Reinforcement Learning

Abir Das
Assistant Professor
Computer Science and Engineering Department
Indian Institute of Technology Kharagpur

http://cse.iitkgp.ac.in/~adas/

Logistics

- **Course Name and code**: Reinforcement Learning, CS60077
- **Time**: Thursday (3:00-4:55 pm), Friday(3:00-3:55 pm) [Slot V3]
- **Office Hours**: Saturday(3:00-4:00 pm) [Keep it free for in-class quizzes. It will be announced earlier]
- **Venue**: Your rooms!!
- **Course website**: http://cse.iitkgp.ac.in/~adas/courses/rl_aut2021/rl_aut2021.php
- **TAs**: Aadarsh Sahoo (sahoo_aadarsh@iitkgp.ac.in), Siddhant Agarwal (agarwalsiddhant10@iitkgp.ac.in)

Logistics

- **Microsoft team**: https://tinyurl.com/yy3hmt53 (Imp: Download the app)
- **Piazza Forum**: https://piazza.com/iitkgp.ac.in/fall2021/cs60077/home
- **Google Cloud Platform (GCP)**: Google has kindly agreed to provide limited free cloud credit for the compute needed for this course. Details will be posted in Piazza soon.
- **Moodle**: For quizzes and assignments, moodle will be used. The moodle link is: https://moodlecse.iitkgp.ac.in/moodle/login/index.php

Course Information

- **Prerequisites**: 1. CS60010: Deep Learning.

- **Python Proficiency**: Proficiency in Python. Familiarity with some Deep Learning tools (Tensorflow, Pytorch etc.) could be handy. A few links to get started.
  - https://docs.python.org/3/tutorial/
  - http://cs231n.github.io/python-numpy-tutorial/

Course Information

- **Books and References**:
  1. "Reinforcement Learning: An Introduction", Richard S. Sutton and Andrew G. Barto, 2nd Edition, Free link.
  2. "An Introduction to Deep Reinforcement Learning", Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare and Joelle Pineau, Free link.
  3. "Algorithms for Reinforcement Learning", Csaba Szepesvari, Free link.
  4. [For Probability Primer] - "Probability, Statistics, and Random Processes for Electrical Engineering", 3rd Edition, Alberto Leon-Garcia, Free link.
  5. [For Probability Primer] - Lecture Notes on Introduction to Probability, Dimitri P. Bertsekas and John N. Tsitsiklis, Free link
- More references specific to the lectures will be added in the course

Course Information

- Online lecture Videos: The following courses will be closely followed in this course
  - Reinforcement Learning by David Silver (Link)
  - Deep Reinforcement Learning by Sergey Levine (Link)
  - NPTEL Reinforcement Learning by Balaraman Ravindran (Link)

Course Information

- **Evaluation**: Homework (40%) ; In-class quizes (40%); Project (20%).
  - Project
    - Each project will be done by a 3 member team. Start forming the team.
    - Tentative deadline to submit project title and half a page abstract along with the team member names is [Sept 30, 2020].
    - Coming up with your own project idea is highly recommended. You can discuss with the TAs and mail me if you need to discuss.
    - The project deliverable is a 6 page report plus bibliography [CVPR Style paper with reduced page limit] at the end of the course (Tentatively Pre final week) and a team presentation.

Reinforcement Learning

- "**Goal-directed** learning from **interaction** in an **uncertain** environment" – Reinforcement Learning: An Introduction, Sutton and Barto, Second Edition

Video courtesy: Brendan Cole, youtube link

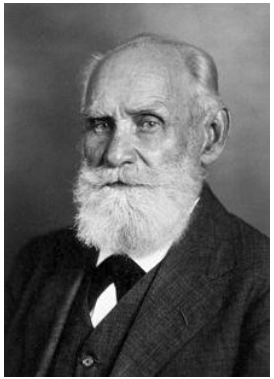
Reinforcement Learning

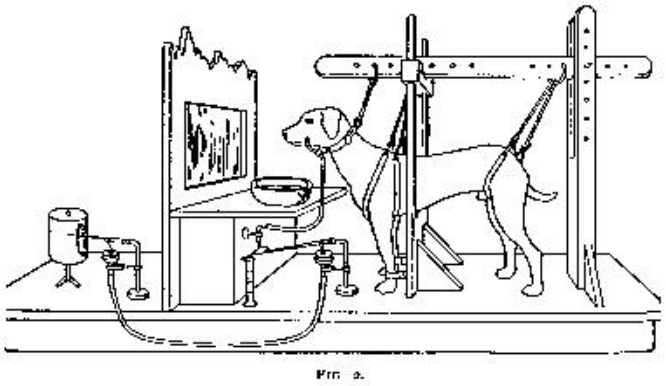
- Definition of Learning from Psychology: Learning is a relatively **permanent** change in behavior that occurs through **experience**. It's a **continuous** and **gradual** process.

Image credit: Melbourne Child Psychology & School Psychology Services

Some more of Psychology

- **Classical Conditioning**: Theory developed by Ivan Pavlov (Nobel prize in 1904).



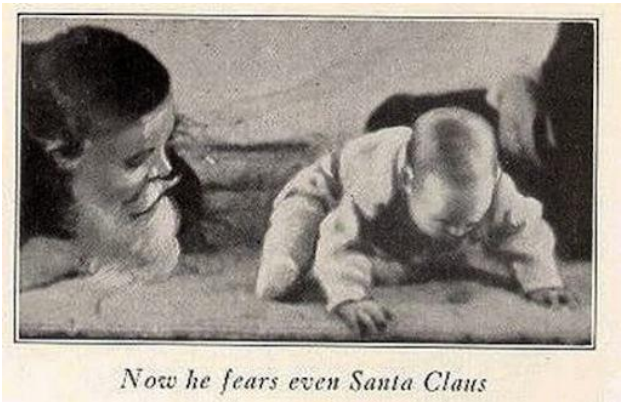


- Dog salivates seeing the food (Unconditioned response (salivating) to Unconditioned Stimulus (food))
- Dog salivates hearing the bell (Conditioned response (salivating) to Conditioned Stimulus (bell)). Learning to associate bell to salivation via food.

Images credit: Wikipedia

Some more of Psychology

- **Classical Conditioning**: In an experiment with human babies, it was learnt that classical conditioning can occur in humans too.
  - John B. Watson in 1920 performed his famous "Little Albert experiment"



Now he fears even Santa Claus

- Mary Cover Jones in 1924 performed her "Little Peter experiment" to show that "counter conditioning" is possible for human subjects.

Image credit: Wikipedia

Some more of Psychology

- **In Classical Conditioning**: The reward (e.g., food) is present and that caused the response (e.g., salivation).
- At a similar time, Edward Thorndike was conducting experiments on cats to see how positive/negative feedback affects goal directed learning.

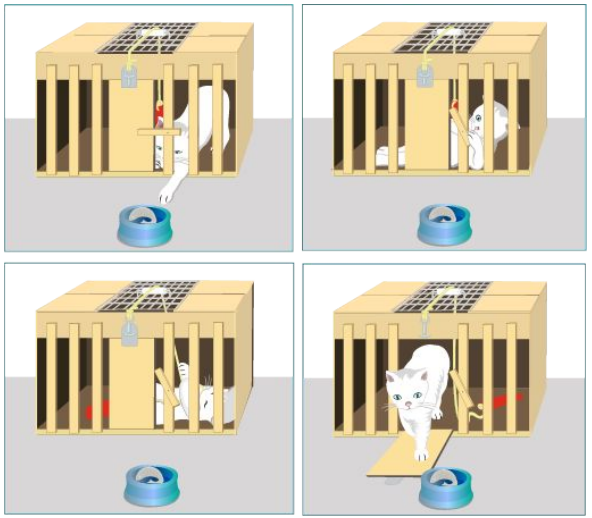


Image Link

Some more of Psychology

- **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."
- This is later studied more by B. F. Skinner and he popularized the term **Operant Conditioning**.
- Pigeons were provided food grains as reward for some "appropriate" behavior.
- Skinner also came up with "Pigeon guided missiles" using similar principle.

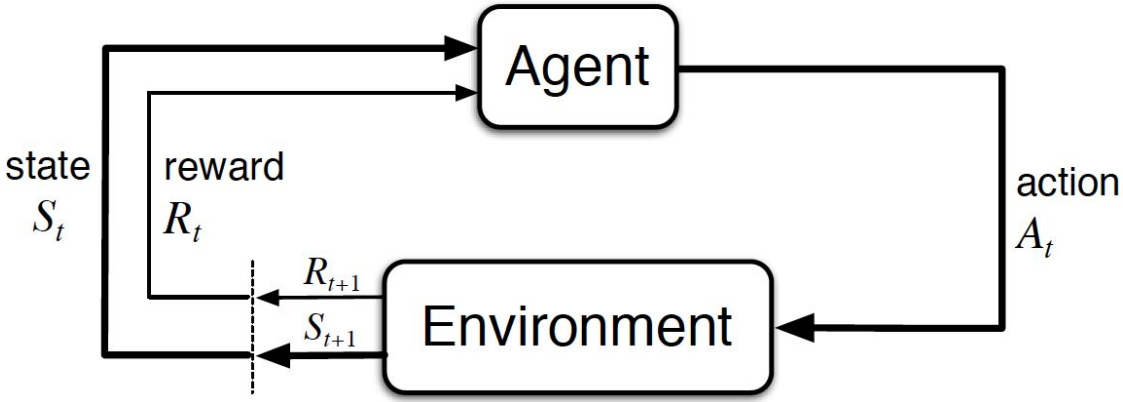
Some more of Psychology

- **Reinforcement/Punishment:-** In Psychological terms, **Reinforcement** is a consequence that causes a behavior to occur with greater frequency. **Punishment** is a response or consequence that causes a behavior to occur with less frequency.

- Reinforcement can also mean removal of aversive consequence (Negative Reinforcement).

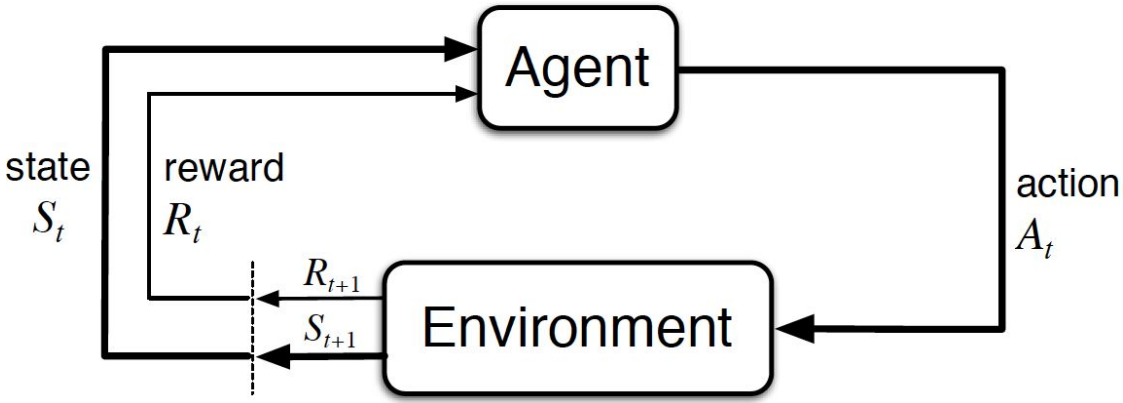
- Similarly, Punishment can also mean removal of rewarding consequence (Negative Punishment).

Reinforcement Learning Setting



- **Agent**:- The learner or the decision maker.
- **Environment**:- The thing the learner interacts with, comprising everything outside the agent.
- They interact continually. The agent selects actions. The environment responds to these actions by presenting new situation and giving rewards for the action

Reinforcement Learning Setting

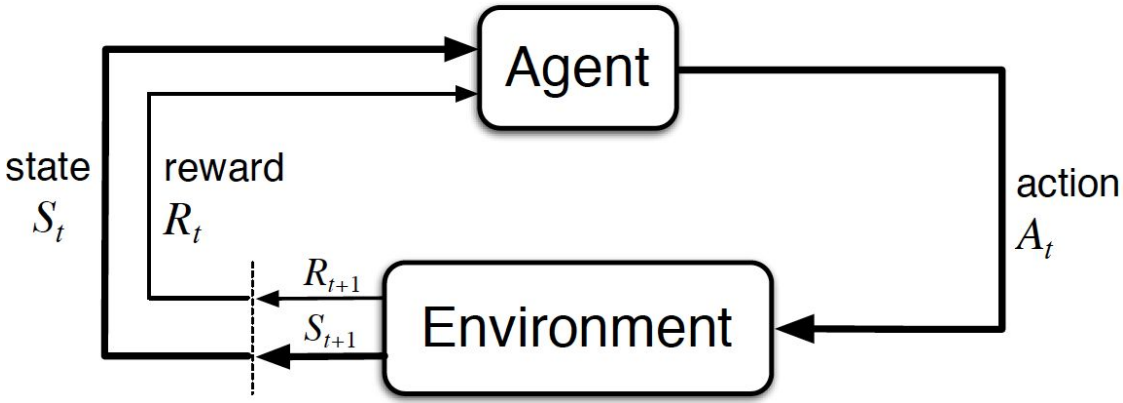


- **Reward**:- Rewards are scalar measures defining what are the good and bad events for the agent.
- **Value**:- Value of a state is the total amount of reward an agent is expected to get in future starting from that state.

To make a human analogy, rewards are somewhat like pleasure (if high) and pain (if low), whereas values correspond to a more refined and farsighted judgment of how pleased or displeased we are that our environment is in a particular state.

Sutton and Barto

Reinforcement Learning Setting



- **Goal in RL Problem**:- to maximize the total reward "in expectation" over the long run.
- $\tau \stackrel{\text{def}}{=} (s_1, a_1, s_2, a_2, \dots), p(\tau) = p(s_1) \prod_t p(a_t|s_t) p(s_{t+1}|s_t, a_t)$
- $\max \mathbb{E}_{\tau \sim p(\tau)} [\sum_t R(s_t, a_t)]$

Distinguishing Features of Reinforcement Learning

- **Trial and error:-** The learner is not told which actions to take, but instead must discover which actions are most rewarding by trying them.

- **Delayed rewards:-** actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards.

- **Exploration-vs-exploitation:-**

To obtain a lot of reward, a reinforcement learning agent must prefer actions that it has tried in the past and found to be effective in producing reward. But to discover such actions, it has to try actions that it has not selected before. - *Sutton and Barto*

When to use Reinforcement Learning

- **Trajectories**:- Data comes in the form trajectories.
- **Decisions**:- Need to take decisions that affect the trajectory data.
- **Feedback**:- Need to get feedback about the choice of actions.

Supervised, Unsupervised and Reinforcement Learning

- **Supervised Learning**:- Learn $y = f(x)$ – You are given a bunch of (x, y) pairs and your goal is to find the function f mapping x to y.

- **Unsupervised Learning**:- Learn $f(x)$ – You don't have access to any y, you are given a bunch of $x's$ and your goal is to find some f that gives a "compact description" of these $x's$.

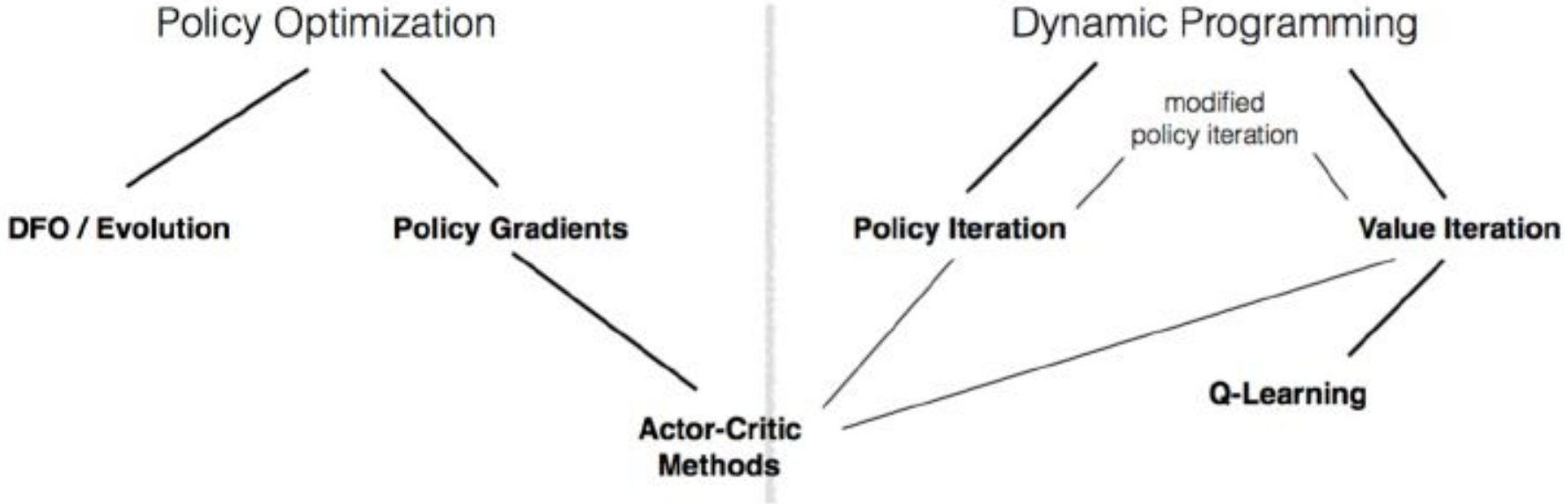
- **Reinforcement Learning**:- Learn $y = f(x)$, given z – You are given a bunch of (x, z) pairs and your goal is to find the function f mapping x to y.
  - x is state, y is action, z is reward

Supervised, Unsupervised and Reinforcement Learning

- **Supervised Learning**:- Does not involve the problem of temporal credit assignment and exploration.

- **Unsupervised Learning**:- Here, in addition, the correct labels are not available

RL Algorithm Landscape



Slide credit: Deep RL Bootcamp, 2017

Tentative Plans for the Course

- Dynamic Programming based Methods
- Policy Optimization
- Actor-Critic Methods
- Advanced Policy Optimization

Concluding with a quote from 1968 paper

The learning program is thus in the same situation as pre-scientific man, with a fixed repertoire of acts (control signals), a growing repertoire of experiences (state signals) and a basic assumption that there exist some unknown "laws of nature" which enable predictions to be made and strategies for survival to be devised. Progress in such a situation proceeds via the construction of a succession of models (partial descriptions) of the environment, proceeding from the crudest by successive refinement. If a given model can be used for prediction we call it a theory.

Donald Michie, "Memo Functions and Machine Learning", *Nature, 196*

Thank You!!