16.410-13 Recitation 12 Problems

Problem 1: MDP Navigation

Captain Jack Sparrow, infamous pirate, has sailed his ship to the side of the island of Tortuga. See the figure below.

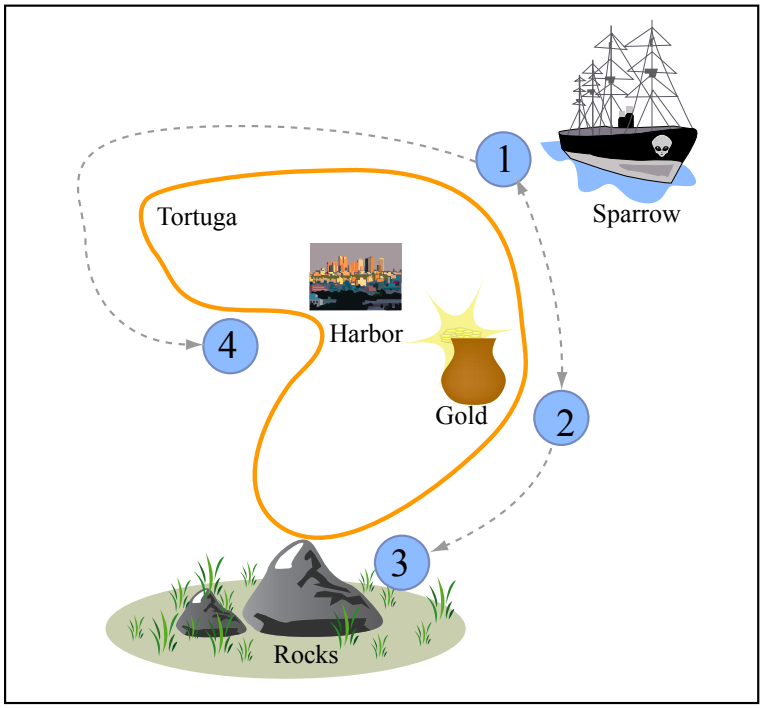

Image by MIT OpenCourseWare.

Captain Sparrow would like to anchor in the harbor on the western side of the island. Let's help him by using an ancient navigation technique that is known all sailors worth their salt: value iteration.

Consider the figure. There are four locations. The dotted arrows denote the valid moves between them. Ultimately Captain Sparrow wants to reach location 4, the harbor of Tortuga. Although, he would be very happy, if he could collect the gold at location 2 before reaching the harbor. However, there is a risk of a thunderstorm which may drag Sparrow's ship to location 3, which is near several rocks that can sink his ship. Assume that ship takes the gold, the first time it reaches location 2.

Let's assume that the goal location, i.e., location 4, has reward big reward. Also, let's assume that location 3 has a big negative reward. Finally, let's assume that Sparrow gets some positive reward when he travels to location 2 for the first time, since he collects the gold.

Part A: Modeling

In this part, you should formulate an MDP model of the system. Explicitly note how you handle the gold being in the ship. Provide a reasonable model of the system by writing the transition function and the reward function.

Part B: Value iteration

Say the discount factor is 0.5. Start with the value $V_0(s) = 0$ for all s, and execute the value iteration for one step, i.e., compute $V_1(s)$ for all $s \in S$.

Part C: Discussion on policies

How many policies are there?

Assume that the discount factor is γ, the reward for collecting the gold is R_G, the reward for reaching the harbor is R_H, and the reward for colliding with the rocks is R_R. Also, assume that once Sparrow starts traveling out from location 2, the probability that he ends up at the rocks at location 3 is p. Compute the value function for all the policies keeping γ, R_G, R_H, and R_H as parameters. Can you assess which one of these policies is better? *(HINT: Policy iteration algorithm was doing something along those lines).*

Part D: Discussion on the discount factor

How would the solution look for a discount factor close to one? How about when the discount factor is close to zero? What can you say when the discount factor is exactly one and exactly zero?

Problem 2: Markov decision processes[1]

Consider a planning planning problem modeled as an MDP with states $X = \{a, b, c\}$ and controls $U = \{1, 2, u_T\}$. The state $X_G = \{c\}$ is identified as the goal state. The actions $u = 1$ and $u = 2$ are shown in the figure below. Taking the action u_T, the agent can stay in its current state with probability one, i.e., $T(x, u_T, x) = 1$ for all $x \in \{a, b, c\}$.

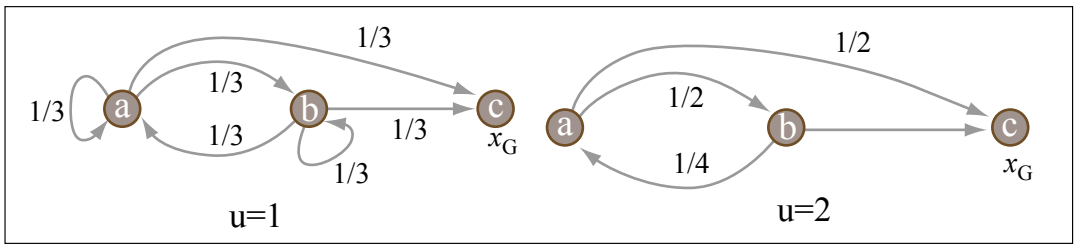

Image by MIT OpenCourseWare.

The reward structure is such that the agent gets reward 1 when it is in state c and takes action u_T, otherwise the reward is zero. That is,

$$R(x, a, y) = \begin{cases} 1 & \text{when } x = c, y = c, a = u_T; \\ 0 & \text{otherwise.} \end{cases}$$

The discount factor is 0.9. Please answer the following questions.

- How would you compute the value function using the value iteration and the optimal policy? Write down the equations for value iteration. Calculate the first iteration of the value iteration by hand.
- How would you compute the value function and the optimal policy using the policy iteration? Write down the equations for policy evaluation and policy improvement. Calculate the first iteration of the policy iteration by hand.

[1] this problem is based on an exercise in PA

16.410 / 16.413 Principles of Autonomy and Decision Making

Fall 2010