CSE 571 Fall 2020

Homework 5

Due *Nov 25, Wednesday* online

Only typed answers will be accepted.

Be precise and concise in your answers. You may add hand-drawn figures when necessary.

Exercise 1.1 (8pt)

Sometimes MDPs are formulated with a reward function R(s, a) that depends only on the current state and action taken or with a reward function R(s) that only depends on the current state.

a. (2pt) Write the Bellman equations for these formulations ***for the optimal value function.***

b. (3pt) Show how an MDP with reward function R(s, a, s′) can be transformed into a different MDP with reward function R(s, a), such that optimal policies in the new MDP correspond exactly to optimal policies in the original MDP. ***You must formally define the new MDP (its components) based on the old MDP.***

c. (3pt) Now do the same to convert MDPs with R(s, a, s') into MDPs with R(s). ***You must formally define the new MDP (its components) based on the old MDP.***

Exercise 1.2 (10pt)

Consider the 3 × 3 world shown below. The transition model is the same as in our robot domain: 80% of the time the agent goes in the direction it selects; the rest of the time it moves at right angles to the intended direction.

r	-1	+10
-1	-1	-1
-1	-1	-1

Use discounted rewards with a discount factor of 0.99. Show the policy obtained in each case. ***Explain intuitively*** why the value of r leads to each policy.

a. r = 100

b. r = −3

c. r = 0

d. r = +3

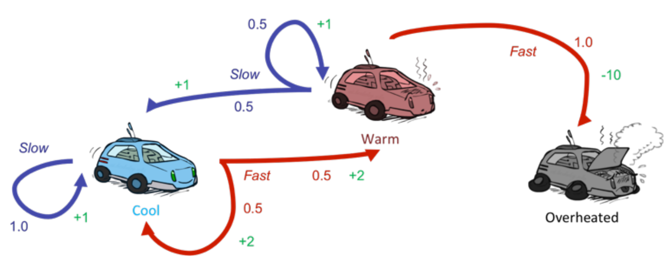
Exercise 1.3 (8pt)

Consider the 101 × 3 world shown below. In the start state the agent has a choice of two deterministic actions, Up or Down, but in the other states the agent has one deterministic action, Right. Assuming a discounted reward function, for what values of the discount γ should the agent choose Up and for which Down? Compute the utility of each action as a function of γ. (Note that this simple example actually reflects many real-world situations in which one must weigh the value of an immediate action versus the potential continual long-term consequences, such as choosing to dump pollutants into a lake.)

+50	-1	-1	-1	...	-1	-1	-1	$\boxed{-1}$
Start				...				
-50	+1	+1	+1	...	+1	+1	+1	$\boxed{+1}$

Exercise 1.4 (12pt)

Apply policy iteration, showing each step in full, to determine the optimal policy when the ***initial policy*** is $\boldsymbol{\pi}$(cool) = Slow and $\boldsymbol{\pi}$(warm) = Fast. ***Show both the policy evaluation and policy improvement steps clearly until convergence.*** Assuming a discount factor of 0.5.



Exercise 1.5 (12pt)

Consider the car domain above (without knowing the T or R) and given the following experiences:

Episode 1:
cool, slow, cool, +1
cool, slow, cool, +1
cool, fast, cool, +2
cool, fast, cool, +2
cool, fast, warm, +2
warm, fast, overheated, -10

Episode 2:
warm, slow, warm, +1
warm, slow, cool, +1
cool, slow, cool, +1
cool, fast, warm, +2
warm, slow, warm, +1
warm, slow, warm, +1
warm, fast, overheated, -10

a. (2pt) Estimating the parameters for T and R for model-based reinforcement learning.

b. (2pt) Use MC reinforcement learning method (direct evaluation) to estimate the V function, assuming $\gamma = 1.0$.

c. (4pt) Assuming that the initial state values are all zeros, compute the updates ***in TD learning for policy evaluation (passive RL)*** to the state values after running through episode 1 and 2 in sequence. Show steps for $\alpha = 0.5$ and $\gamma = 1.0$.

d. (4pt) Assuming that the initial Q values are all zeros, compute the updates ***in Q learning (active RL)*** to the Q values after running through episode 1 and 2 in sequence. Show steps for $\alpha = 0.5$ and $\gamma = 1.0$.