

A2DI: Modèles pour l'apprentissage par renforcement

John Klein

Université de Lille - CRIStAL UMR CNRS 9189





- d'après un document d'Alessandro Lazaric -

Visitez sa [homepage](#)

[Home](#) [Publications](#) [Teaching](#) [Activities](#) [Projects](#)

Alessandro Lazaric



Alessandro Lazaric
Junior Researcher
SequeL Team

Welcome to my site

I received my PhD from the Electronic and Informatics Department of Politecnico di Milano, under the supervision of [Andrea Bonarini](#) and [Marcello Restelli](#).

I'm currently a Junior Researcher (CR1) at INRIA Lille - Nord Europe in the SequeL team led by [Philippe Preux](#) and [Rémi Munos](#).

You can find my (almost) updated CV [here](#).

My mean research topics are:

- [Reinforcement Learning](#)
- [Transfer Learning](#)
- [Multi-arm Bandit](#)
- [Online Learning](#)

I also keep on eye on:

- [Multiagent Learning](#)
- [Game Theory](#)
- [Mechanism Design](#)

Changeons un peu de sujet et parlons d'apprentissage par renforcement !

- On a plus d'exemples $\mathbf{x}^{(i)}$.
- On a plus de solutions associées $y^{(i)}$.
- .. mais on a une fonction de récompense r .

Dans ce chapitre, nous allons voir comment modéliser un problème d'apprentissage de cette nature.

Plan du chapitre

- 1 Généralités sur l'apprentissage par renforcement
- 2 Procéssus de décision Markovien
- 3 Fonction de Valeur
- 4 Conclusions

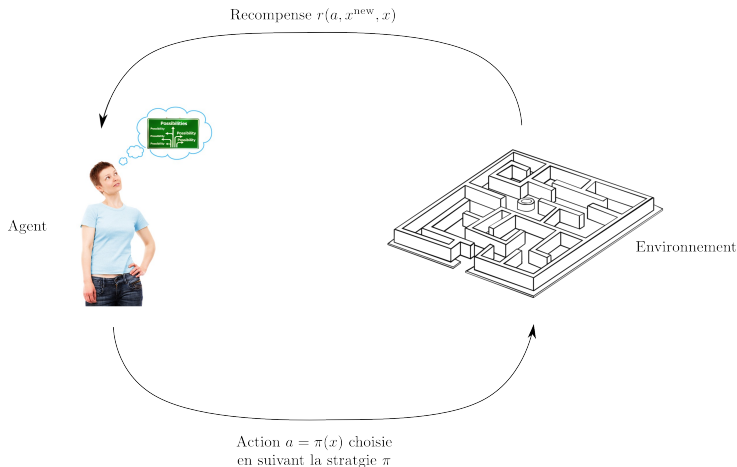
Définition :

Reinforcement learning is learning what to do – how to map situations to actions – so as to **maximize** a numerical **reward** signal in an **unknown uncertain** environment. The learner is not told which actions to take, as in most forms of machine learning, but she must discover which actions yield the most reward by **trying them** (**trial-and-error**). In the most interesting and challenging cases, actions may affect not only the immediate reward but also the **next situation** and, through that, all **subsequent rewards** (**delayed reward**).

“An introduction to reinforcement learning”,
Sutton and Barto (1998).

RL : un concept né de la **psychologie** animale (Pavlov)

.. puis adapté en **automatique** et en **informatique**



RL : comparaison avec les autres paradigmes :

- L'apprentissage supervisé offre de bonnes performances, mais la supervision coûte cher !
- L'apprentissage non supervisé n'offre pas de garanties de performances.
- L'apprentissage par renforcement offre des garanties tout en nécessitant une forme de supervision beaucoup plus faible.

Plutôt que d'opposer les paradigmes → comprendre qu'ils ne s'attaquent pas au même problème de départ !

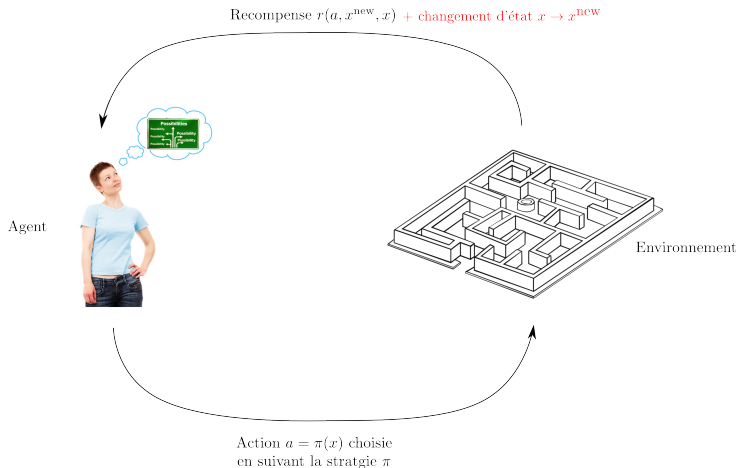
RL : Quelle démarche ?

- Trouver comment **modéliser** le problème.
- Trouver comment **résoudre exactement** le problème.
- Trouver comment **résoudre incrémentalement** le problème.
- Trouver comment **résoudre efficacement** le problème.
- Trouver comment **résoudre approximativement** le problème.

Plan du chapitre

- 1 Généralités sur l'apprentissage par renforcement
- 2 Processus de décision Markovien**
- 3 Fonction de Valeur
- 4 Conclusions

RL : soyons un peu plus précis sur le mécanisme :



a est une **action**. x_t est l'**état** dans lequel on est à l'instant t (suite à nos actions). π est la **politique** (stratégie) qu'on a suivi pour choisir nos actions.

Chaînes de Markov



Pause proba !

Définition (Chaîne de Markov)

Soit X l'espace d'état un sous-ensemble borné compact d'un espace Euclidien, le système dynamique à temps discret $(x_t)_{t \in \mathbb{N}} \in X$ est une chaîne de Markov si il satisfait la propriété de Markov :

$$\mathbb{P}(x_{t+1} = x \mid x_t, x_{t-1}, \dots, x_0) = \mathbb{P}(x_{t+1} = x \mid x_t),$$

Etant donné un état initial $x_0 \in X$, la chaîne de Markov est entièrement caractérisée par les probabilité de transition p

$$p(y|x) = \mathbb{P}(x_{t+1} = y \mid x_t = x).$$

Processus de Décision Markovien

Définition (Processus de Décision Markovien)

Un **Processus de Décision Markovien** est un quadruplet $M = (X, A, p, r)$ où

- X est l'espace d'état,
- A est l'espace d'action,
- $p(y|x, a)$ est la probabilité de transition avec

$$p(y|x, a) = \mathbb{P}(x_{t+1} = y | x_t = x, a_t = a),$$

- $r(x, a, y)$ est la fonction de récompense suite à la transition (x, a, y) .

Processus de Décision Markovien : hypothèses nécessaires

- **hypothèse temporelle** : le temps est discret (non continu)

$$t \rightarrow t + 1$$

- **hypothèse Markovienne** : l'état courant x et l'action a sont des statistiques suffisantes pour le prochain état y

$$p(y|x, a) = \mathbb{P}(x_{t+1} = y | x_t = x, a_t = a)$$

- **hypothèse de récompense** : la récompense est entièrement définie par tout ou partie des transitions

$$r(x, a, y)$$

- **hypothèse de stationnarité** : la dynamique et la récompense n'évoluent pas dans le temps

$$p(y|x, a) = \mathbb{P}(x_{t+1} = y | x_t = x, a_t = a) \quad r(x, a, y)$$

Est-ce que le formalisme du MDP¹ est suffisamment puissant ?

\Rightarrow Essayons !

Processus de Décision Markovien : Exemple

Exemple : gestion du stock d'un magasin

A chaque mois t , un magasin contient x_t exemplaires d'un article et la demande pour cet article est D_t . A la fin de chaque mois, le manager peut commander a_t exemplaires de plus à son fournisseur. De plus, on sait que :

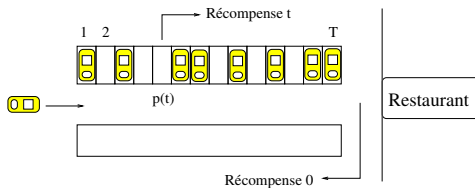
- le coût du stockage de x exemplaires est $h(x)$,
- le coût de l'achat de a exemplaires est $C(a)$,
- le revenu des ventes de q exemplaires est $f(q)$,
- si la demande D est supérieure à la disponibilité x , les clients qui n'ont pu être servis partent.
- le nombre d'exemplaires à la fin de l'année est $g(x)$.
- **Contrainte** : la capacité maximum de stockage est M .

Processus de Décision Markovien : Exemple

- **Espace d'état** : $x \in X = \{0, 1, \dots, M\}$.
- **Espace d'action** : il n'est pas possible de commander plus d'exemplaires que la capacité du magasin, donc l'espace d'action dépend de l'état courant. Formellement, à l'état x , $a \in A(x) = \{0, 1, \dots, M - x\}$.
- **Dynamique** : $x_{t+1} = [x_t + a_t - D_t]^+$.
Problème : la dynamique doit être Markovienne et stationnaire !
- La demande D_t est **stochastique et dépend du temps**. Formellement, $D_t \stackrel{i.i.d.}{\sim} \mathcal{L}oi$.
- **Récompense** : $r_t = -C(a_t) - h(x_t + a_t) + f([x_t + a_t - x_{t+1}]^+)$.

Processus de Décision Markovien : Exercice

Le problème du parking Un conducteur veut garer sa voiture le plus près possible du restaurant.



- Le conducteur ne peut voir si une place est libre avant d'être devant.
- Il y a T places.
- A chaque place i , le conducteur peut soit continuer soit se garer (si la place est libre).
- Plus on est près du restaurant, plus la satisfaction est grande.
- Si le conducteur ne trouve pas de place, il part chercher un autre restaurant.

Processus de Décision Markovien : Exercice

Le problème du parking

Processus de Décision Markovien : Politique (ou stratégie)

Définition (Politique)

Une **règle de décision** π_t peut être

- **Déterminististe** : $\pi_t : X \rightarrow A$,
- **Stochastique** : $\pi_t : X \rightarrow \Delta(A)$,

Une **politique** (strategie, plan) peut être

- **Non-stationnaire** : $\pi = (\pi_0, \pi_1, \pi_2, \dots)$,
- **Stationnaire (Markoviennne)** : $\pi = (\pi, \pi, \pi, \dots)$.

Rq : MDP M + politique stationnaire $\pi \Rightarrow$ **chaîne de Markov** des états X_t avec pour probabilité de transition $p(y|x) = p(y|x, \pi(x))$.

Processus de Décision Markovien : Politique (ou stratégie)

Exemple : gestion du stock d'un magasin

- Politique stationnaire n°1

$$\pi(x) = \begin{cases} M - x & \text{si } x < M/4 \\ 0 & \text{sinon} \end{cases}$$

- Politique stationnaire n°2

$$\pi(x) = \max\{(M - x)/2 - x; 0\}$$

- Politique non-stationnaire

$$\pi_t(x) = \begin{cases} M - x & \text{si } t < 6 \\ \lfloor (M - x)/5 \rfloor & \text{sinon} \end{cases}$$

Plan du chapitre

- 1 Généralités sur l'apprentissage par renforcement
- 2 Procéssus de décision Markovien
- 3 Fonction de Valeur**
- 4 Conclusions

Comment évaluer une politique - comparer deux politiques ?

⇒ fonction de valeur !

Optimalité à horizon d'une politique

- Horizon finie T : *deadline* au temps T , l'agent se focalise sur la récompense cumulée jusqu'à T .
- Horizon infinie avec affaiblissement : le problème est sans fin mais les récompenses à court terme ont une importance plus élevée.
- Horizon infinie avec états terminaux : le problème prend fin mais l'agent va atteindre un état terminal.
- Horizon infinie avec récompense moyenne : le problème est sans fin mais l'agent se focalise seulement sur la moyenne des récompenses (attendue).

Fonction de valeur d'une politique

- **Horizon finie T** : *deadline* au temps T , l'agent se focalise sur la récompense cumulée jusqu'à T .

$$V^\pi(t, x) = \mathbb{E} \left[\sum_{s=t}^{T-1} r(x_s, \pi_s(x_s)) + R(x_T) \mid x_t = x; \pi \right],$$

où R est la fonction de valeur de l'état final.

- **Utile quand** : il y a une *deadline* explicite à respecter.

Fonction de valeur d'une politique

- **Horizon infini avec affaiblissement** : le problème est sans fin mais les récompenses à **court terme** ont une importance plus élevée.

$$V^{\pi}(x) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(x_t, \pi(x_t)) \mid x_0 = x; \pi \right],$$

avec facteur d'affaiblissement $0 \leq \gamma < 1$:

- **petit** = récompense à court terme, **grand** = récompense à long terme
- pour tout $\gamma \in [0, 1)$ la suite converge toujours (pour des récompenses bornées)
- **Utile quand** : la *deadline* est incertaine et/ou le problème fait explicitement référence à un facteur d'affaiblissement.

Fonction de valeur d'une politique

- **Horizon infini avec états terminaux** : le problème prend fin mais l'agent va atteindre un **état terminal**.

$$V^{\pi}(x) = \mathbb{E} \left[\sum_{t=0}^T r(x_t, \pi(x_t)) | x_0 = x; \pi \right],$$

où T est le 1^{er} temps (**aléatoire**) où un **état terminal** est atteint.

- **Utile quand** : il y a un but connu ou une condition d'échec.

Fonction de valeur d'une politique

- **Horizon infinie avec récompense moyenne** : le problème est sans fin mais l'agent se focalise seulement sur la **moyenne des récompenses** (attendue).

$$V^\pi(x) = \lim_{T \rightarrow \infty} \mathbb{E} \left[\frac{1}{T} \sum_{t=0}^{T-1} r(x_t, \pi(x_t)) \mid x_0 = x; \pi \right].$$

- **Utile quand** : le système a besoin d'être constamment contrôlé au fil du temps.

Fonction de valeur d'une politique

Note : les espérances sont prises sous $p_{X_1, \dots, X_t | X_0 = x_0; \pi}$.

Une politique non-stationnaire π partant de x_0 donne

$$(x_0, r_0, x_1, r_1, x_2, r_2, \dots)$$

où $r_t = r(x_t, \pi_t(x_t))$ et $x_t \sim p(\cdot | x_{t-1}, a_t = \pi(x_t))$ sont des réalisations **aléatoires**. La fonction de valeur (affaiblie à horizon infinie) est

$$V^\pi(x) = \mathbb{E}_{(x_1, x_2, \dots)} \left[\sum_{t=0}^{\infty} \gamma^t r(x_t, \pi(x_t)) \mid x_0 = x; \pi \right],$$

Fonction de valeur d'une politique

Exemple : gestion du stock d'un magasin

Simulation

Fonction de valeur optimale et politique optimale

Definition (Politique optimale et fonction de valeur optimale)

La solution d'un MDP est une *politique optimale* π^* satisfaisant

$$\pi^* \in \arg \max_{\pi \in \Pi} V^\pi$$

en tout état $x \in X$, où Π l'ensemble des politiques considérées.

La fonction de valeur correspondante est la *fonction de valeur optimale*

$$V^* = V^{\pi^*}$$

Fonction de valeur optimale et politique optimale

Remarques

- 1 $\pi^* \in \arg \max(\cdot)$ et non $\pi^* = \arg \max(\cdot)$ car un MDP peut admettre **plusieurs** politiques optimales.
- 2 π^* atteint la plus grande fonction de valeur possible en **chaque** état.
- 3 Il existe toujours une politique optimale **deterministe**.
- 4 Hormis pour des problèmes à horizon finie, il existe toujours une politique optimale **stationnaire**.

Messages importants du chapitre :

- un MDP est un modèle puissant pour représenter l'interaction entre un agent et un environnement stochastique.
- La fonction de valeur est l'objectif que l'on devra maximiser (voir Chap. suivant).