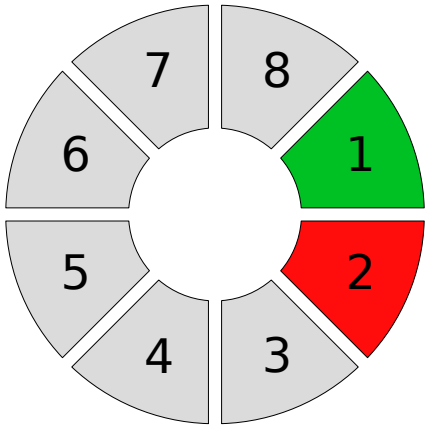
Artificial Intelligence
Exercise 3

Marc Toussaint
Machine Learning & Robotics lab, U Stuttgart
Universitätsstraße 38, 70569 Stuttgart, Germany

21. November 2018

1 Votieraufgabe: Value Iteration

(Teilaufgaben werden separat votiert.)



Consider the circle of states above, which depicts the 8 states of an MDP. The green state (#1) receives a reward of $r = 4096$ and is a 'tunnel' state (see below), the red state (#2) is punished with $r = -512$. Consider a discounting of $\gamma = 1/2$.

Description of $P(s'|s, a)$:

- The agent can choose between two actions: going one step clock-wise or one step counter-clock-wise.
- With probability 3/4 the agent will transition to the desired state, with probability 1/4 to the state in opposite direction.
- Exception: When $s = 1$ (the green state) the next state will be $s' = 4$, independent of a. The Markov Decision Process never ends.

Description of $R(s, a)$:

- The agent receives a reward of $r = 4096$ when $s = 1$ (the green state).
- The agent receives a reward of $r = -512$ when $s = 2$ (the red state).
- The agent receives zero reward otherwise.

(a) Perform three steps of Value Iteration: Initialize $V_{k=0}(s) = 0$, what is $V_{k=1}(s)$, $V_{k=2}(s)$, $V_{k=3}(s)$?

(b) How can you compute the value function $V^\pi(s)$ of a GIVEN policy (e.g., always walk clock-wise) in closed form? Provide an explicit matrix equation.

(c) Assume you are given $V^*(s)$. How can you compute the optimal $Q^*(s, a)$ form this? And assume $Q^*(s, a)$ is given, how can you compute the optimal $V^*(s)$ from this? Provide general equations.

(d) What is $Q_{k=3}(s, a)$ for the example above? What is the "optimal" policy given $Q_{k=3}$?

2 Programmieraufgabe: Value Iteration

In the repository you find python code to load the probability table $P(s'|a,s)$ and the reward function $R(a,s)$ for the maze of Exercise 1. In addition, the MDP is defined by $\gamma = 0.5$.

(a) Implement Value Iteration to reproduce the results of Exercise 1(a). Tip: An easy way to implement this is to iterate the two equations:

$$Q(s,a) \leftarrow R(s,a) + \gamma \sum_{s'} P(s'|s,a)\ V(s') \tag{1}$$

$$V(s) \leftarrow \max_a Q(s,a) \tag{2}$$

Compare with the value functions $V_{k=1}(s)$, $V_{k=2}(s)$, $V_{k=3}(s)$ computed by hand. Also compute the $V_{k=100} \approx V^*$.

(b) Implement Q-Iteration for exactly the same setting. Check that $V(s) = \max_a Q(s,a)$ converges to the same optimal value.

WARNING: The test you have in your repository only tests for the specific world of Exercise 1. However, our evaluation will test your method also for other MDPs with other states, actions, rewards, and transitions! Implement general methods.

3 Präsenzaufgabe: The Tiger Problem

Assume that the tiger is truly behind the left door. Consider an agent that always chooses to listen.

a) Compute the belief state after each iteration.

b) In each iteration, estimate the expected reward of open-left/open-right based only on the current belief state.

c) When should the agent stop listening and open a door for a discount factor of $\gamma = 1$? (How would this change if there were zero costs for listening?)