Master MVA: Reinforcement Learning **Lecture: 1**

Introduction to Reinforcement Learning

Lecturer: *Alessandro Lazaric* *http://researchers.lille.inria.fr/∼lazaric/Webpage/Teaching.html*

Plan of the Lecture

1. Brief historical overview of reinforcement learning
2. A multi-disciplinary field
3. The agent-environment model

1 Reinforcement Learning: from Psychology to Machine Learning

1.1 Animal Learning, Experimental Psychology, and Neuroscience

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].
- *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].

- *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).

1.2 Optimal Control Theory, Dynamic Programming, and Machine Learning

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

- *Objective*: **learning** of a behavior strategy (a *policy*) which maximizes the long term sum of rewards (**delayed reward**) by a direct interaction (**trial-and-error**) with an unknown and uncertain (e.g., stochastic) environment.
- *Examples*: sensorimotor learning, autonomous robotics, portfolio management, games (e.g., backgammon, chess, go), web advertising.

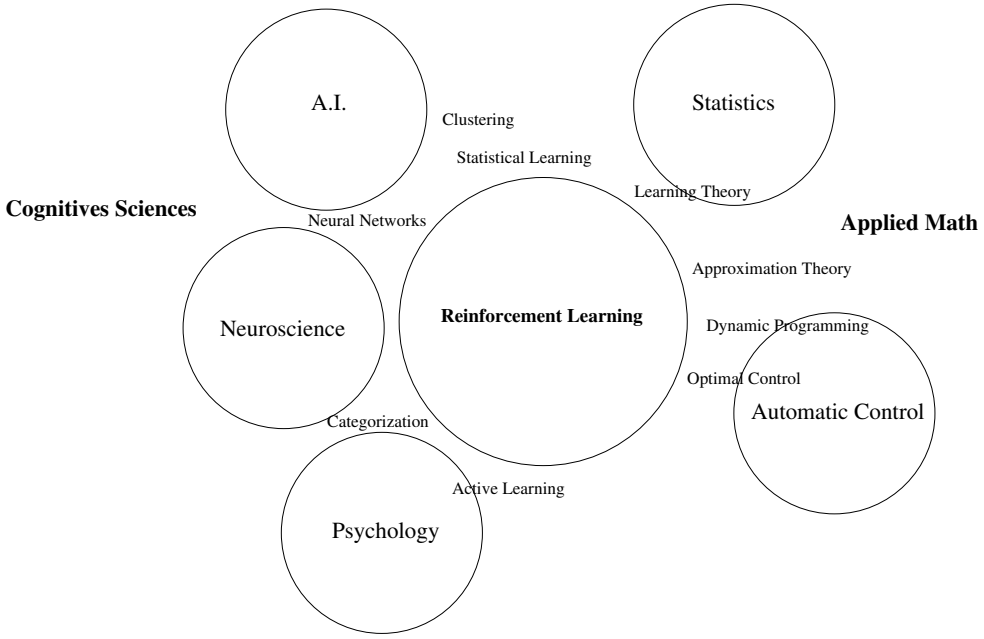
Reinforcement learning milestones

- "Programming a computer for playing chess" [Shannon, 1988].
- "Theory of Neural-Analog Reinforcement Systems" [Minsky, 1954].
- "Studies in machine learning using the game of checkers" [Samuel, 1959].
- "Trial and error" (tic-tac-toe player) [Michie, 1961].
- "BOXES: An experiment in adaptive control" (inverted pendulum) [Michie and Chambers, 1968].
- "Punish/Reward: Learning with a Critic in Adaptive Threshold Systems" (neural network) [Widrow et al., 1973].
- "Associative search network: A reinforcement learning associative memory" [Barto et al., 1981].
- "Temporal credit assignment in reinforcement learning" (temporal difference learning) [Sutton, 1984].
- "Learning from delayed rewards" (Q-learning) [Watkins, 1989].
- "Temporal difference learning and TD-Gammon" [Tesauro, 1995].

A machine learning paradigm

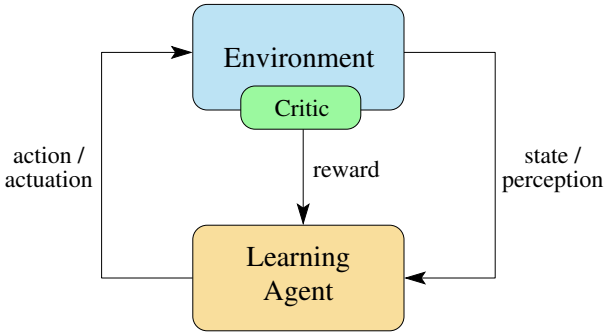
- *Supervised learning:* an expert (*supervisor*) provides examples of the right strategy (e.g., classification of clinical images). **Supervision is expensive.**
- *Unsupervised learning:* different objects are clustered together by similarity (e.g., clustering of images on the basis of their content). **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.**

A multi-disciplinary field



2 The Reinforcement Learning Model

2.1 The Agent-Environment Interaction Protocol



for $t = 1, \ldots, n$ **do**
  The agent perceives state s_t
  The agent performs action a_t
  The environment evolves to s_{t+1}
  The agent receives reward r_t
end for

The environment: depending on how it reacts to the agent's actions and how it can be perceived by the agent, we have

- Fully controllable (e.g., chess), partially controllable (e.g., portfolio optimization)
- Deterministic (e.g., chess) or stochastic (e.g., backgammon)
- Adversarial (e.g., chess) or fixed (e.g., tetris)
- Fully observable (e.g., chess) or partially observable (e.g., robotics)
- Known (e.g., chess) or unknown (e.g., robotics)

The critic: it returns the **reinforcement**, which evaluates the quality of the action taken by the agent depending on its actual state.

The agent: defines the concept of *state* and *action* and it acts according to

- Open loop control
- Close loop control (i.e., *adaptive*)
- Non-stationary close loop control (i.e., *learning*)

2.2 The Learning Problem

Objective: find the close loop strategy which maximizes the long term (average) sum of rewards.

Agent-environment model: How do we formalize the rules defining the dynamics of the environment and its interaction with the agent? ⇒ Markov decision processes ***[Lecture 2]***.

The credit assignment problem: How do we assign rewards to actions? How do we learn to give up short term rewards to get long term rewards? ⇒ Value functions ***[Lecture 2]***.

The representation problem: How do we represent complex solutions? can we somehow approximate them? ⇒ Function approximation ***[Lecture 5/6]***.

Information collection: What is the most effective way to collect information about the environment and the rewards? ⇒ Exploration/exploitation dilemma ***[Lecture 4]***.

The learning process: How do we learn the optimal strategy?

- Known model ⇒ Dynamic programming ***[Lecture 2]***.
- Unknown model incremental ⇒ Temporal difference algorithms ***[Lecture 3]***.
- Unknown model batch ⇒ Approximate dynamic programming ***[Lecture 5]***.

3 Successful stories in Reinforcement Learning

- TD-Gammon [Tesauro, 1995]: best backgammon player.
- KnightCap [Baxter et al., 1998]: chess player with $\simeq$2500 ELO.

- Robotics [Schaal and Atkeson, 1994]: jugglers, balancers, acrobots.
- Mobile robotics [Thrun et al., 1999]: robot guide at theSmithonian museum.
- Elevators control system [Crites and Barto, 1996].
- Package routing [Littman and Boyan, 1993].
- Job-shop scheduling [Zhang and Dietterich, 1995].
- Production manufacturing control [Mahadevan and Theocharous, 1998]
- Computer poker[University of Alberta, 2006]: first computer poker player to beat professional players.
- Computer go: computer players constantly beat amateurs and approaching master players.

References

[Barto et al., 1981] Barto, A., Sutton, R., and Brouwer, P. (1981). Associative search network: A reinforcement learning associative memory. *Biological Cybernetics*, 40(3):201–211.

[Baxter et al., 1998] Baxter, J., Tridgell, A., and Weaver, L. (1998). Knightcap: A chess program that learns by combining td() with game-tree search. In *Proceedings of the 15th International Conference on Machine Learning*, pages 28–36. Morgan Kaufmann.

[Bellman, 2003] Bellman, R. (2003). *Dynamic Programming*. Dover Books on Computer Science Series. Dover Publications, Incorporated.

[Crites and Barto, 1996] Crites, R. and Barto, A. (1996). Improving elevator performance using reinforcement learning. In *Advances in Neural Information Processing Systems 8*, pages 1017–1023. MIT Press.

[Damasio, 1994] Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. Grosset/Putnam.

[Doya, 1999] Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex. *Neural Networks*, 12:961–974.

[Hebb, 1961] Hebb, D. O. (1961). Distinctive features of learning in the higher animal. In Delafresnaye, J. F., editor, *Brain Mechanisms and Learning*. Oxford University Press.

[Littman and Boyan, 1993] Littman, M. and Boyan, J. (1993). A distributed reinforcement learning scheme for network routing. In *In Proceedings of the 1993 International Workshop on Applications of Neural Networks to Telecommunications*, pages 45–51. Erlbaum.

[Mahadevan and Theocharous, 1998] Mahadevan, S. and Theocharous, G. (1998). Optimizing production manufacturing using reinforcement learning. In *In Eleventh International FLAIRS Conference*, pages 372–377. AAAI Press.

[Michie, 1961] Michie, D. (1961). Trial and error. *Science Survey*, 2:129–145.

[Michie and Chambers, 1968] Michie, D. and Chambers, R. A. (1968). Boxes: An experiment in adaptive control. *Machine Intelligence 2*.

[Minsky, 1954] Minsky, M. (1954). *Theory of Neural-analog Reinforcement Systems and Its Application to the Brain Model Problem*. Princeton University.

[Pavlov, 1927] Pavlov, I. (1927). *Conditioned reflexes*. Oxford University Press.

[Pontryagin and Neustadt, 1962] 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.

[Samuel, 1959] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229.

[Schaal and Atkeson, 1994] Schaal, S. and Atkeson, C. (1994). Robot juggling: implementation of memory-based learning. *Control Systems, IEEE*, 14(1):57–71.

[Shannon, 1988] Shannon, C. E. (1988). Computer chess compendium. chapter Programming a computer for playing chess, pages 2–13. Springer-Verlag New York, Inc., New York, NY, USA.

[Skinner, 1938] Skinner, B. F. (1938). *The behavior of organisms*. Appleton-Century-Crofts.

[Sutton, 1984] Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning*. PhD thesis. AAI8410337.

[Tesauro, 1995] Tesauro, G. (1995). Temporal difference learning and td-gammon. *Commun. ACM*, 38(3):58–68.

[Thorndike, 1911] Thorndike, E. (1911). *Animal Intelligence: Experimental Studies*. The animal behaviour series. Macmillan.

[Thrun et al., 1999] Thrun, S., Bennewitz, M., Burgard, W., Cremers, A., Dellaert, F., Fox, D., Hähnel, D., Rosenberg, C., Roy, N., Schulte, J., and Schulz, D. (1999). MINERVA: A second generation mobile tour-guide robot. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

[University of Alberta, 2006] University of Alberta (2006). Computer poker player.

[Watkins, 1989] Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. PhD thesis, King's College, Cambridge.

[Widrow et al., 1973] Widrow, B., Gupta, N. K., and Maitra, S. (1973). Punish/reward: Learning with a critic in adaptive threshold systems. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-3(5):455–465.

[Zhang and Dietterich, 1995] Zhang, W. and Dietterich, T. G. (1995). A reinforcement learning approach to job-shop scheduling. In *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1114–1120. Morgan Kaufmann.