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.

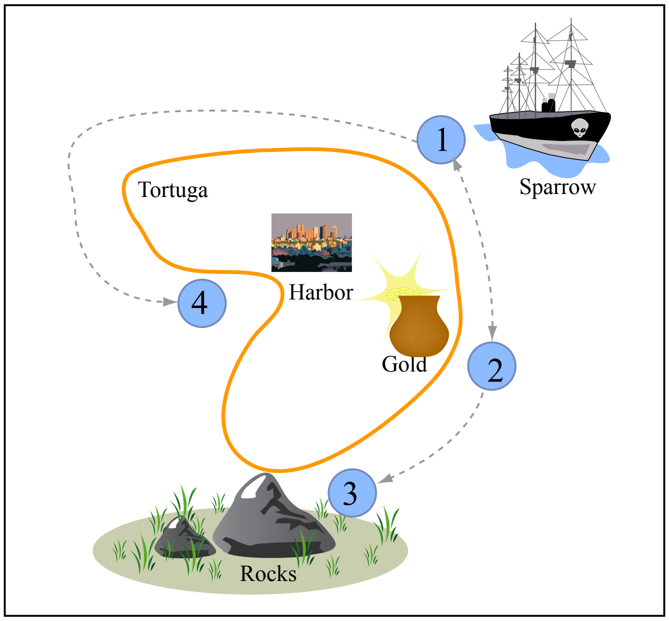


Image by MIT OpenCourseWare.

CCaptain 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.

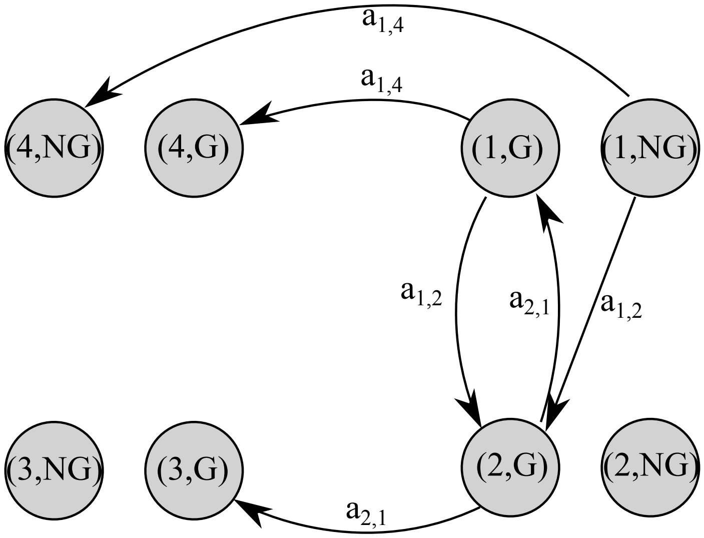
Let's assume that the goal location, i.e., location 4, has reward big reward. Also, let's assume that location 3 has a big negative reward. Finally, let's assume that Sparrow gets some positive reward when he travels to location 2 for the first time, since he collects the gold.

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.

The state must include whether Captain Sparrow has gold or not. Hence the state space will be $\{1, 2, 3, 4\} \times \{NG, G\}$, where 1, 2, 3, and 4 indicate the location whereas NG and G indicate that the captain has does not have the gold or has the gold, respectively.

Let us define the set of actions as follows: $\{a_{1,4}, a_{1,2}, a_{2,1}, a_T\}$. Intuitively, $a_{1,4}$ steers the ship from location 1 to location 4; $a_{1,2}$ steers the ship from location 1 to location 2, and $a_{2,1}$ steers the ship from location 2 to location 1. Applying action u_T, the Captain can stay at his current location with probability one. Let's assume that the Captain can steer his ship deterministically from location 1 to any other state. The only non-deterministic action is $a_{2,1}$, in which case the ship ends up at state 1 with probability p and state 3 with probability $1 - p$. Let us assign $p = 0.5$. To model collecting the gold, we will assume taking action $a_{1,2}$ when in state $(1, NG)$ the ship ends up at state $(2, G)$, with probability one.



Note that u_T is not shown in the graph. Notice that the states $(2, NG)$ and $(3, NG)$ are not reachable, with probability one. You can safely remove these states from your model.

To ensure that Captain Sparrow ends up at the harbor we assign the actions that reach location 4 reward 1000. To ensure that the Captain does not drift towards the rocks, let's assign reward -1000 to the triplet $((2, G), a_{2,1}, (3, G))$. Finally, let us assign the triplet $((1, NG), a_{1,2}, (2, G))$ reward 200 to represent the value of the gold. For all the other triplets, we can assign reward 0.

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 $s \in S$.

Recall that the value iteration is carried out by the following update equation:

$$V_{i+1}(s) \leftarrow \sum_{s' \in S} T(s, a, s')(R(s, a, s') + \gamma V_i(s')).$$

In the following, we provide the update for the first state only. The rest can be carried out similarly.

$$\begin{aligned} V_1((1, NG)) = \max_{a_{1,2}, a_{1,4}} \{ & T((1, NG), a_{1,2}, (2, G))(R((1, NG), a_{1,2}, (2, NG)) + \gamma V_0((2, NG))), \\ & T((1, NG), a_{1,4}, (4, NG))(R((1, NG), a_{1,4}, (4, NG)) + \gamma V_0((4, NG)))\} = 200. \end{aligned}$$

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).

The way the problem is set up, the only decision that the Captain can make comes up in two states: $(1, NG)$ and $(1, G)$. Sparrow can choose from one of the two options in both states. Hence, there are exactly 4 policies.

Notice that in state $(1, G)$, it does not make sense to go back to state $(2, G)$ since the only that can be collected is the negative reward when going into state $(3, G)$. Hence, the only state that there is a nontrivial choice is the initial state, $(1, NG)$, which reduces our policy space to only two policies: Sparrow can either choose to go to state $(2, G)$ and collect the gold and then head out to the harbor, or alternatively Sparrow can directly go to the harbor without collecting the gold. Let us denote these policies by π_1 and π_2, respectively. More precisely, we have

$$\pi_1((1, NG)) = a_{1,4}, \qquad \pi_1((1, G)) = a_{1,4}, \qquad \pi_1((2, G)) = a_{2,1}$$

and $\pi_1(\cdot) = a_T$ for all other states. The only difference in π_2 is that $\pi_2((1, NG)) = a_{1,2}$.

We can evaluate the value function for these two policies using the policy evaluation. Recall that for policy evaluation we need to solve the following system of linear equations.

$$V_\pi(s) = \sum_{s' \in S} T(s, \pi(s), s')[R(s, \pi(s), s') + \gamma V(s')] \qquad \text{for all } s \in S.$$

First consider the policy π_1, let us calculate the value function under this policy using policy evaluation. Then,

$$V_{\pi_1}((4, NG)) = 1[R((4, NG), a_T, (4, NG)) + \gamma V_{\pi_1}((4, NG))] = R_H + \gamma V_{\pi_1}((4, NG)))$$

Hence,

$$V_{\pi_1}((4, NG))) = \frac{R_H}{1 - \gamma}.$$

Then, we have

$$V_{\pi_1}((1, NG)) = 1[R((1, NG), a_{1,4}, (4, NG)) + \gamma V_{\pi_1}((4, NG))] = \gamma V_{\pi_1}((4, NG))$$

Hence, the value at the initial state is

$$V_{\pi_1}((1, NG)) = \frac{\gamma}{1 - \gamma} R_H. \tag{1}$$

Next, consider the policy π_2. Notice that

$$V_{\pi_2}((4, NG))) = V_{\pi_2}((4, G))) = \frac{R_H}{1 - \gamma},$$

as before. Also,

$$V_{\pi_2}((1, G)) = 1[R((1, G), a_{1,4}, (4, G)) + \gamma V_{\pi_2}((4, G))] = \gamma V_{\pi_2}((4, G)) = \frac{\gamma}{1 - \gamma} R_H.$$

Moreover,

$$V_{\pi_2}((3, G)) = 1[R((3, G), a_T, (3, G)) + \gamma V_{\pi_2}((3, G))].$$

Then,

$$V_{\pi_2}((3, G)) = \frac{1}{1 - \gamma} R_R.$$

Also,

$$V_{\pi_2}((2, G)) = p[R((2, G), a_{2,1}, (1, G)) + \gamma V_{\pi_2}((2, G))] + (1 - p)[R((2, G), a_{2,1}, (3, G) + \gamma V_{\pi_2}(3, G))],$$

which yields,

$$V_{\pi_2}((2, G)) = p \frac{\gamma}{1 - \gamma} R_R + (1 - p) \frac{\gamma^2}{1 - \gamma} R_H.$$

Finally,

$$V_{\pi_2}((1, NG)) = 1[R((1, NG), a_{1,2}, (2, G)) + \gamma V_{\pi_2}((2, G))] = R_G + \gamma V_{\pi_2}((2, G)),$$

which yields

$$V_{\pi_2}((1, NG)) = R_G + \gamma \left(p \frac{\gamma}{1 - \gamma} R_R + (1 - p) \frac{\gamma^2}{1 - \gamma} R_H \right). \tag{2}$$

Hence, the discounted reward that Sparrow get with the first policy is given in Eqn (1), and that he can collect using the second policy is given in Eqn (2).

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?

A high discount factor will value the future reward, hence will cause Sparrow to go directly to the Harbor. A low discount on the other hand will value immediate awards and will have Sparrow go try to collect the reward. We can compute the rewards that Sparrow can collect for the two policies of available to him using Equations (1) and (2).

When the discount factor is exactly one, the problem is not well-defined. For instance, for state $(4, G)$, the value function would take value infinity. When the discount factor is zero, the future is not taken into account at all. Hence, Sparrow, in that case, would directly go to location 2 to collect the gold (convince yourself that no other policy achieves a better reward).

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* $x \in \{a, b, c\}$.

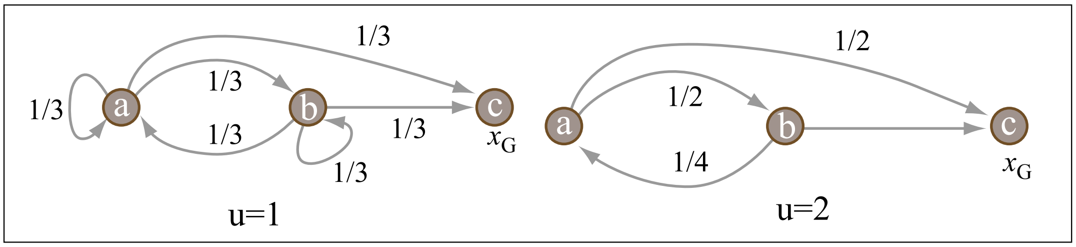


Image by MIT OpenCourseWare.

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,*

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

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

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

- *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.*

The value iteration algorithm is carried out by repeatedly applying the following update:

$$V_{i+1}(s) \leftarrow \max_a \sum_{s' \in S} T(s,a,s')(R(s,a,s') + \gamma V_i(s')).$$

Initially, you can set $V_i(s) = 0$ for all $s \in S$. The calculation is quite straightforward, and not provided here.

- *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.*

The policy iteration consists of two main steps: policy evaluation and policy improvement. The algorithm starts with an initial policy, and in each iteration, it first computes of the value of this policy by policy evaluation, and then computes a new policy that improves up on the current policy by policy improvement. These steps are repeated in an iterative manner, until the value function does not change.

Given a policy π, the policy evaluation step is carried out by solving the following linear system of equations:

$$V_\pi(s) = \sum_{s' \in S} T(s,\pi(s),s')[R(s,\pi(s),s') + \gamma V(s')]$$

The solution to these equations is the value function V that maps each state to a real number (its value under the policy π).

In the policy improvement step, the policy is refined to a better one, by using the value function of the current policy, i.e., the value function found in the policy evaluation step. The policy improvement step is carried out as follows:

$$\pi(s) \leftarrow \arg\max_a \sum_{s' \in S} T(s,a,s')[R(s,a,s') + \gamma V(s')]. \tag{3}$$

To solve the problem presented in the figure, first pick an initial policy, say $\pi(a) = 1$, $\pi(b) = 1$, and $\pi(c) = u_T$. Then, go through the policy evaluation and the policy improvement steps until the policy converges (i.e., it does not change anymore). The calculations are straightforward and are not provided here.

16.410 / 16.413 Principles of Autonomy and Decision Making
Fall 2010