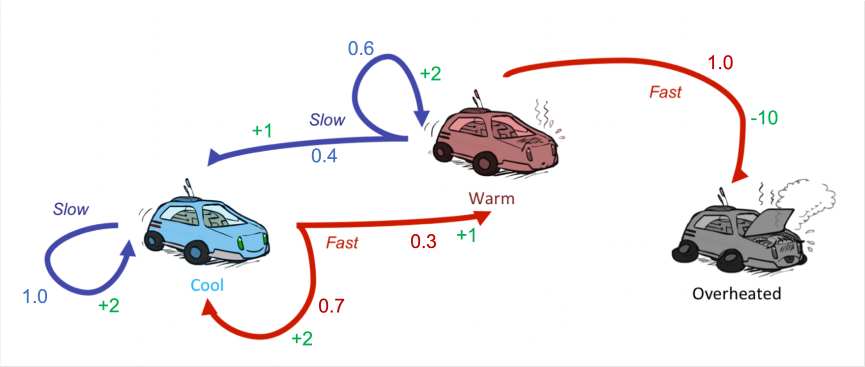
CS 4300 **Homework 3** **Prof. Alan Kuntz**

Place your answers in the spaces provided, if you need more room write on the back of the pages. Round to two decimal places when rounding is required. ***SHOW ALL WORK.***

1 Value Iteration

Consider the following MDP example that is similar to, *but not identical to* the one discussed in class. In this MDP, the states are *Cool*, *Warm*, and *Overheated*. You can take two actions from both Cool and Warm: Fast and Slow. You can take no actions from Overheated. The transition model is represented by the arrows with corresponding probability (red and blue) and reward (green). For example, if the car is in state *Warm* and takes action Fast, with a probability of 1.0 it will transition to state *Overheated* and receive a reward of -10. For this problem use a discount factor of $\gamma = 1$.



i	$V_i(Cool)$	$V_i(Warm)$	$V_i(Overheated)$
0	0	0	0
1			
2			

(a) Perform the first two iterations of value iteration and fill in the value table above. Show all work.

(b) Suppose after k iterations of value iteration we end up with $V_k(Cool) = 20$, $V_k(Warm) = 10$, $V_k(Overheated) = 0$. Using these values do policy extraction to find $\pi^*(Cool)$ and $\pi^*(Warm)$. Please show all work.

2 Policy Iteration

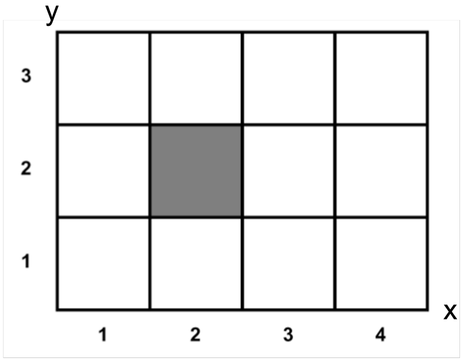
(a) For the same MDP as above: Run two iterations of policy evaluation (starting with all state values initialized to zero) for the potentially suboptimal policy $\pi_0(Cool) = Fast$ and $\pi_0(Warm) = Slow$ to determine the state values under the policy. Please fill in the table below, showing all work.

i	$V_i(Cool)$	$V_i(Warm)$	$V_i(Overheated)$
0	0	0	0
1			
2			

(b) Use the state values obtained from 2(a) to perform one iteration of policy extraction, i.e., determine $\pi_1(Cool)$ and $\pi_1(Warm)$. Please show all work.

3 Reinforcement Learning

Consider the following grid-world:



Suppose that we run two episodes that yield the following sequences of (state (x, y), action, reward) tuples:

Episode 1			Episode 2		
(1,1)	up	-1	(1,1)	up	-1
(2,1)	up	-1	(1,2)	up	-1
(1,1)	right	-1	(1,3)	right	-1
(2,1)	right	-1	(2,3)	right	-1
(3,1)	up	-1	(3,3)	right	-1
(4,1)	left	-1	(4,3)	exit	+50
(3,1)	up	-1	(done)		
(3,2)	up	-1			
(3,3)	up	-1			
(4,3)	exit	+50			
(done)					

(a) Perform the direct evaluation method for both episodes, what are the resulting values for every state in the grid? Fill in the grid above with the values. Don't bother listing the values for which we have no information.

(b) According to model-based learning, what are the transition probabilities for every (state, action, state) tuple computed from these episodes? Don't bother listing the tuples for which we have no information.

(c) Suppose that the values of all states are initialized to zero. Run temporal difference learning for Episode 1 only. Use a discount factor and learning rate of $\gamma = 1$ and $\alpha = 0.5$, respectively. What are the resulting values for each state? Fill in the table below with the resulting values. Don't bother listing the values for which we have no information. (For this problem treat '(done)' as a state that always has value 0.)