Programming Assignment 2

EE932: Introduction to Reinforcement Learning

May 10, 2024

Instructions

- **Assignment Deadline - 22nd May 2024**
- Kindly name your submission files as '`RollNo_Name_A1.ipynb`'.
- You are required to work out your answers and submit only the iPython Notebook. The code should be well commented and easy to understand as there are marks for this.
- You may use the notebook given along with the assignment as a template. You are free to use parts of the given base code but may also choose to write the whole thing on your own.
- Submissions are to be made through iPearl portal. Submissions made through mail will not be graded.
- Answers to the theory questions, if any, should be included in the notebook itself. While using special symbols use the *LaTeX* mode
- Make sure your plots are clear and have title, legends and clear lines, etc.
- Plagiarism of any form will not be tolerated. If your solutions are found to match with other students or from other uncited sources, there will be heavy penalties and the incident will be reported to the disciplinary authorities.
- In case you have any doubts, feel free to reach out to TAs for help.

Dynamic Programming for MDP

Part A: Deterministic Career Path

Consider a simple Markov Decision Process below with four states and two actions available at each state. In this simplistic setting actions have deterministic effects, i.e., taking an action in a state always leads to one next state with transition probability equal to one. There are two actions out of each state for the agent to choose from: D for development and R for research. The _ultimately-care-only-about-money_ reward scheme is given along with the states.

The jupyter-notebook accompanied with the assignment implements the environment and demonstrates and tests the Policy Iteration (PI) method to obtain an optimal policy.

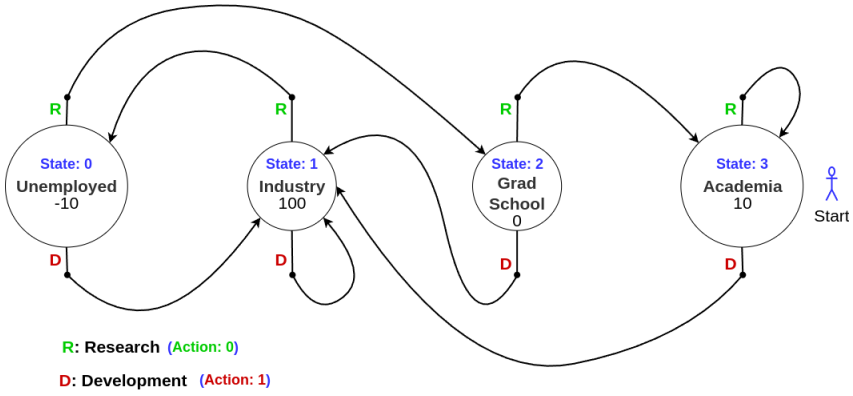

Figure 1: A deterministic career path MDP

Go thorough the sample code for the above example here: https://tinyurl.com/ee932-assignments and answer the following questions.

A1 Value Iteration

A1.1 Find an optimal policy to navigate the given environment using Value Iteration (VI) [8 Marks]

A1.2 Compare PI and VI in terms of convergence. Is the policy obtained by both same? [2 Marks]

Part B: Stochastic Career Path

Now consider a more realistic Markov Decision Process below with four states and two actions available at each state. In this setting Actions have nondeterministic effects, i.e., taking an action in a state always leads to one next state, but which state is the one next state is determined by transition probabilities. These transition probabilites are shown in the figure attached to the transition arrows from states and actions to states. There are two actions out of each state for the agent to choose from: D for development and R for research. The same _ultimately-care-only-about-money_ reward scheme is given along with the states.

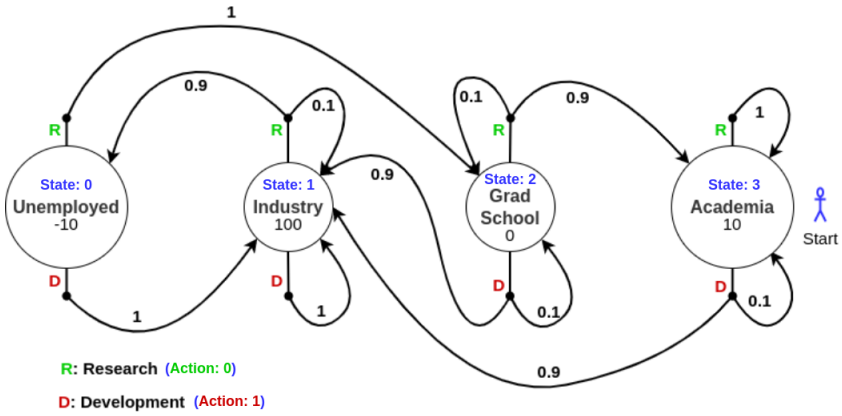

Figure 2: A stochastic career path MDP

B1 Determining the optimal policy (Bonus Questions)

B1.1 Find an optimal policy to navigate the given environment using Policy Iteration (PI) [4 Marks]

B1.2 Find an optimal policy to navigate the given environment using Value Iteration (VI) [4 Marks]

B1.3 Compare PI and VI in terms of convergence. Is the policy obtained by both same? [2 Marks]

[Hint] Notice what would change in the code due to the stochastic nature of the environment?