COMP0037 2021 / 2022 Robotic Systems
Lab Week 05: Value Functions, Policies and Policy Iteration

COMP0037 Teaching Team

February 10, 2022

Overview

In this lab, we will explore some of the ideas associated with FMDPs, value functions and policy evaluation. This lab covers Lectures 05–08.

Installation Instructions

This code uses NumPy, which should be installed in your system by default.

FMDP Model

The problem of a cleaning robot is modelled as an episodic Markov Decision Problem (MDP):

1. The state is the battery level of the robot. It can be in one of four states: `HIGH`, `MEDIUM`, `LOW`, `DISCHARGED`. If the robot reaches the `DISCHARGED`, the robot can only transition to the virtual terminal state and the episode ends.

2. The action space consists of: `TERMINATE`, `SEARCH`, `WAIT` and `RECHARGE`. The `TERMINATE` action can only be called from the `DISCHARGED` state.

3. The state transition probabilities are described below.

4. The reward function is only dependent upon the action.

5. The discount factor $\gamma = 1$.

At the start of each episode, the robot's initial state is `HIGH`. The episode terminates when the robot reaches the `DISCHARGED` state.

The battery state transition depends upon the activity the robot undertakes:

- **Search:** If the battery level is high, the probability that it will remain high is $1 - \alpha$, the probability that it becomes medium is $\frac{2\alpha}{3}$ and the probability that it becomes low is $\frac{\alpha}{3}$. If the battery level is medium, there is a probability $1 - \beta$ that it remains at medium, and a probability β that it becomes low. Finally, if the robot battery level is low, there is a probability $1-\mu$ that it remains in a low state and a probability μ that the battery becomes flat. If the battery becomes flat, the robot has to be manually taken to a charging station, and the episode ends.

- **Wait:** The battery state remains the same with probability 1.

- **Recharge:** If the battery state is high, it remains high with probability 1. If the battery state is medium, there is a probability $1 - \delta$ that it will remain at medium (does not charge), and a probability δ that it will become high. Finally, if the battery state is low, the battery will change its state from low to medium with probability δ and stay low with probability $1 - \delta$.

At each time step, the robot receives a reward signal. The reward signal depends upon the robot's current activity. When searching for rubbish, the reward is r_{search} per time step. While waiting for rubbish to accumulate, the reward is r_{wait}. The rewards for recharging is $r_{recharge}$. If the robot runs out of charge, a human has to push it to a charging station. The reward is $r_{discharged}$.

The parameters used in this lab are in Table 1.

FMDP Construction

1. Complete the implementation of the state transition and rewards for the system in class `RecyclingRobotEnvironment`. You will need to modify two things. First, you

Parameter	Value	Parameter	Value
α	0.4	β	0.1
γ	0.1	δ	0.9
r_{search}	10	r_{wait}	5
$r_{discharged}$	-20	$r_{recharge}$	0

Table 1: The parameter values.

will need to modify `_state_transition_table`. Second, you will need to fill in values in the `_action_reward_table`.

Policy Evaluation

2. Implement the policy evaluation class. Using the uniformly distributed policy, verify that your state value function is

$$v_\pi(s_) = \begin{bmatrix} -20 & 1226.65 & 1348.50 & 1345.38 \end{bmatrix}.$$

What do you think the large positive values mean?

Searching for Better Policies

3. Explore other choices of policies to see if you can identify a better one than taking actions at random.