+1}$

La fonction de valeur $Q^\pi(s_t, a_t)$

R ponds   la question :   quel point prendre cette action   partir de cet   tat est bon ?

Comment : En calculant   partir de l'  tat s_t la valeur attendue du retour G_t si on prend l'action a tout en suivant la politique π jusqu'  la fin de la trajectoire.

$$\begin{aligned} Q^\pi(s_t, a_t) &= \mathbb{E}_{\tau \sim \pi} [G_t | S_t = s, A_t = a] \\ &= \mathbb{E}_{\tau \sim \pi} [r_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\ &= \mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} \left[r_{t+1} + \gamma \mathbb{E}_{a_{t+1} \sim \pi(\cdot | s_{t+1})} [Q^\pi(s_{t+1}, a_{t+1})] \right] \\ &= \sum_{s' \in \mathcal{S}_{t+1}} p(s' | s_t, a_t) \left[r_{t+1} + \gamma \sum_{a' \in \mathcal{A}_{t+1}} \pi(a' | s') Q^\pi(s', a') \right] \end{aligned}$$

La valeur d'une action   partir d'un   tat [Q-function, action-value function]

Rappel : $r_{t+1} = r(s_t, a_t, s_{t+1})$

Remarque : $G_t(\tau) = \sum_{k=t}^{\tau} \gamma^{k-t} r(s_k, a_k, s_{k+1}) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = r_{t+1} + \gamma G_{t+1}$

La fonction de valeur $Q^\pi(s_t, a_t)$

R ponds   la question :   quel point prendre cette action   partir de cet   tat est bon ?

Comment : En calculant   partir de l'  tat s_t la valeur attendue du retour G_t si on prend l'action a tout en suivant la politique π jusqu'  la fin de la trajectoire.

$$\begin{aligned} Q^\pi(s_t, a_t) &= \mathbb{E}_{\tau \sim \pi} [G_t | S_t = s, A_t = a] \\ &= \mathbb{E}_{\tau \sim \pi} [r_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\ &= \mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} \left[r_{t+1} + \gamma \mathbb{E}_{a_{t+1} \sim \pi(\cdot | s_{t+1})} [Q^\pi(s_{t+1}, a_{t+1})] \right] \\ &= \sum_{s' \in \mathcal{S}_{t+1}} p(s' | s_t, a_t) \left[r_{t+1} + \gamma \sum_{a' \in \mathcal{A}_{t+1}} \pi(a' | s') Q^\pi(s', a') \right] \end{aligned}$$

La valeur d'une action   partir d'un   tat [Q-function, action-value function]

Rappel : $r_{t+1} = r(s_t, a_t, s_{t+1})$

Remarque : $G_t(\tau) = \sum_{k=t}^{\tau} \gamma^{k-t} r(s_k, a_k, s_{k+1}) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = r_{t+1} + \gamma G_{t+1}$

La fonction de valeur $Q^\pi(s_t, a_t)$

R ponds   la question :   quel point prendre cette action   partir de cet   tat est bon ?

Comment : En calculant   partir de l'  tat s_t la valeur attendue du retour G_t si on prend l'action a tout en suivant la politique π jusqu'  la fin de la trajectoire.

$$\begin{aligned} Q^\pi(s_t, a_t) &= \mathbb{E}_{\tau \sim \pi} [G_t | S_t = s, A_t = a] \\ &= \mathbb{E}_{\tau \sim \pi} [r_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\ &= \mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} \left[r_{t+1} + \gamma \mathbb{E}_{a_{t+1} \sim \pi(\cdot | s_{t+1})} [Q^\pi(s_{t+1}, a_{t+1})] \right] \\ &= \sum_{s' \in \mathcal{S}_{t+1}} p(s' | s_t, a_t) \left[r_{t+1} + \gamma \sum_{a' \in \mathcal{A}_{t+1}} \pi(a' | s') Q^\pi(s', a') \right] \end{aligned}$$

La valeur d'une action   partir d'un   tat [Q-function, action-value function]

Rappel : $r_{t+1} = r(s_t, a_t, s_{t+1})$

Remarque : $G_t(\tau) = \sum_{k=t}^T \gamma^{k-t} r(s_k, a_k, s_{k+1}) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = r_{t+1} + \gamma G_{t+1}$

La fonction de valeur $Q^\pi(s_t, a_t)$

R ponds   la question :   quel point prendre cette action   partir de cet   tat est bon ?

Comment : En calculant   partir de l'  tat s_t la valeur attendue du retour G_t si on prend l'action a tout en suivant la politique π jusqu'  la fin de la trajectoire.

$$\begin{aligned} Q^\pi(s_t, a_t) &= \mathbb{E}_{\tau \sim \pi} [G_t | S_t = s, A_t = a] \\ &= \mathbb{E}_{\tau \sim \pi} [r_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\ &= \mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} \left[r_{t+1} + \gamma \mathbb{E}_{a_{t+1} \sim \pi(\cdot | s_{t+1})} [Q^\pi(s_{t+1}, a_{t+1})] \right] \\ &= \sum_{s' \in \mathcal{S}_{t+1}} p(s' | s_t, a_t) \left[r_{t+1} + \gamma \sum_{a' \in \mathcal{A}_{t+1}} \pi(a' | s') Q^\pi(s', a') \right] \end{aligned}$$

Diagramme de la valeur d'une action   partir d'un   t t

$$Q^\pi(s_t, a_t) = \sum_{s' \in \mathcal{S}_{t+1}} p(s'|s_t, a_t) \left[r_{t+1} + \gamma \sum_{a' \in \mathcal{A}_{t+1}} \pi(a'|s') Q^\pi(s', a') \right]$$

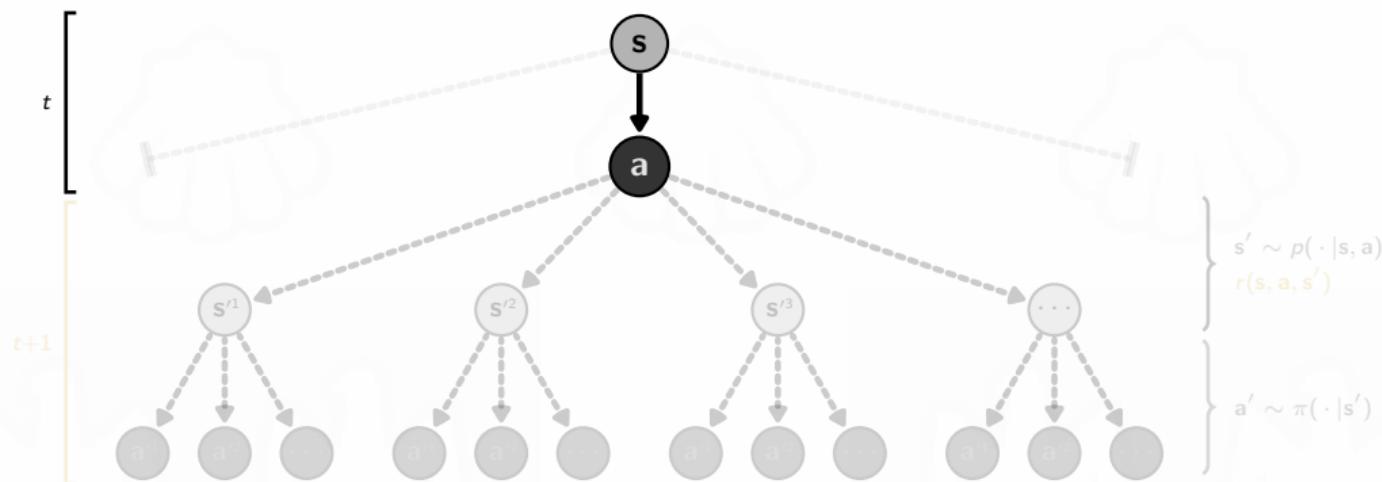


Diagramme de la valeur d'une action   partir d'un   tat

$$Q^\pi(s_t, a_t) = \sum_{s' \in \mathcal{S}_{t+1}} p(s'|s_t, a_t) \left[r_{t+1} + \gamma \sum_{a' \in \mathcal{A}_{t+1}} \pi(a'|s') Q^\pi(s', a') \right]$$

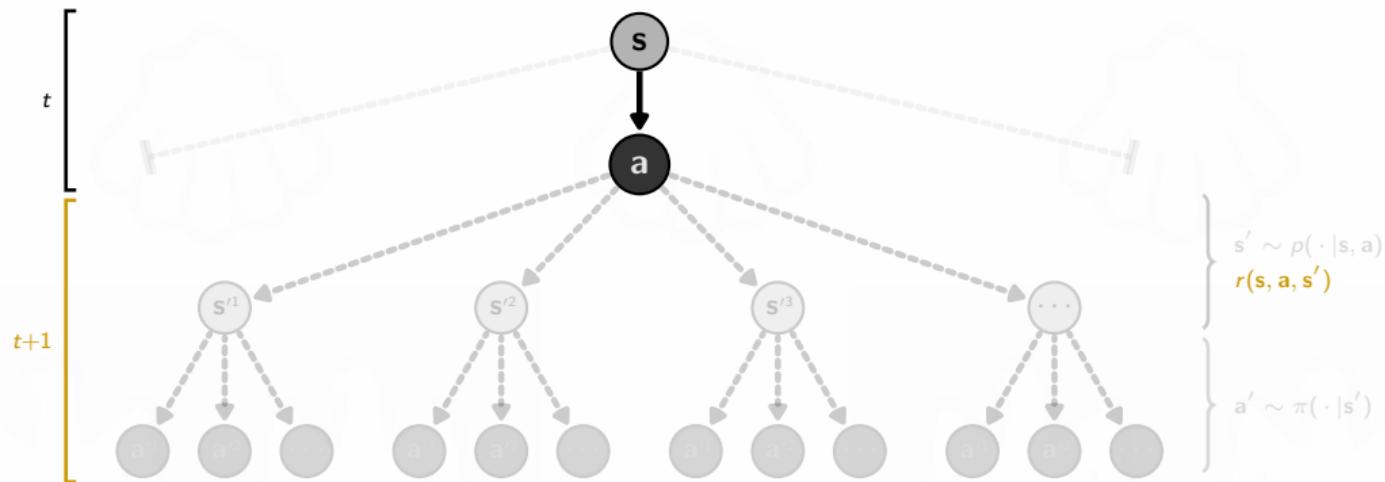


Diagramme de la valeur d'une action   partir d'un   t t

$$Q^\pi(s_t, a_t) = \sum_{s' \in S_{t+1}} p(s'|s_t, a_t) \left[r_{t+1} + \gamma \sum_{a' \in A_{t+1}} \pi(a'|s') Q^\pi(s', a') \right]$$

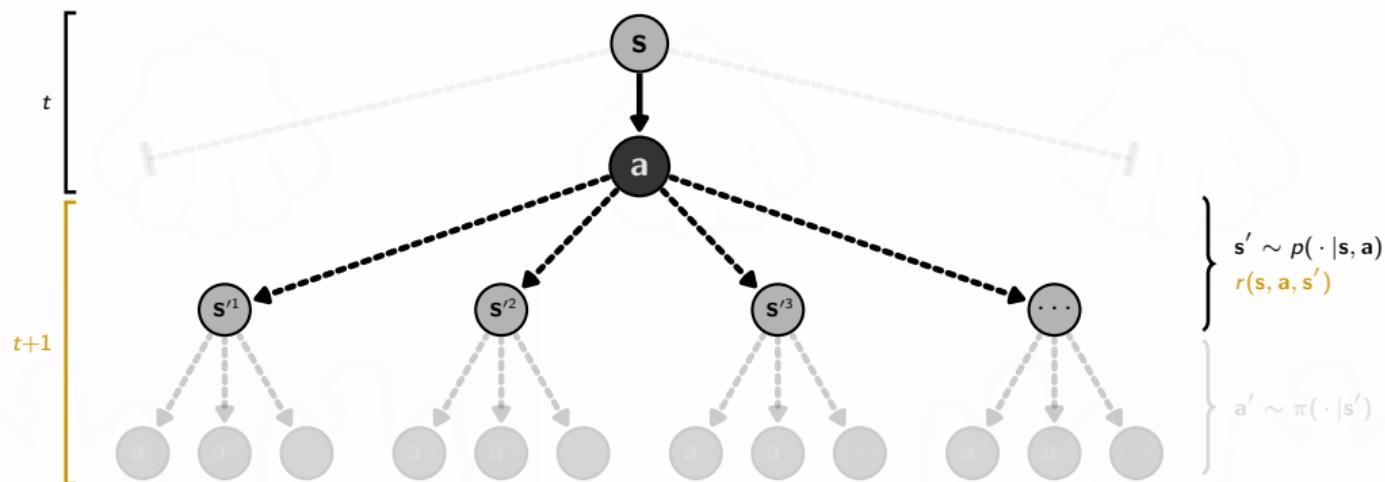
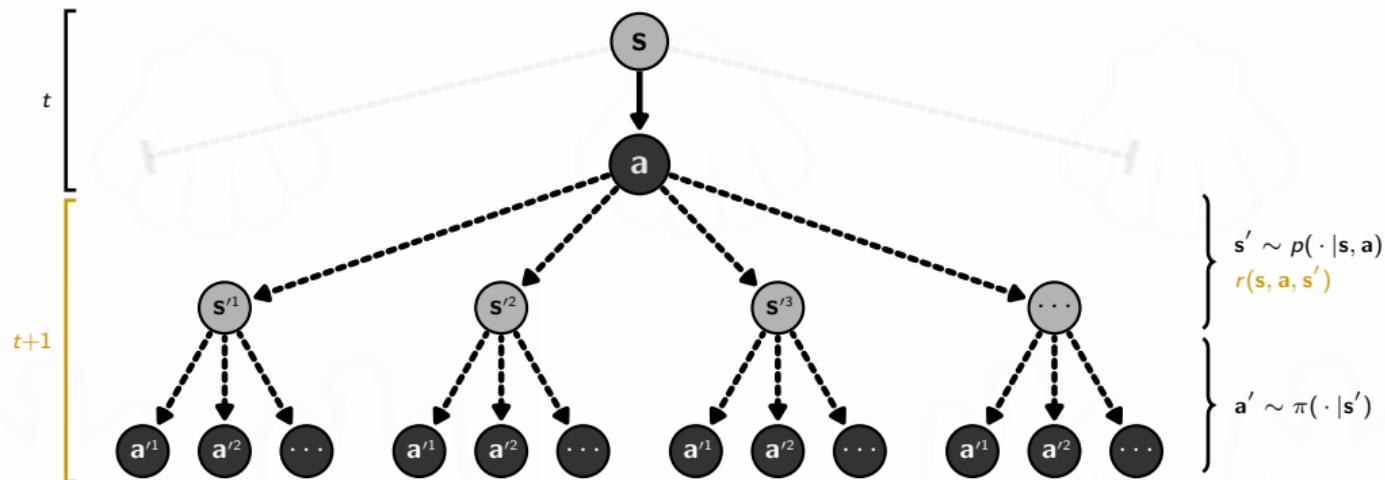


Diagramme de la valeur d'une action   partir d'un   tat

$$Q^\pi(s_t, a_t) = \sum_{s' \in S_{t+1}} p(s'|s_t, a_t) \left[r_{t+1} + \gamma \sum_{a' \in A_{t+1}} \pi(a'|s') Q^\pi(s', a') \right]$$



La valeur d'un  tat [Value function, state-value function]

Rappel : $r_{t+1} = r(s_t, a_t, s_{t+1})$

Remarque : $G_t(\tau) = \sum_{k=t}^{\tau} \gamma^{k-t} r(s_k, a_k, s_{k+1}) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = r_{t+1} + \gamma G_{t+1}$

La fonction de valeur $V^\pi(s_t)$

R pons   la question :  A quel point cet  tat est bon ?

Comment : En calculant   partir de l' tat s_t la valeur attendue du retour G_t tout en suivant la politique π jusqu'  la fin de la trajectoire.

$$\begin{aligned} V^\pi(s_t) &= \mathbb{E}_{\tau \sim \pi(\tau)} [G_t | S_t = s] \\ &= \mathbb{E}_{\tau \sim \pi(\tau)} [r_{t+1} + \gamma G_{t+1} | S_t = s] \\ &= \mathbb{E}_{a_t \sim \pi(\cdot | s_t)} [Q^\pi(s_t, a_t)] \\ &= \mathbb{E}_{a_t \sim \pi(\cdot | s_t)} [\mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} [r_{t+1} + \gamma V^\pi(s_{t+1})]] \\ &= \sum_{a \in \mathcal{A}_t} \pi(a | s_t) \sum_{s' \in \mathcal{S}_{t+1}} p(s' | s_t, a) [r_{t+1} + \gamma V^\pi(s')] \end{aligned}$$

La valeur d'un  tat [Value function, state-value function]

Rappel : $r_{t+1} = r(s_t, a_t, s_{t+1})$

Remarque : $G_t(\tau) = \sum_{k=t}^{\tau} \gamma^{k-t} r(s_k, a_k, s_{k+1}) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = \color{red}{r_{t+1} + \gamma G_{t+1}}$

La fonction de valeur $V^\pi(s_t)$

R pons   la question :  A quel point cet  tat est bon ?

Comment : En calculant  partir de l' tat s_t la valeur attendue du retour G_t tout en suivant la politique π jusqu'  la fin de la trajectoire.

$$\begin{aligned} V^\pi(s_t) &= \mathbb{E}_{\tau \sim \pi(\tau)} [G_t | S_t = s] \\ &= \mathbb{E}_{\tau \sim \pi(\tau)} [\color{red}{r_{t+1} + \gamma G_{t+1}} | S_t = s] \\ &= \mathbb{E}_{a_t \sim \pi(\cdot | s_t)} [Q^\pi(s_t, a_t)] \\ &= \mathbb{E}_{a_t \sim \pi(\cdot | s_t)} \left[\mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} [r_{t+1} + \gamma V^\pi(s_{t+1})] \right] \\ &= \sum_{a \in \mathcal{A}_t} \pi(a | s_t) \sum_{s' \in \mathcal{S}_{t+1}} p(s' | s_t, a) [r_{t+1} + \gamma V^\pi(s')] \end{aligned}$$

La valeur d'un  tat [Value function, state-value function]

Rappel : $r_{t+1} = r(s_t, a_t, s_{t+1})$

Remarque : $G_t(\tau) = \sum_{k=t}^{\tau} \gamma^{k-t} r(s_k, a_k, s_{k+1}) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = \color{red}{r_{t+1} + \gamma G_{t+1}}$

La fonction de valeur $V^\pi(s_t)$

R pons   la question :  A quel point cet  tat est bon ?

Comment : En calculant  partir de l' tat s_t la valeur attendue du retour G_t tout en suivant la politique π jusqu'  la fin de la trajectoire.

$$\begin{aligned} V^\pi(s_t) &= \mathbb{E}_{\tau \sim \pi(\tau)} [G_t | S_t = s] \\ &= \mathbb{E}_{\tau \sim \pi(\tau)} [\color{red}{r_{t+1} + \gamma G_{t+1}} | S_t = s] \\ &= \mathbb{E}_{a_t \sim \pi(\cdot | s_t)} [Q^\pi(s_t, a_t)] \\ &= \mathbb{E}_{a_t \sim \pi(\cdot | s_t)} \left[\mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} [r_{t+1} + \gamma V^\pi(s_{t+1})] \right] \\ &= \sum_{a \in \mathcal{A}_t} \pi(a | s_t) \sum_{s' \in \mathcal{S}_{t+1}} p(s' | s_t, a) [r_{t+1} + \gamma V^\pi(s')] \end{aligned}$$

La valeur d'un  tat [Value function, state-value function]

Rappel : $r_{t+1} = r(s_t, a_t, s_{t+1})$

Remarque : $G_t(\tau) = \sum_{k=t}^{\tau} \gamma^{k-t} r(s_k, a_k, s_{k+1}) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = r_{t+1} + \gamma G_{t+1}$

La fonction de valeur $V^\pi(s_t)$

R pons   la question :  A quel point cet  tat est bon ?

Comment : En calculant  partir de l' tat s_t la valeur attendue du retour G_t tout en suivant la politique π jusqu'  la fin de la trajectoire.

$$\begin{aligned} V^\pi(s_t) &= \mathbb{E}_{\tau \sim \pi(\tau)} [G_t | S_t = s] \\ &= \mathbb{E}_{\tau \sim \pi(\tau)} [r_{t+1} + \gamma G_{t+1} | S_t = s] \\ &= \mathbb{E}_{a_t \sim \pi(\cdot | s_t)} [Q^\pi(s_t, a_t)] \\ &= \mathbb{E}_{a_t \sim \pi(\cdot | s_t)} [\mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} [r_{t+1} + \gamma V^\pi(s_{t+1})]] \\ &= \sum_{a \in \mathcal{A}_t} \pi(a | s_t) \sum_{s' \in \mathcal{S}_{t+1}} p(s' | s_t, a) [r_{t+1} + \gamma V^\pi(s')] \end{aligned}$$

La valeur d'un  tat [Value function, state-value function]

Rappel : $r_{t+1} = r(s_t, a_t, s_{t+1})$

Remarque : $G_t(\tau) = \sum_{k=t}^{\tau} \gamma^{k-t} r(s_k, a_k, s_{k+1}) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = r_{t+1} + \gamma G_{t+1}$

La fonction de valeur $V^\pi(s_t)$

R pons   la question :  A quel point cet  tat est bon ?

Comment : En calculant  partir de l' tat s_t la valeur attendue du retour G_t tout en suivant la politique π jusqu'  la fin de la trajectoire.

$$\begin{aligned} V^\pi(s_t) &= \mathbb{E}_{\tau \sim \pi(\tau)} [G_t | S_t = s] \\ &= \mathbb{E}_{\tau \sim \pi(\tau)} [r_{t+1} + \gamma G_{t+1} | S_t = s] \\ &= \mathbb{E}_{a_t \sim \pi(\cdot | s_t)} [Q^\pi(s_t, a_t)] \\ &= \mathbb{E}_{a_t \sim \pi(\cdot | s_t)} [\mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} [r_{t+1} + \gamma V^\pi(s_{t+1})]] \\ &= \sum_{a \in \mathcal{A}_t} \pi(a | s_t) \sum_{s' \in \mathcal{S}_{t+1}} p(s' | s_t, a) [r_{t+1} + \gamma V^\pi(s')] \end{aligned}$$

Diagramme de la valeur d'un  t t

$$V^\pi(s_t) = \sum_{a \in \mathcal{A}_t} \pi(a|s_t) \sum_{s' \in \mathcal{S}_{t+1}} p(s'|s_t, a) [r_{t+1} + \gamma V^\pi(s')]$$

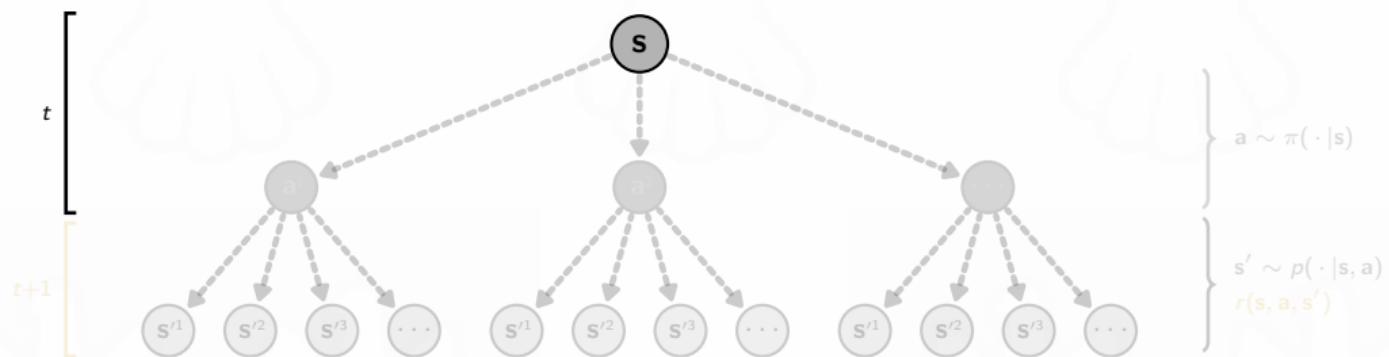


Diagramme de la valeur d'un  t t

$$V^\pi(s_t) = \sum_{a \in \mathcal{A}_t} \pi(a|s_t) \sum_{s' \in \mathcal{S}_{t+1}} p(s'|s_t, a) [r_{t+1} + \gamma V^\pi(s')]$$

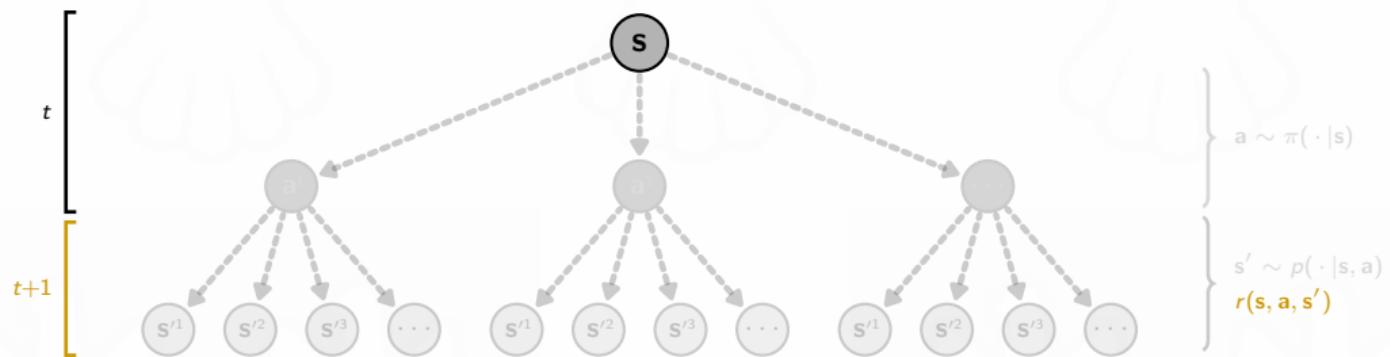


Diagramme de la valeur d'un  t t

$$V^\pi(s_t) = \sum_{a \in \mathcal{A}_t} \pi(a|s_t) \sum_{s' \in \mathcal{S}_{t+1}} p(s'|s_t, a) [r_{t+1} + \gamma V^\pi(s')]$$

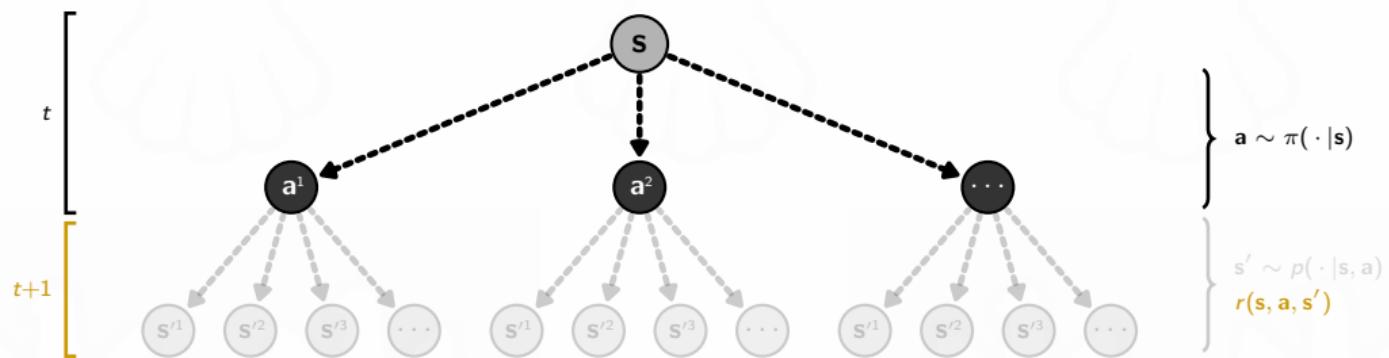
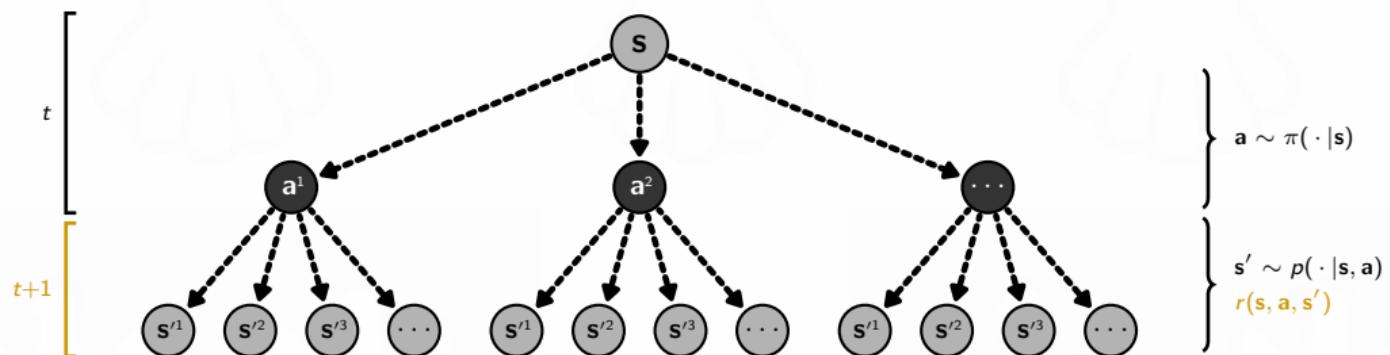


Diagramme de la valeur d'un  t t

$$V^\pi(s_t) = \sum_{a \in \mathcal{A}_t} \pi(a|s_t) \sum_{s' \in \mathcal{S}_{t+1}} p(s'|s_t, a) [r_{t+1} + \gamma V^\pi(s')]$$



L'univers du point de vue de l'agent

Équation d'optimalité de Bellman

La fonction de valeur optimale $Q^*(s_t, a_t)$

R pons   la question :   quel point prendre cette action   partir de cet  tat est bon si on agit optimalement   partir de ce point jusqu'  la fin de la trajectoire ?

$$\begin{aligned} Q^*(s_t, a_t) &= \max_{\pi} Q^{\pi}(s_t, a_t) \quad \text{pour tous } s \in \mathcal{S} \text{ et } a \in \mathcal{A} \\ &= \mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a_t)} \left[r_{t+1} + \gamma \max_{a' \in \mathcal{A}_{t+1}} Q^*(s_{t+1}, a') \right] \\ &= \sum_{s' \in \mathcal{S}_{t+1}} p(s' | s_t, a_t) \left[r_{t+1} + \gamma \max_{a' \in \mathcal{A}_{t+1}} Q^*(s', a') \right] \end{aligned}$$

Diagramme de la valeur optimale d'une action   partir d'un   tat

$$Q^*(\mathbf{s}_t, \mathbf{a}_t) = \sum_{\mathbf{s}' \in \mathcal{S}_{t+1}} p(\mathbf{s}' | \mathbf{s}_t, \mathbf{a}_t) \left[r_{t+1} + \gamma \max_{\mathbf{a}' \in \mathcal{A}_{t+1}} Q^*(\mathbf{s}', \mathbf{a}') \right]$$

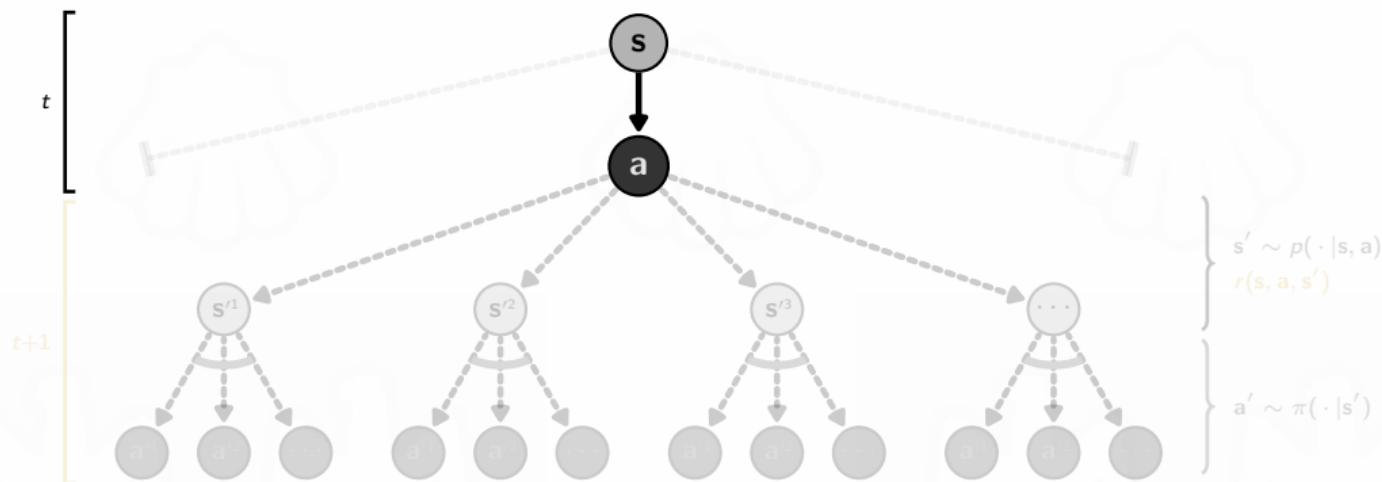


Diagramme de la valeur optimale d'une action   partir d'un   tat

$$Q^*(s_t, a_t) = \sum_{s' \in \mathcal{S}_{t+1}} p(s'|s_t, a_t) \left[r_{t+1} + \gamma \max_{a' \in \mathcal{A}_{t+1}} Q^*(s', a') \right]$$

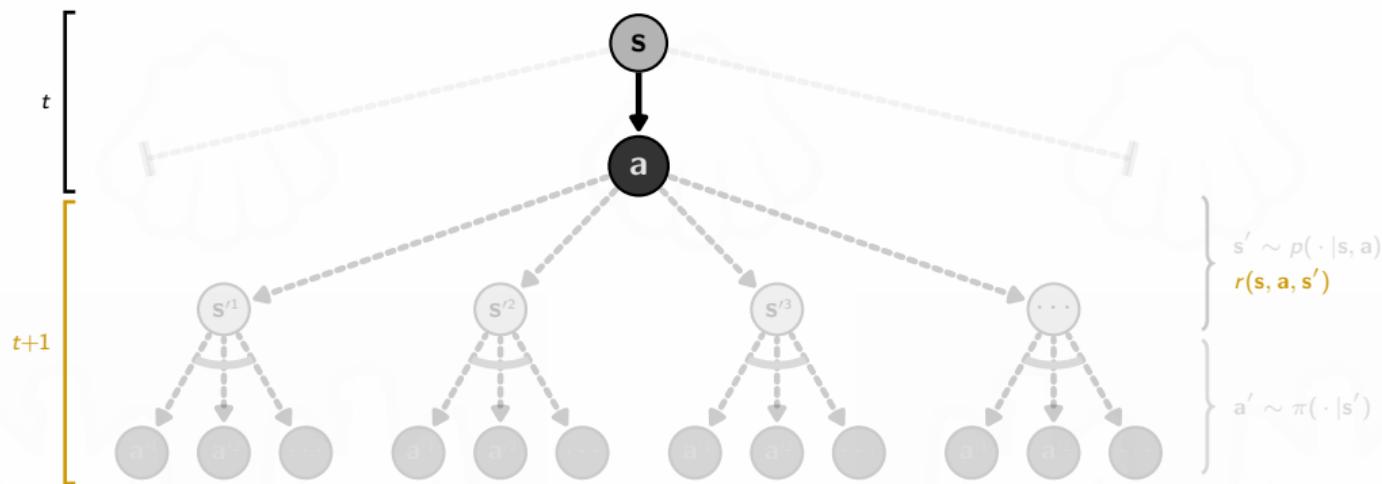


Diagramme de la valeur optimale d'une action   partir d'un   tat

$$Q^*(\mathbf{s}_t, \mathbf{a}_t) = \sum_{\mathbf{s}' \in \mathcal{S}_{t+1}} p(\mathbf{s}' | \mathbf{s}_t, \mathbf{a}_t) \left[r_{t+1} + \gamma \max_{\mathbf{a}' \in \mathcal{A}_{t+1}} Q^*(\mathbf{s}', \mathbf{a}') \right]$$

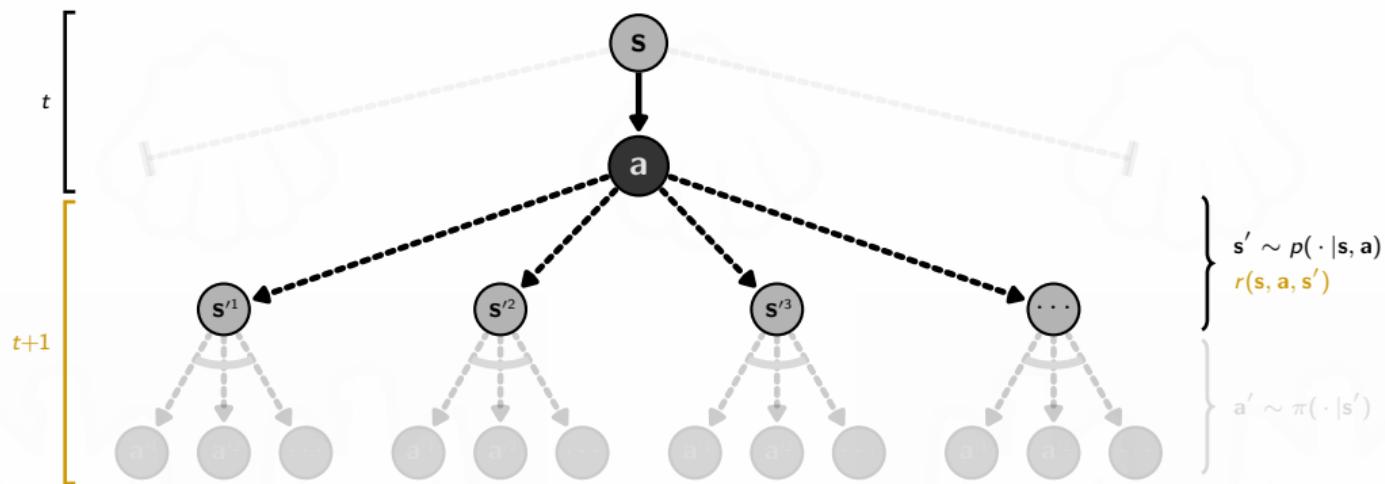


Diagramme de la valeur optimale d'une action   partir d'un   tat

$$Q^*(\mathbf{s}_t, \mathbf{a}_t) = \sum_{\mathbf{s}' \in \mathcal{S}_{t+1}} p(\mathbf{s}' | \mathbf{s}_t, \mathbf{a}_t) \left[r_{t+1} + \gamma \max_{\mathbf{a}' \in \mathcal{A}_{t+1}} Q^*(\mathbf{s}', \mathbf{a}') \right]$$

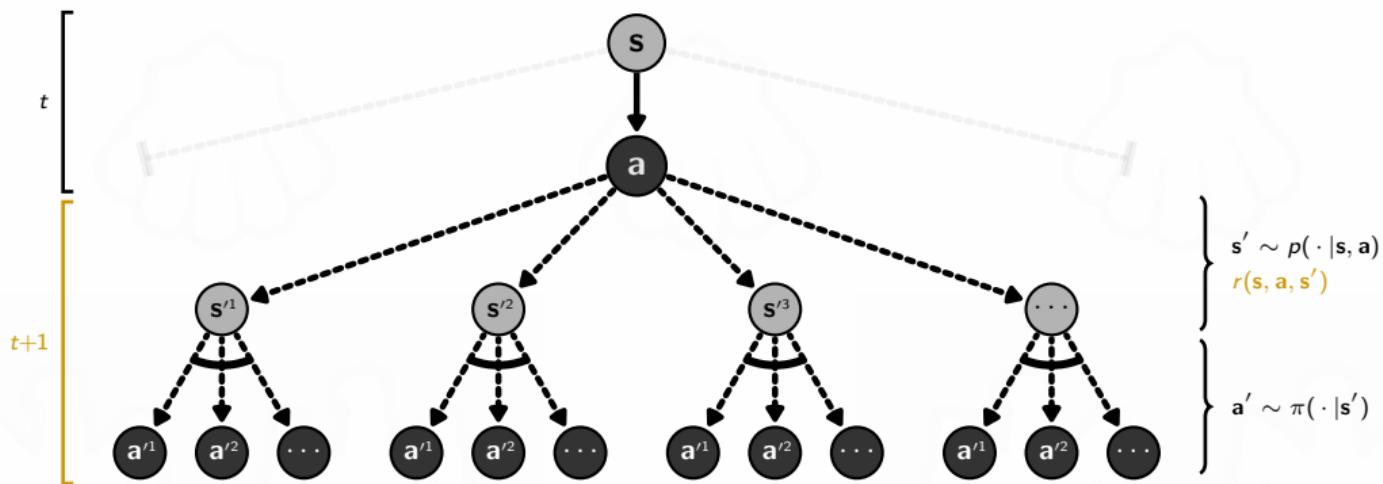
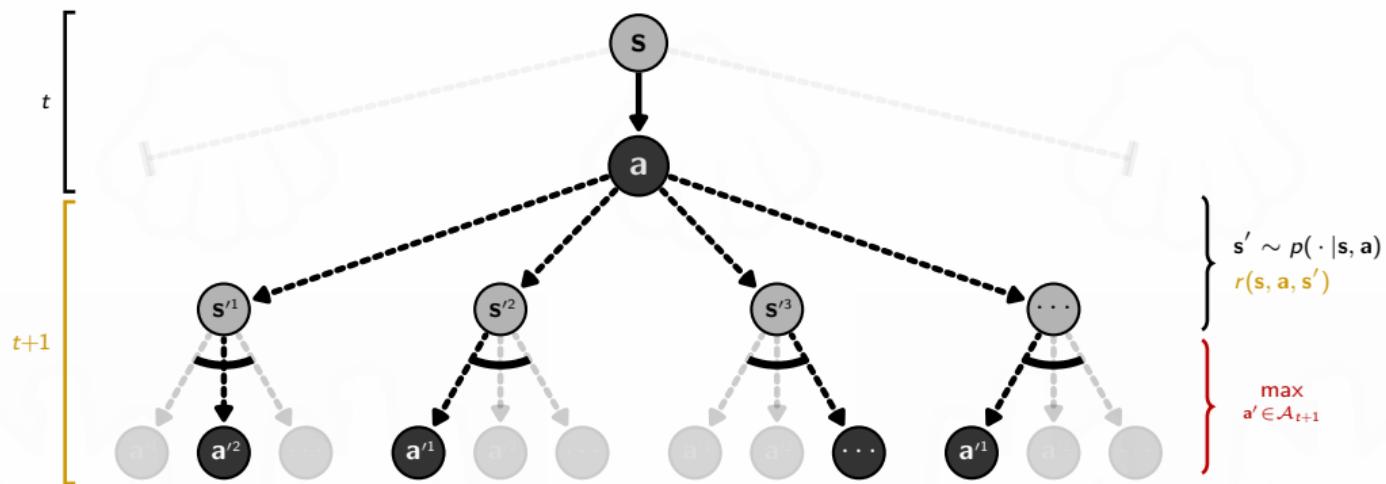


Diagramme de la valeur optimale d'une action   partir d'un   tat

$$Q^*(\mathbf{s}_t, \mathbf{a}_t) = \sum_{\mathbf{s}' \in \mathcal{S}_{t+1}} p(\mathbf{s}' | \mathbf{s}_t, \mathbf{a}_t) \left[r_{t+1} + \gamma \max_{\mathbf{a}' \in \mathcal{A}_{t+1}} Q^*(\mathbf{s}', \mathbf{a}') \right]$$



La fonction de valeur optimale $V^*(s_t)$

R pons   la question :   quel point cet  tat est bon **si on agit optimalement   partir de ce point** jusqu'  la fin de la trajectoire ?

$$\begin{aligned} V^*(s_t) &= \max_{\pi} V^\pi(s_t) \quad \text{pour tous } s \in \mathcal{S} \\ &= \max_{a \in \mathcal{A}_t} \mathbb{E}_{s_{t+1} \sim p(\cdot | s_t, a)} \left[r_{t+1} + \gamma V^*(s_{t+1}) \right] \\ &= \max_{a \in \mathcal{A}_t} \sum_{s' \in \mathcal{S}_{t+1}} p(s'|s_t, a) [r_{t+1} + \gamma V^*(s')] \end{aligned}$$

Diagramme de la valeur optimale d'un  t t

$$V^*(\mathbf{s}_t) = \max_{a \in A_t} \sum_{s' \in S_{t+1}} p(s'|s_t, a) [r_{t+1} + \gamma V^*(s')]$$

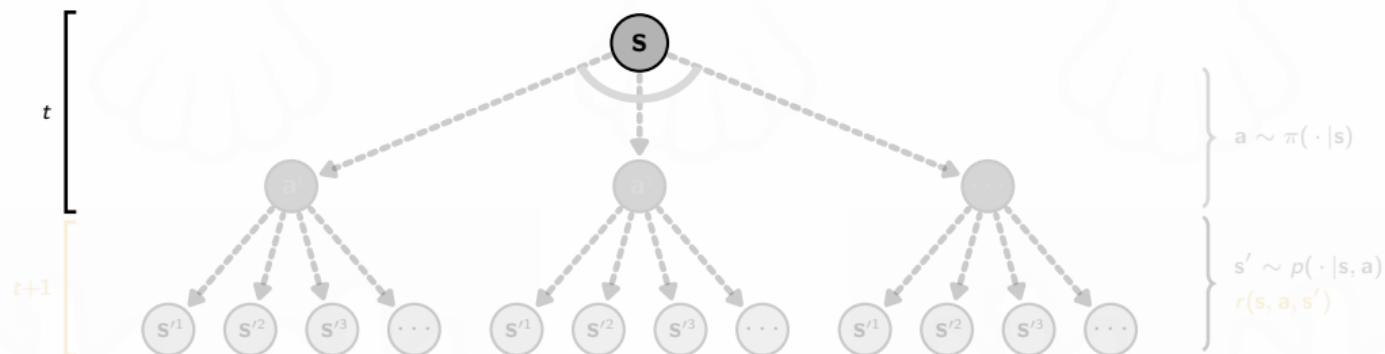


Diagramme de la valeur optimale d'un  t t

$$V^*(s_t) = \max_{a \in A_t} \sum_{s' \in S_{t+1}} p(s'|s_t, a) [r_{t+1} + \gamma V^*(s')]$$

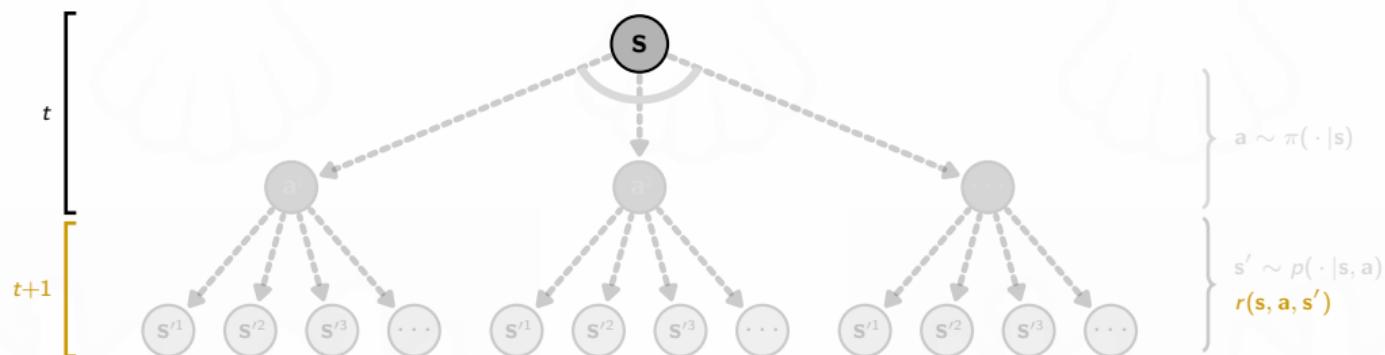


Diagramme de la valeur optimale d'un  t t

$$V^*(\mathbf{s}_t) = \max_{a \in \mathcal{A}_t} \sum_{s' \in \mathcal{S}_{t+1}} p(s'|s_t, a) [r_{t+1} + \gamma V^*(s')]$$

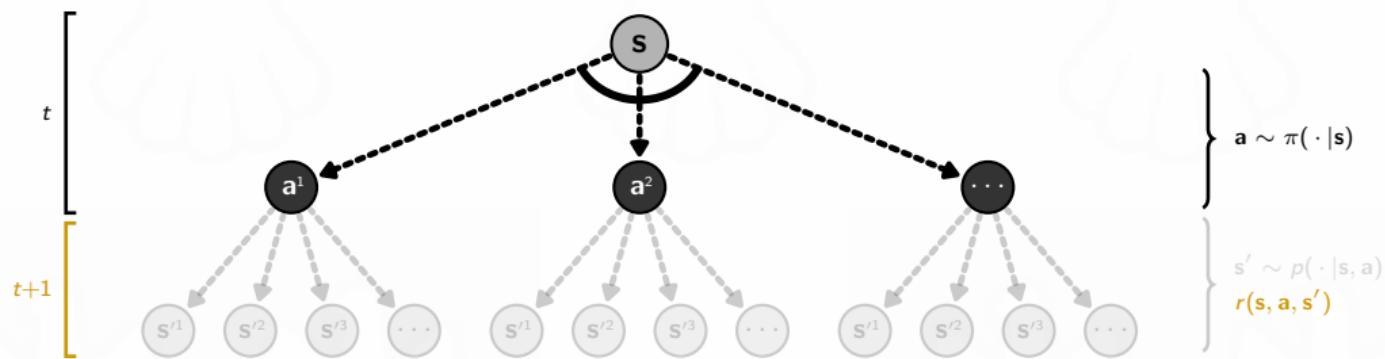


Diagramme de la valeur optimale d'un  t t

$$V^*(\mathbf{s}_t) = \max_{\mathbf{a} \in \mathcal{A}_t} \sum_{\mathbf{s}' \in \mathcal{S}_{t+1}} p(\mathbf{s}' | \mathbf{s}_t, \mathbf{a}) [r_{t+1} + \gamma V^*(\mathbf{s}')]$$

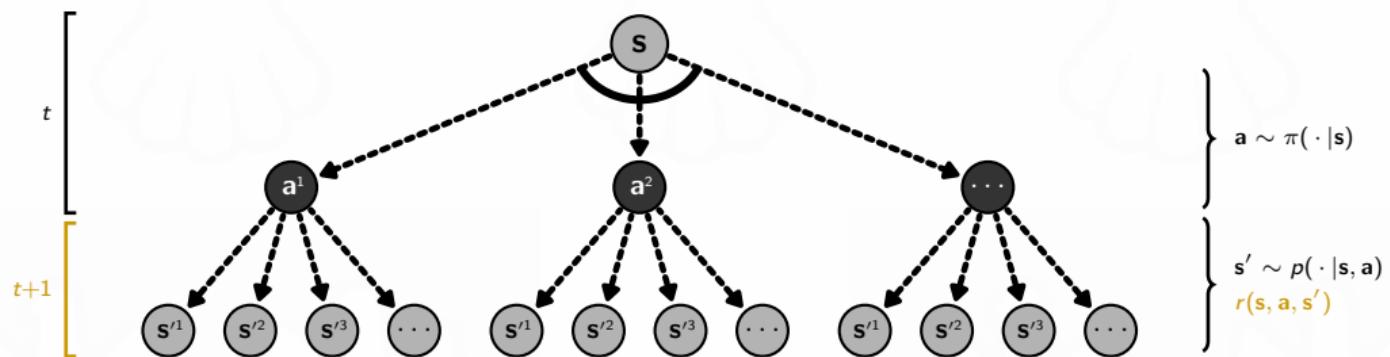
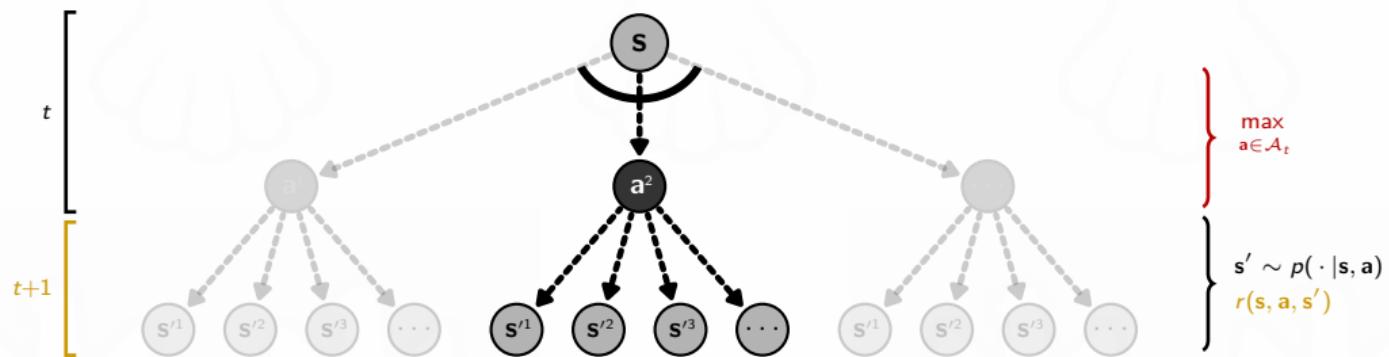


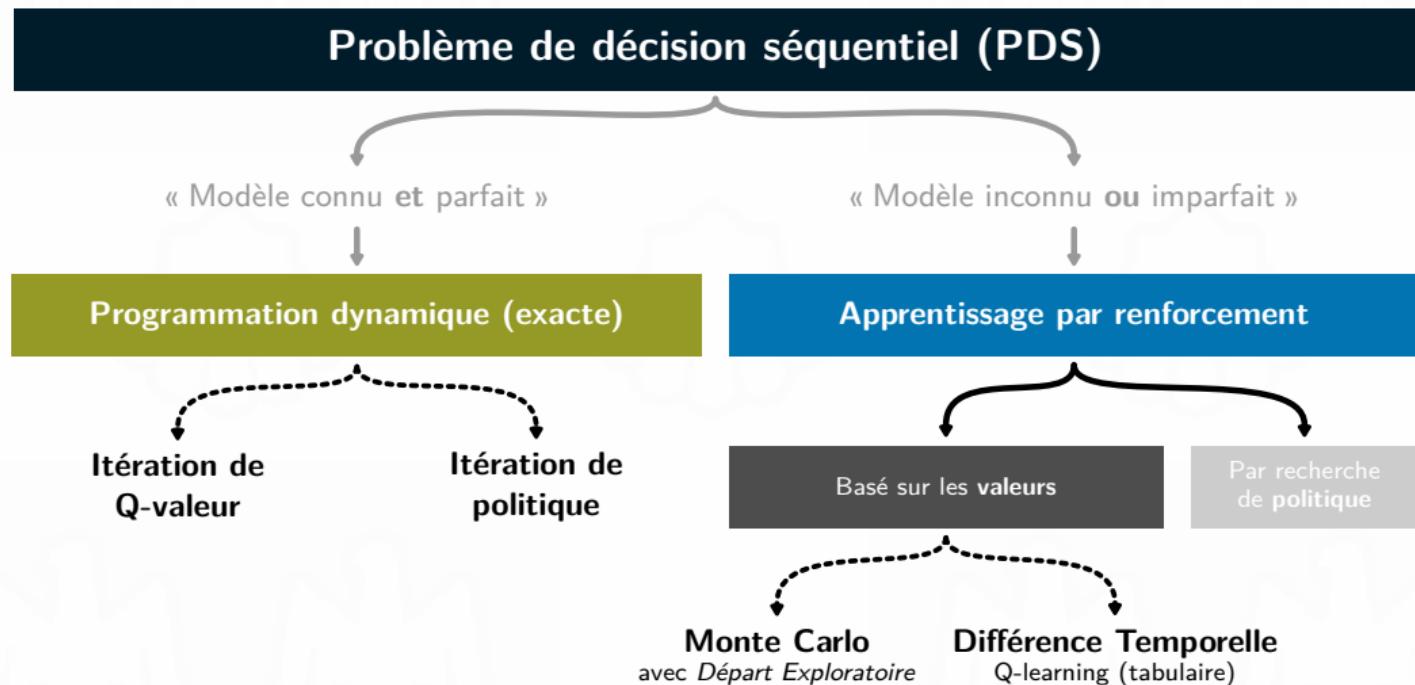
Diagramme de la valeur optimale d'un  t t

$$V^*(\mathbf{s}_t) = \max_{\mathbf{a} \in \mathcal{A}_t} \sum_{\mathbf{s}' \in \mathcal{S}_{t+1}} p(\mathbf{s}' | \mathbf{s}_t, \mathbf{a}) [r_{t+1} + \gamma V^*(\mathbf{s}')]$$



Prochaine partie

Prochaine partie



Pour aller plus loin

1 Vue d'ensemble

2 Formalisme et outil de l'apprentissage par renforcement

3 L'univers du point de vue de l'agent

4 Pour aller plus loin

- Complément théorique
- Références

Le futur prometteur de l'apprentissage par renforcement

- The Power of Self-Learning Systems - Pr sentation au *MIT + Center for Brains, Minds & Machines*, 2019 par Demis Hassabis, Co-Founder & CEO, *DeepMind*

▶ Vid o: The Power of Self-Learning Systems - 63 minutes

Pour aller plus loin

Complément théorique

Le Retour G sur un horizon infini

Contexte : cas o  l'environnement n'a **pas de terminaison** (par « design » ou en cas limite).

$T = \infty \implies$ la trajectoire τ ne peut pas  tre segment  en ** pisode**
 \implies **t che continue**

Probl me no 1 : Le retour $G_t(\tau)$ pourrait  tre infini

Solution : Utiliser un « discounted factor γ » tel que $\gamma < 1$
(Normalement : $0 \leq \gamma \leq 1$)

Dans la mesure o  il existe une limite   l'espace de r compense \mathcal{R} ,
alors le retour sur un horizon infini aura une valeur finie.¹

$$G_t(\tau) \doteq \sum_{t'=t}^{\infty} \gamma^{t'-t} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}, \mathbf{s}_{t'+1})$$

1. « Reinforcement Learning : An introduction », section 3.3 « Returns and Episodes » et 3.4 « Unified Notation for Episodic and Continuing Tasks » par Sutton & Barto [8]

Le Retour G sur un horizon infini

Contexte : cas o  l'environnement n'a **pas de terminaison** (par « design » ou en cas limite).

$T = \infty$ \implies la trajectoire τ ne peut pas  tre segment  en ** pisode**
 \implies **t che continue**

Probl me no 2 :  tant donn  que la trajectoire τ ne termine jamais,

-   quel moment commencer l'apprentissage ?
- comment structurer l'exp rience en vue de l'apprentissage ?

Solution : *Online learning method, Continual learning, Meta-RL, ...*

Lecture recommand 

Online learning : « Reinforcement Learning : An introduction », section 6.2 « Advantages of TD Prediction Methods » par Sutton & Barto comme exemple de m thode *Online*. [8]

Continual learning : L'introduction de « Progress & Compress : A scalable framework for continual learning » par Schwarz et al. (*DeepMind*) donne un bon aper  du domaine. [9]

Meta-RL : « Meta Reinforcement Learning » par Lilian Weng donne un bon aper  du domaine. Lil'Log blog est son blogue personnel. Elle est pr sentement chercheuse chez OpenAI (2020) [10]

Pour aller plus loin

Références

References I

1. BERSETH, G., PENG, X. B. & van de PANNE, M. Terrain RL Simulator. 1-10. arXiv : 1804.06424. <http://arxiv.org/abs/1804.06424> (2018).
2. PENG, X. B., BERSETH, G. & VAN DE PANNE, M. Terrain-adaptive locomotion skills using deep reinforcement learning. *ACM Transactions on Graphics* **35**, 1-15. ISSN : 15577368 (2016).
3. MNIH, V. *et al.* Playing Atari with Deep Reinforcement Learning. 1-9. arXiv : 1312.5602. <http://arxiv.org/abs/1312.5602> (2013).
4. MNIH, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529-533. ISSN : 14764687. <https://web.stanford.edu/class/psych209/Readings/MnihEtAlHassabis15NatureControlDeepRL.pdf> (2015).
5. VINYALS, O. *et al.* StarCraft II: A New Challenge for Reinforcement Learning. arXiv : 1708.04782. <http://arxiv.org/abs/1708.04782> (2017).
6. ABBEEL, P., COATES, A., QUIGLEY, M. & NG, A. Y. An Application of Reinforcement Learning to Aerobatic Helicopter Flight. *Advances in Neural Information Processing Systems* (2006).

References II

7. MAO, H., ALIZADEH, M., MENACHE, I. & KANDULA, S. Resource management with deep reinforcement learning. *HotNets 2016 - Proceedings of the 15th ACM Workshop on Hot Topics in Networks*, 50-56 (2016).
8. SUTTON, R. S. & BARTO, A. G. *Reinforcement learning: An introduction*. 2^e ´ed. (´ed. MIT PRESS) ISBN : 978-0262039246.
<http://incompleteideas.net/book/RLbook2018.pdf> (Cambridge, MA, 2018).
9. SCHWARZ, J. et al. Progress & compress: A scalable framework for continual learning. *35th International Conference on Machine Learning, ICML 2018* **10**, 7199-7208. arXiv : 1805.06370. <https://deepmind.com/research/publications/progress-compress-scalable-framework-continual-learning> (2018).
10. WENG, L. *Meta Reinforcement Learning (Lil'Log)*. 2019.
<http://lilianweng.github.io/lil-log/2019/06/23/meta-reinforcement-learning.html>.