CS 7.603

Reinforcement Learning

Tejas Bodas

Assistant Professor, IIIT Hyderabad

Logistics

- Second half of the course by Prof. Harikumar K.
- Feel free to contact me anytime at tejas.bodas@iiit.ac.in.
- Office @ A5304b.
- TA list: Sukhjinder, Shikhar & Mukund
- Communication: Moodle

Resources

- Wont be following any one particular book.
- Lecture slides will have material from variety of sources.
- Some popular books:
  1. Reinforcement learning by Sutton & Barto
  2. Reinforcement learning and Optimal control by Bertsekas
  3. Applied probability models with optimization applications by Sheldon Ross (for MDP's)
  4. Other recent books by Warren Powell, Sean Meyn, Sham Kakde, Abhijit Gosavi, Ashwin Rao.
- Some Course notes
  1. `https://appliedprobability.files.wordpress.com/2021/01/stochastic_control_jan29.pdf`
  2. `https://adityam.github.io/stochastic-control/notes/`
  3. `https://www.deepmind.com/learning-resources/introduction-to-reinforcement-learning-with-david-silver`

Evaluation scheme

- Quiz 1 : 10%.
- Midsem exam: 25%.
- Project 1 25%
- Quiz 2: 10%
- Project 2 30%

Course Outline

- Module 1 (3-4 Lectures)
  Motivation & Probability & Markov Chains
- Module 2 (5-6 Lectures)
  Markov Decision Processes
- Module 3 (4-5 Lectures)
  Introduction to Reinforcement Learning
- Module 4 & 5 (12-14 lectures)
  Advanced Reinforcement learning (Prof. Harikumar)

Homework for today!

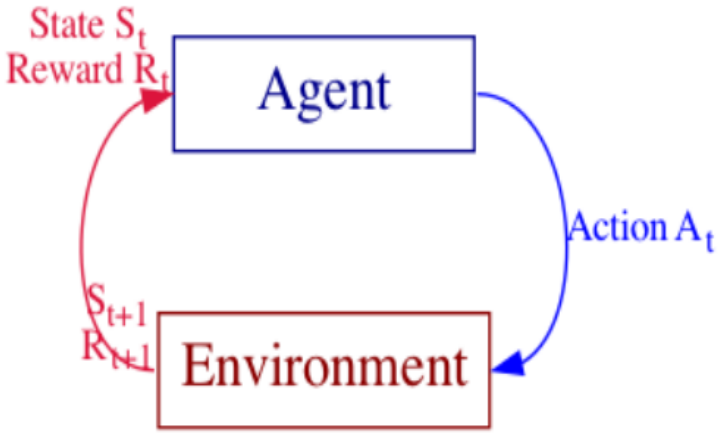
Watch the following

- AlphaGo (2017 documentary, 90min)
- David Silver Lecture 1.

What is RL? Mathematical Viewpoint

- It is essentially an MDP where the Markovian transitions are unknown.
- What is an MDP ? It is essentially a Markov Chain that you control with actions for maximising your accumulated reward.
- What is a Markov Chain ? Basically a collection of dependent random variables (stochastic process).
- Given $X_{present}$ and X_{past}, X_{future} depends only on $X_{present}$ and is independent of X_{past}.
- To predict the future evolution, I only need to know the present state and your past experience is irrelevant.
- What is a random variable ? MA6.101.

Viewpoint 2: Sequential decision problem

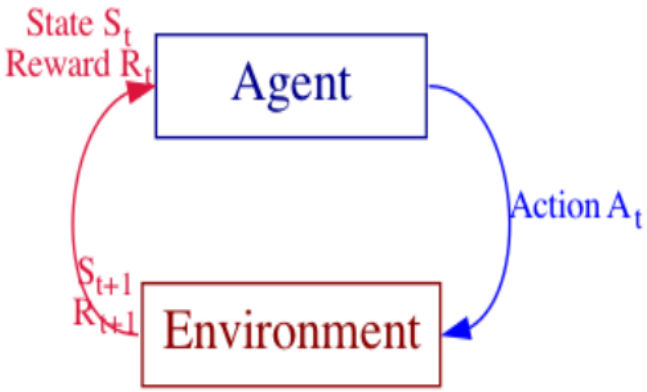


- RL studies how an agent learns to perform tasks by trial and error while interacting with an environment.
- Notion of State, Action, Reward, next State (SARSA)
- Agent has to select sequence of actions to maximize total reward under environment uncertainty.
- Balance immediate gains and long term gains.
- Balance exploration and exploitation.

Key ingredients of MDP/RL

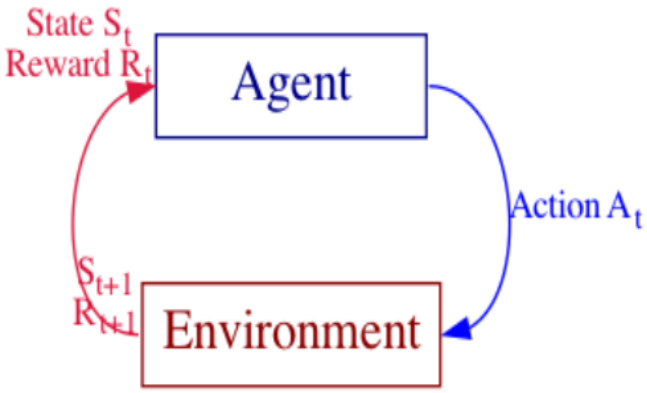
- Model for the environment
  1. Transition model
  2. Reward model
- Policy of the agent
- Value function for the policy and/or states

Model, Policy and Value function



- **Transition Model:** represents dynamics of the environment.
- $S_{t+1} = f(\mathcal{H}_t, W_t)$ represents the model for state transitions where history $\mathcal{H}_t = \{S_1, A_1, \ldots, S_t, A_t\}$ and W_t represents possible source for randomness.
- Markovian Model: $S_{t+1} = f(S_t, A_t, W_t)$ where W_t is i.i.d noise. In this case, the following Markov property is true
- $P[S_{t+1} = s' | S_t = s, A_t = a, S_{t-1}, A_{t-1}, \ldots] =$
  $P[S_{t+1} = s' | S_t = s, A_t = a]$

Model, Policy and Value function



- **Reward Model:** R_{t+1}represents the reward received at time t.
- Typically we assume that $R_{t+1} = g(S_t, A_t)$, i.e, the reward depends on current state and action.
- Other models for reward include $R_{t+1} = g(S_t, A_t, S_{t+1})$
- **Reward Hypothesis:** Goal in RL is to maximize the expected total rewards under *model uncertainty*.
- Other related criteria include, finite time expected total reward, time average reward and discounted total expected reward.

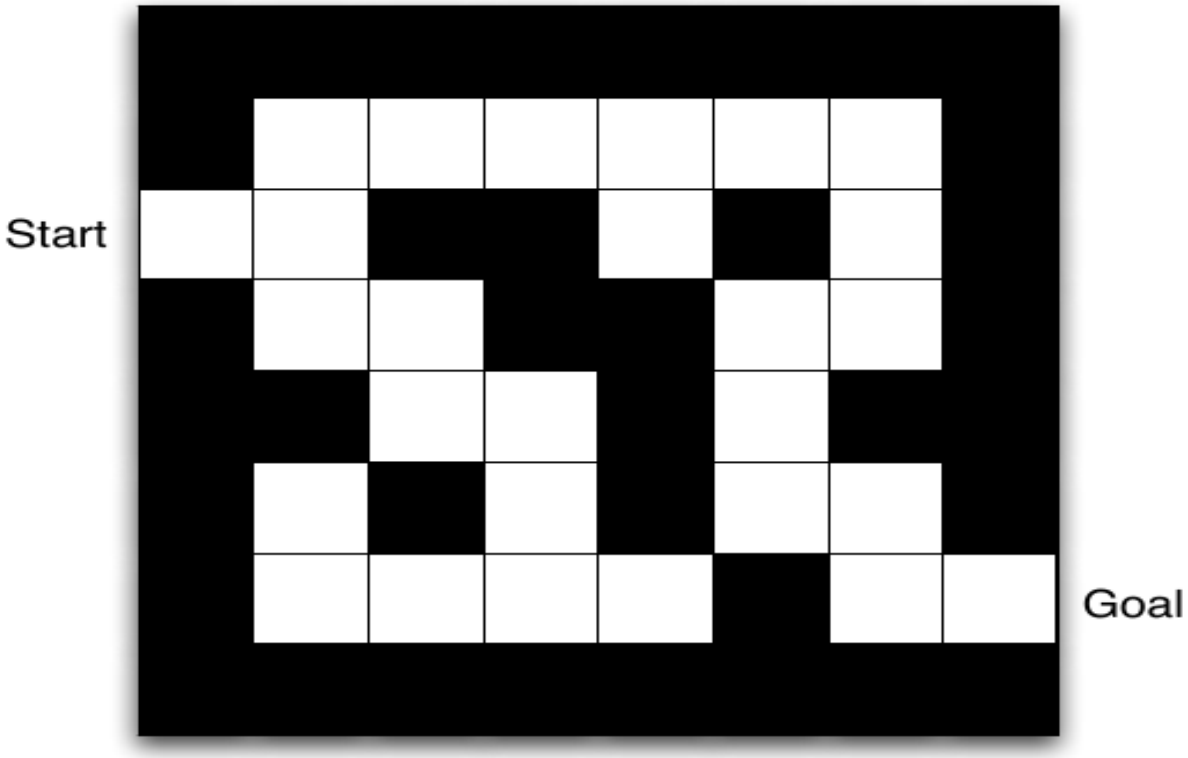
Model, Policy and Value function

- **Policy π:** A policy π represents a strategy for choosing action at each time.
- Policies/strategies could be history based, Markovian, deterministic, randomized, stationary etc.
- An optimal policy π^* is one which offers the highest expected total reward.
- When the model is known (in case of MDP's), the *optimal policies* often turn out to be Markovian, deterministic and even stationary (more later).
- However in RL, the model is unknown and you therefore do not know the optimal policy.
- RL is all about figuring out the optimal policy without incurring much *regret*.

Model, Policy and Value function

- When following a policy π, we may want to know the value or quality of states being visited.
- The **value function** $V^{\pi}(s)$ quantifies the expected total reward from policy π when starting in state s.
- Another important quantity is the state action value function $Q^{\pi}(s, a)$ for policy π.
- Main objective in MDP is to obtain expressions for the optimal value function $V(s) := \max_{\pi \in \Pi} V^{\pi}(s)$ and $\pi^* = \arg\max_{\pi \in \Pi} V^{\pi}(s)$
- Under model uncertainty, the objective in RL is to quickly learn $Q(s, a)$ and π^*

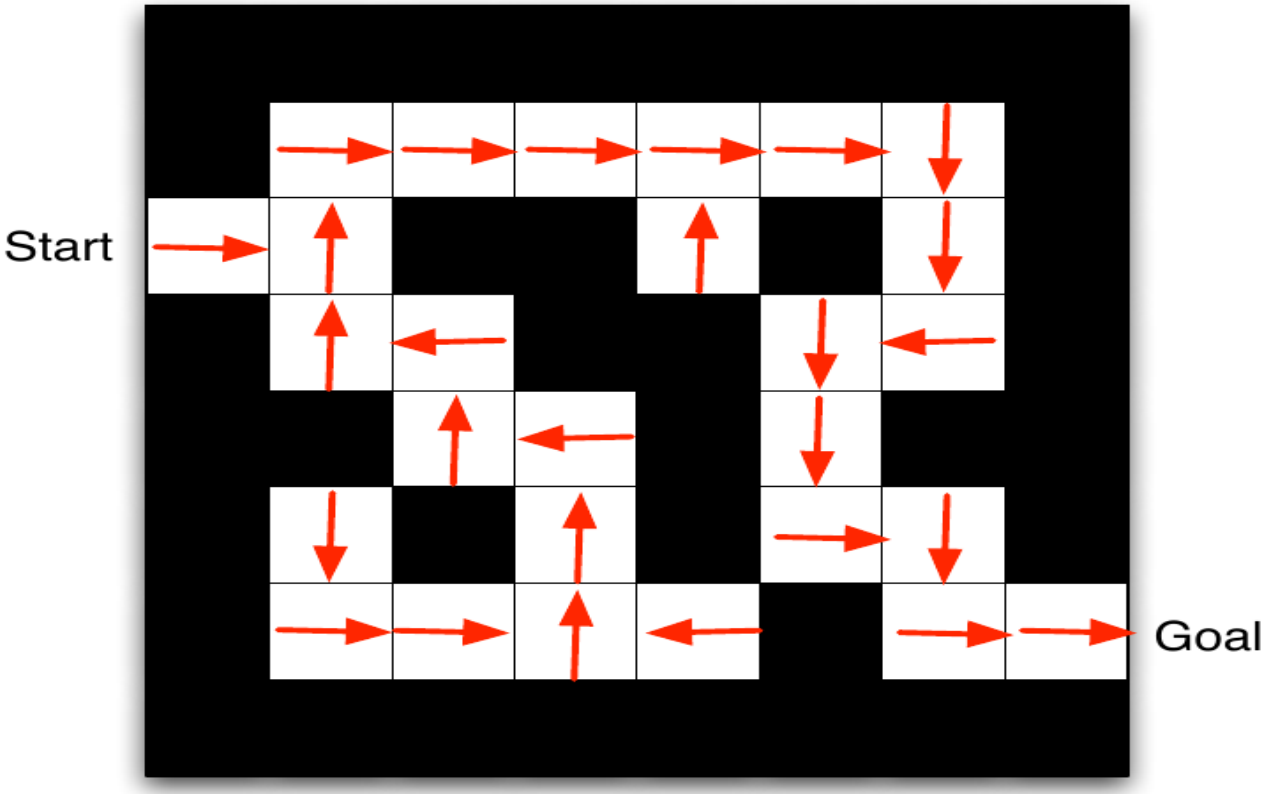
Maze Runner [1]



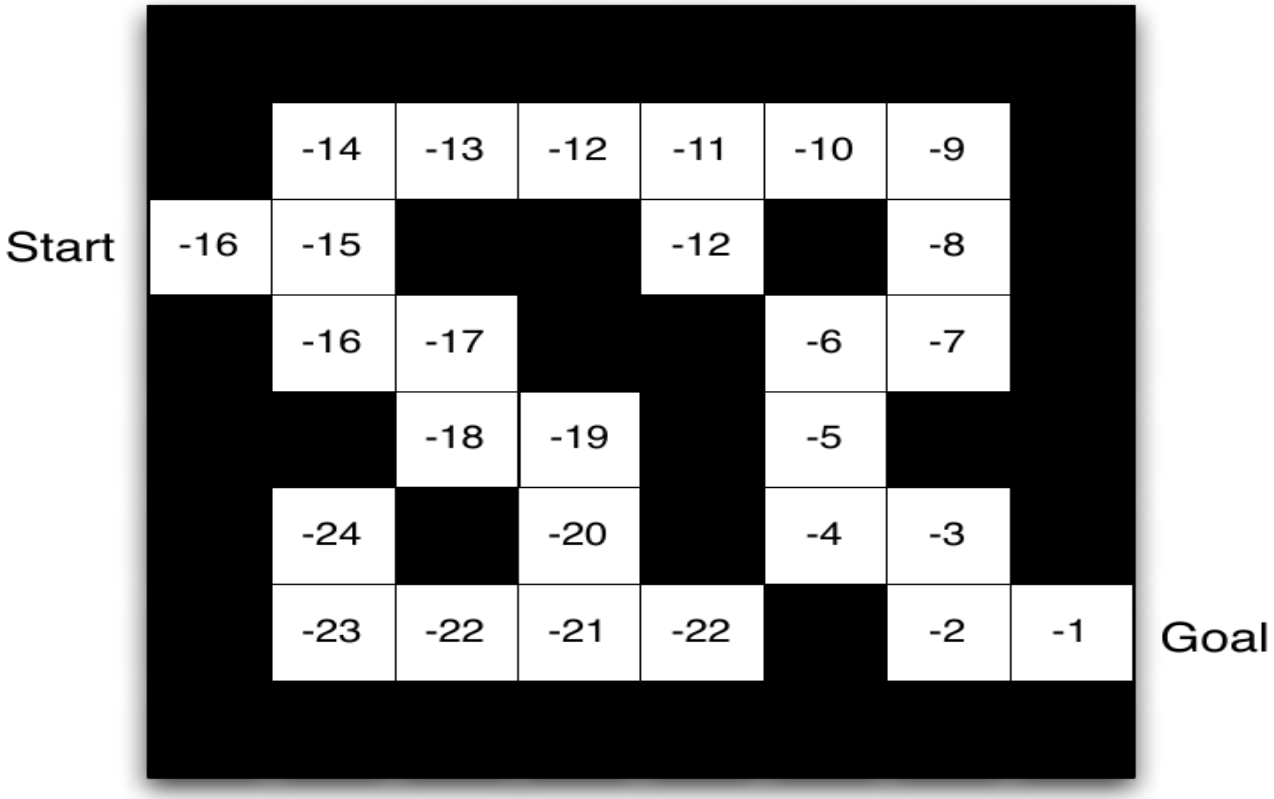
- States are locations in the maze
- Rewards are -1 per time step
- Actions are L,R,U,D.

[1] Example from David Silver's slides

Maze Runner – Optimal policy π^*



Maze Runner – Optimal value $V(s)$



- Optimal value $V(s)$ captures the expected total discounted reward when starting in state s and following the optimal policy.

Classification of RL problems

- Under model uncertainty, the objective in RL is to quickly learn $Q^*(s, a)$ and/or π^*.
- RL algorithms that focus on learning $Q^*(s, a)$ quickly are called value function based algorithms, e.g. Value iteration.
- RL algorithms that focus on learning π^* quickly are called policy based algorithms, e.g. policy iteration.
- Some RL algorithms do both, e.g. actor-critic algorithms.
- Note of these algorithms try to learn the model $f(.,.)$ explicitly and hence are called model free algorithms.
- Some algorithms try to explicitly learn the model first and then solve MDP for the learnt model. Such algos are called Model based.

Where is MDP/RL used?

- Robotics.
- Autonomous/Self driving vehicles.
- Finance (Management of Investment Portfolio)
- Inventory control in Operations Research
- Dynamic pricing in Operations Management.

Agenda for next 2 lectures

- Basic Probability
- Random variables
- Sequence of Random variables and some Limit theorems
- Markov Chains
- Markov Reward Process