Reinforcement Learning and Decision Making Under Uncertainty IN-STK 5100/9100

Spring 2022
A. George, UiO
22nd February, 2022

Mandatory Assignment 5

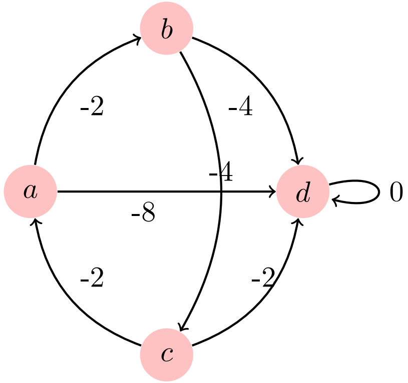
Submission Deadline: 8th March, 2022

Exercise 8. Modelling MDPs

Give two example tasks that fit into the MDP framework. Make the two examples as different from each other as possible. What are the states, actions, rewards, transitions and appropriate discount factor? Are these MDPs finite? Are the tasks episodic, finite horizon or infinite horizon tasks?

Exercise 9. Policy Iteration

Consider the MDP on the right, where states correspond to nodes and actions to edges in the graph, and where all transitions are deterministic, i.e., when taking action (a, b) in state a the next state is b and the reward is -2 with probability 1. (When taking an action (x, y) in state z with $x \neq z$ you can assume that the next state is z and the reward is $-\infty$.) Furthermore, let the discount factor be $\gamma = 1$.



a) Let π_0 be the policy that selects an outgoing edge (action) uniformly at random for every node (state). Compute the approximated state values for the policy π_0 using Policy Evaluation (2-array version) with $\theta = 0.35$ and initial values $V(s) = 0$ for all states s. Write down the state values of all states for every round of updates in a table.

b) Given π_0 and it's approximated values as computed in a), what would be the improved policy π_1 according to Policy Iteration?

Exercise 10. Policy Improvement

Show that the (deterministic) greedy policy chosen in the policy improvement step

$$\pi'(s) \in \arg\max_a q_\pi(s, a) \text{ for all } s \in \mathcal{S}$$

satisfies the condition of the policy improvement theorem:

$$q_\pi(s, \pi'(s)) \geq v_\pi(s) \text{ for all } s \in \mathcal{S}.$$

Argue why (the special case of) the Policy Improvement Theorem as given in the lecture is relevant for Dynamic Programming approaches.