

# Apprentissage par renforcement

Sergey Kirgizov

Université de Bourgogne

October 2016

# Exemple motivant

## Jeu de loup



Apprentissage par  
renforcement

Sergey Kirgizov

Intro

Historique

Théorie

TD-learning

Q-learning

Applications

Jeu du loup en R

En TP / Home  
work

Recherche actuelle

# Exemple motivant

## Jeu de loup

Apprentissage par  
renforcement

Sergey Kirgizov

Intro

Historique

Théorie

TD-learning

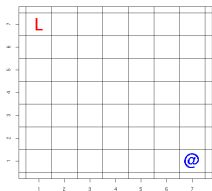
Q-learning

Applications

Jeu du loup en R

En TP / Home  
work

Recherche actuelle



Pas bien pour le loup ↑

# Exemple motivant

## Jeu de loup

Apprentissage par  
renforcement

Sergey Kirgizov

### Intro

Historique

Théorie

TD-learning

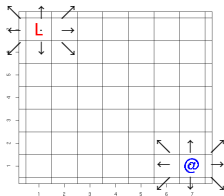
Q-learning

Applications

Jeu du loup en R

En TP / Home  
work

Recherche actuelle



Pas bien pour le loup ↑

# Exemple motivant

## Jeu de loup

Apprentissage par  
renforcement

Sergey Kirgizov

Intro

Historique

Théorie

TD-learning

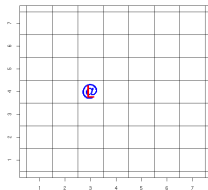
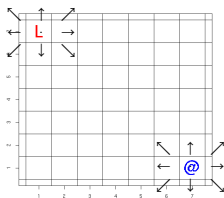
Q-learning

Applications

Jeu du loup en R

En TP / Home  
work

Recherche actuelle



Pas bien pour le loup ↑

↑ Bien pour le loup

# Idée l'apprentissage par renforcement

Apprentissage par  
renforcement

Sergey Kirgizov

**Intro**

Historique

Théorie

TD-learning

Q-learning

Applications

Jeu du loup en R

En TP / Home  
work

Recherche actuelle

1988 TD-learning, Richard Sutton



1989 Q-learning, Chris Watkins



“Toutefois, l’origine de l’apprentissage par renforcement est plus ancienne. Elle dérive de formalisations théoriques de méthodes de contrôle optimal, visant à mettre au point un contrôleur permettant de minimiser au cours du temps une mesure donnée du comportement d’un système dynamique. La version discrète et stochastique de ce problème est appelée un processus de décision markovien et fut introduite par Richard Ernest Bellman en 1957”



[https://fr.wikipedia.org/wiki/Apprentissage\\_par\\_renforcement](https://fr.wikipedia.org/wiki/Apprentissage_par_renforcement)



# Théories de psychologie animale → Intelligence artificielle

Apprentissage par  
renforcement

Sergey Kirgizov

Intro

Historique

Théorie

TD-learning

Q-learning

Applications

Jeu du loup en R

En TP / Home  
work

Recherche actuelle

“D’autre part, la formalisation des problèmes d’apprentissage par renforcement s’est aussi beaucoup inspirée de théories de psychologie animale, comme celles analysant comment un animal peut apprendre par essais-erreurs à s’adapter à son environnement. Ces théories ont beaucoup inspiré le champ scientifique de l’intelligence artificielle et ont beaucoup contribué à l’émergence d’algorithmes d’apprentissage par renforcement au début des années 1980.”

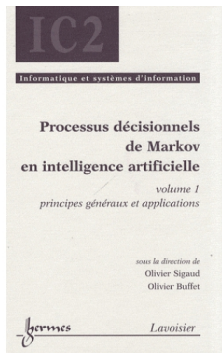
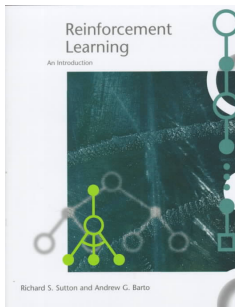
— Wikipedia

## 1. Reinforcement Learning : An Introduction

Richard S. Sutton et Andrew G. Barto

## 2. Processus décisionnels de Markov en intelligence artificielle

Olivier Sigaud et Olivier Buffet



## Markov Decision Process (MDP)

Finite and small are easy. (better than brute force ???)

But large and (infinity) are difficult. How to explore?

balance between explorations and actions

Q-learning is an example of a stochastic approximation algorithm

$S$  est un ensemble d'états

$A$  est un ensemble d'actions

$\Pi : S \rightarrow A$  est une politique

rules of transitioning between states ;

rules that determine the scalar immediate reward of a transition ;

and rules that describe what the agent observes.

TODO : cite some books

## !!! Q-Learning is an Off-Policy algorithm for Temporal Difference learning !!!

Types of Reinforcement Learning :

1. Bruteforce
2. TD-learning
3. Q-learning
4. Bellman equation ?
5. other stuff

bla-bla-bla???? really do this?

https:

[//en.wikipedia.org/wiki/Temporal\\_difference\\_learning](https://en.wikipedia.org/wiki/Temporal_difference_learning)

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t)(1 - \alpha) + \alpha \left( r_{t+1} + \gamma \max_a Q_t(s_{t+1}, a) \right)$$

Intro

Historique

Théorie

TD-learning

**Q-learning**

Applications

Jeu du loup en R

En TP / Home  
work

Recherche actuelle



# Optimal policy...

Apprentissage par  
renforcement

Sergey Kirgizov

Intro

Historique

Théorie

TD-learning

**Q-learning**

Applications

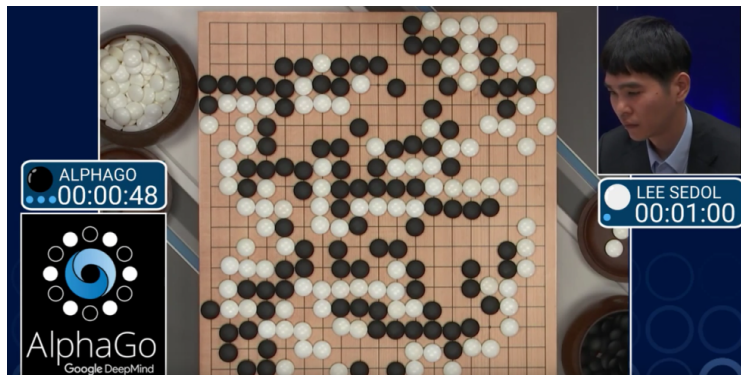
Jeu du loup en R

En TP / Home  
work

Recherche actuelle

Stochastic approximation algorithms by Robbins-Monro,  
or Kiefer-Wolfowitz, Blume from 1950-th..

$\sum_{n=0}^{\infty}$  from 1950-th...



“This program was based on general-purpose AI methods [including reinforcement learning], using deep neural networks to mimic expert players, and further improving the program by learning from games played against itself.”

## Using Reinforcement Learning for Autonomic Resource Allocation in Clouds: Towards a Fully Automated Workflow

Xavier Dutreilh<sup>†</sup>, Sergey Kirgizov\*, Olga Melekhova\*, Jacques Malenfant\*, Nicolas Rivier<sup>‡</sup> and Isis Truck<sup>‡</sup>

\*Université Pierre et Marie Curie – Paris 6 CNRS, UMR 7606 LIP6, 4 place Jussieu, Paris, 75005, France

Email: {Olga.Melekhova, Jacques.Malenfant}@lip6.fr

<sup>†</sup>Orange Labs, 38-40 rue du Général Leclerc, Issy-les-Moulineaux, 92130, France

Email: {xavier.dutreilh, nicolas.rivier}@orange-ftgroup.com

<sup>‡</sup>LIASD – EA 4383, Université Paris 8, 2 rue de la Liberté, Saint-Denis Cedex, 93526, France

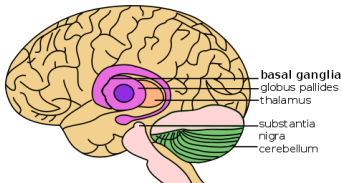
Email: truck@ai.univ-paris8.fr

**Abstract**—Dynamic and appropriate resource dimensioning is a crucial issue in cloud computing. As applications go more and more 24/7, online policies must be sought to balance performance with the cost of allocated virtual machines. Most industrial approaches to date use *ad hoc* manual policies, such as threshold-based ones. Providing good thresholds proved to be tricky and hard to automatize to fit every application requirement. Research is being done to apply automatic decision-making approaches, such as reinforcement learning. Yet, they face a lot of problems to go to the field: having good policies in the early phases of learning, time for the learning to converge to an optimal policy and coping with changes in the application performance behavior over time. In this paper, we propose to deal with these problems using appropriate initialization for the early stages as well as convergence speedups applied throughout the learning phases and we present our first experimental results for these. We also introduce a performance model change detection on which we are currently working to complete the learning process management. Even though some of these proposals were known in the reinforcement learning field, the key contribution of this paper is to integrate them in a real cloud controller and to program them as an automated workflow.

they don't require the *a priori* knowledge of the application performance model, but rather learn it as the application runs. Yet, RL faces a lot of problems to go to the field [3][4], such as: having good policies in the early phases of learning, time for the learning to converge to an optimal policy and coping with changes in the application performance behavior over time. In this paper, we propose to deal with these problems using appropriate initialization for the early stages, convergence speedups applied throughout the learning phases and performance model change detection. Even though some of these proposals were known in the RL field, the key contribution of this paper is to integrate them in a real cloud controller and to program them as an automated workflow.

We present our first results towards this automated learning management workflow. Section II introduces the resource allocation problem for cloud computing. Section III presents the formulation of the problem in the Q-learning framework, as we have modeled it for a private cloud deployed at Orange Labs. Section IV then presents the core contribution of the

Basal Ganglia and Related  
Structures of the Brain



“La collaboration entre neurobiologistes et chercheurs en intelligence artificielle a permis de découvrir qu’une partie du cerveau fonctionnait de façon très similaire aux algorithmes d’apprentissage par renforcement.”

— Wikipedia

**A Model of How the Basal Ganglia Generate and Use Neural Signals that Predict Reinforcement**

James C. Houk, J. L. Adams, A. G. Barto

# Jeu de loup

en R

Apprentissage par  
renforcement

Sergey Kirgizov

Intro

Historique

Théorie

TD-learning

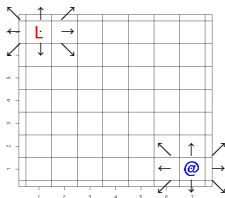
Q-learning

Applications

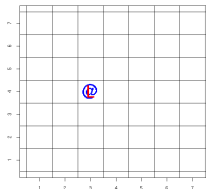
**Jeu du loup en R**

En TP / Home  
work

Recherche actuelle



Pas bien pour le loup ↑



↑ Bien pour le loup

1. Jouer avec le code
2. Trouvez et décrivez les différences entre `q-loup.R` et `q-loup-1.R`
3. Pourquoi `q-loup-1.R` apprend mieux le comportement cyclique du chat?
4. Visualisez l'évolution de valeurs de  $Q$
5. Change  $\text{ALPHA} = 0.6$  dans le code par quelque chose qui respecte la condition  $\sum_{i=0}^{\infty} \alpha_i < \infty$
6. Ajouter quelques murs dans l'environnement.

- ▶ Environnements sont plus grands
- ▶ Il faut augmenter la vitesse de convergence
- ▶ Combinaison avec d'autres techniques d'apprentissage par machine (les réseaux de neurones, etc)