



Introduction to Reinforcement Learning

A. LAZARIC (*SequeL Team @INRIA-Lille*)
ENS Cachan - Master 2 MVA

SequeL – INRIA Lille

A Bit of History

From Psychology to Machine Learning



The law of effect [Thorndike, 1911]

*“Of several responses made to the same situation, those which are accompanied or closely followed by **satisfaction** to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by **discomfort** to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur.*

The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.”

Experimental psychology

- ▶ *Classical (human and) animal conditioning*: “the magnitude and timing of the conditioned response changes as a result of the contingency between the conditioned stimulus and the unconditioned stimulus” [Pavlov, 1927].

Experimental psychology

- ▶ *Classical (human and) animal conditioning*: “the magnitude and timing of the conditioned response changes as a result of the contingency between the conditioned stimulus and the unconditioned stimulus” [Pavlov, 1927].
- ▶ *Operant conditioning (or instrumental conditioning)*: process by which humans and animals *learn* to behave in such a way as to obtain *rewards* and avoid *punishments* [Skinner, 1938].

Experimental psychology

- ▶ *Classical (human and) animal conditioning*: “the magnitude and timing of the conditioned response changes as a result of the contingency between the conditioned stimulus and the unconditioned stimulus” [Pavlov, 1927].
- ▶ *Operant conditioning (or instrumental conditioning)*: process by which humans and animals *learn* to behave in such a way as to obtain *rewards* and avoid *punishments* [Skinner, 1938].

Remark: **reinforcement** denotes any form of conditioning, either positive (*rewards*) or negative (*punishments*).

Computational neuroscience

- ▶ *Hebbian learning*: development of formal models of how the synaptic weights between neurons are reinforced by simultaneous activation. *“Cells that fire together, wire together.”* [Hebb, 1961].

Computational neuroscience

- ▶ *Hebbian learning*: development of formal models of how the synaptic weights between neurons are reinforced by simultaneous activation. “*Cells that fire together, wire together.*” [Hebb, 1961].
- ▶ *Emotions theory*: model on how the emotional process can bias the decision process [Damasio, 1994].

Computational neuroscience

- ▶ *Hebbian learning*: development of formal models of how the synaptic weights between neurons are reinforced by simultaneous activation. “*Cells that fire together, wire together.*” [Hebb, 1961].
- ▶ *Emotions theory*: model on how the emotional process can bias the decision process [Damasio, 1994].
- ▶ *Dopamine and basal ganglia model*: direct link with motor control and decision-making (e.g., [Doya, 1999]).

Computational neuroscience

- ▶ *Hebbian learning*: development of formal models of how the synaptic weights between neurons are reinforced by simultaneous activation. “*Cells that fire together, wire together.*” [Hebb, 1961].
- ▶ *Emotions theory*: model on how the emotional process can bias the decision process [Damasio, 1994].
- ▶ *Dopamine and basal ganglia model*: direct link with motor control and decision-making (e.g., [Doya, 1999]).

Remark: **reinforcement** denotes the effect of dopamine (and surprise).

Optimal control theory and dynamic programming

- ▶ *Optimal control*: *formal framework* to define optimization methods to derive control policies in continuous time control problems [Pontryagin and Neustadt, 1962].

Optimal control theory and dynamic programming

- ▶ *Optimal control*: *formal framework* to define optimization methods to derive control policies in continuous time control problems [Pontryagin and Neustadt, 1962].
- ▶ *Dynamic programming*: set of methods used to *solve control problems* by decomposing them into subproblems so that the optimal solution to the global problem is the conjunction of the solutions to the subproblems [Bellman, 2003].

Optimal control theory and dynamic programming

- ▶ *Optimal control*: *formal framework* to define optimization methods to derive control policies in continuous time control problems [Pontryagin and Neustadt, 1962].
- ▶ *Dynamic programming*: set of methods used to *solve control problems* by decomposing them into subproblems so that the optimal solution to the global problem is the conjunction of the solutions to the subproblems [Bellman, 2003].

Remark: **reinforcement** denotes an objective function to maximize (or minimize).

Reinforcement learning

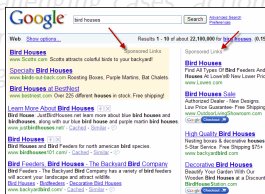
Reinforcement learning is learning what to do – how to map situations to actions – so as to **maximize** a numerical **reward** signal in an **unknown uncertain** environment. The learner is not told which actions to take, as in most forms of machine learning, but she must discover which actions yield 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).*

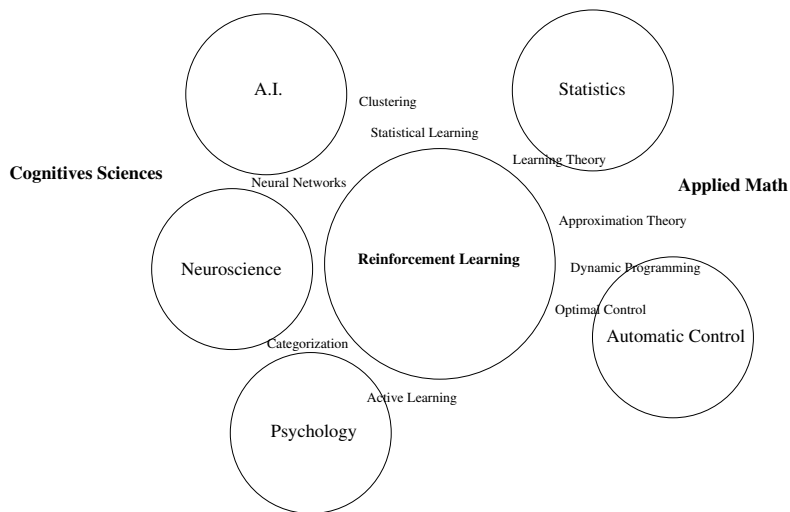
Reinforcement learning



Reinforcement learning is learning with only the immediate and delayed reward). In the most interesting and challenging cases, action selection is only the most interesting and challenging part of the problem (Sutton & Barto (1998)).



A Multi-disciplinary Field



A Machine Learning Paradigm

- ▶ *Supervised learning*: an expert (*supervisor*) provides examples of the right strategy (e.g., classification of clinical images).
Supervision is expensive.

A Machine Learning Paradigm

- ▶ *Supervised learning*: an expert (*supervisor*) provides examples of the right strategy (e.g., classification of clinical images).
Supervision is expensive.
- ▶ *Unsupervised learning*: different objects are clustered together by similarity (e.g., clustering of images on the basis of their content). *No actual performance is optimized.*

A Machine Learning Paradigm

- ▶ *Supervised learning*: an expert (*supervisor*) provides examples of the right strategy (e.g., classification of clinical images). *Supervision is expensive.*
- ▶ *Unsupervised learning*: different objects are clustered together by similarity (e.g., clustering of images on the basis of their content). *No actual performance is optimized.*
- ▶ *Reinforcement learning*: learning by direct interaction (e.g., autonomous robotics). *Minimum level of supervision (reward) and maximization of long term performance.*

The Problems

How to *model* an RL problem

The Problems

How to *model* an RL problem

How to solve *exactly* an RL problem

The Problems

How to *model* an RL problem

How to solve *exactly* an RL problem

How to solve *incrementally* an RL problem

The Problems

How to *model* an RL problem

How to solve *exactly* an RL problem

How to solve *incrementally* an RL problem

How to *efficiently* explore in an RL problem

The Problems

How to *model* an RL problem

How to solve *exactly* an RL problem

How to solve *incrementally* an RL problem

How to *efficiently* explore in an RL problem

How to solve *approximately* an RL problem

Bibliography I



Bellman, R. (2003).

Dynamic Programming.

Dover Books on Computer Science Series. Dover Publications, Incorporated.



Damasio, A. R. (1994).

Descartes' Error: Emotion, Reason and the Human Brain.

Grosset/Putnam.



Doya, K. (1999).

What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex.

Neural Networks, 12:961–974.



Hebb, D. O. (1961).

Distinctive features of learning in the higher animal.

In Delafresnaye, J. F., editor, *Brain Mechanisms and Learning*. Oxford University Press.



Pavlov, I. (1927).

Conditioned reflexes.

Oxford University Press.

Bibliography II



Pontryagin, L. and Neustadt, L. (1962).

The Mathematical Theory of Optimal Processes.

Number v. 4 in Classics of Soviet Mathematics. Gordon and Breach Science Publishers.



Skinner, B. F. (1938).

The behavior of organisms.

Appleton-Century-Crofts.



Thorndike, E. (1911).

Animal Intelligence: Experimental Studies.

The animal behaviour series. Macmillan.

Reinforcement Learning



Alessandro Lazaric

alessandro.lazaric@inria.fr

sequel.lille.inria.fr