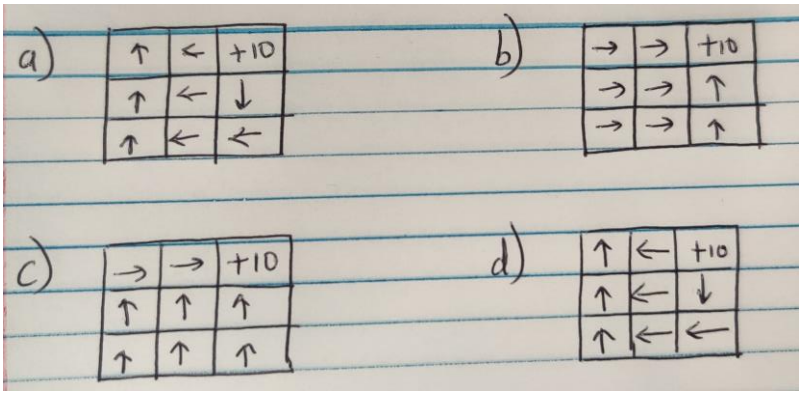
# CSE 571 Fall 2022 HW5

**Submitted By: Aman Peshin (1225476655)**

## Exercise 1.2




a. We know from the question that the bottom left square is the origin.
For r = 100
The agent will stay in the location of (1,3) since its reward in that state is very much higher than it would receive from the final state which is 10. The agent is going to try and get to square (1,3) from any location.
In square (3,2) the agent tends to go down since the probability to get to location (3,3) is zero. It is going to go to any location other than (3,2).
There is a chance that it will end in the final state.
In square (1,3) the agent will go north to increase its total reward and will try to elongate this.

b. For r = -3
Here we see that the agent avoids location (1,3) as its reward is -3 as compared to -1 in other locations.
Since the reward in the final state is 10. The agent moves towards the location (3,3) as quickly as it can)
In location (1,2) the agent moves towards the right with a probability of 0.1 that it would move to location (1,3). Since that state has a lower reward as compared to other squares but is very much lower as compared to -1. That's why we are going right from (1,2) as opposed to going down. As going down has zero chance, the agent will take a lot of time to reach the final state.

c. For r = 0
Here we see that the reward at (1,3) is zero and that value is slightly greater than the

value at other location other than the final location/state.
The reward that we see at the final state is 10, the agent will try and move as quickly towards the target location as fast as possible.
The path would be as follows
(1,1)(1,2)(2,1)(2,2)(3,1)(3,2)
Which is different from the one going from the left through (1,3). It's going to increase the reward as the value at (1,3) is greater than -1.
Since this is not positive the agent will still progress toward the final state.

d. For r = +3
Here, in this case, we see the reward is +3 at (1,3) which is a high reward as compared to a state other than the final state. Even though the reward at the final state is 10 it will try and get the closer reward at (1,3) and stay put at that location as to increase the total reward.
In another square (3,2) The agent tends to go down because the probability of going to location (3,3) i.e. final state is 0. If it goes in rather a direction from location (3,2) it has a non-zero chance it will end up in the final state.

## Exercise 1.3

we are going to solve this problem using the bellman equation.

We are going to check the values of gamma utility when going up and going down
Utility of going UP

$$Utiltiy(up) = 50 * \gamma - \gamma^2 - \gamma^3 - \gamma^4 - \gamma^5 - \gamma^6 \ldots - \gamma^{101}$$

$$\Rightarrow Utiltiy(up) = 50 * \gamma - \sum_{i=2}^{101} \gamma^i$$

We are solving this equation as a sum of Geometric progression

$$\Rightarrow Utiltiy(up) = 50 * \gamma - \gamma^2 * (\frac{1-\gamma^{100}}{1-\gamma})$$

Utility of going DOWN

$$Utiltiy(down) = -(50 * \gamma) + \gamma^2 + \gamma^3 + \gamma^4 + \gamma^5 + \gamma^6 \ldots \mp \gamma^{101} \Rightarrow$$

We are solving this equation as a sum of Geometric progression

$$U(DOWN) = \gamma^2 * \left(\frac{1-\gamma^{100}}{1-\gamma}\right) - 50 * \gamma$$

This will be the final utility function in terms of $\gamma$ for going DOWN.

The agent will decide to move UP if the value is less than $\gamma$ and DOWN if the value is greater than $\gamma$ in order to avoid expensiveness.
Let us try to approximate this value by taking the utility of the value Up and utility down.

Utility (Up) = Utility (Down)
By simplifying we get $\gamma$ = 0.9844
For value > $\gamma$ -> we go down
For value < $\gamma$ -> we go up.

## Exercise 1.4

$\gamma$ which is the discount factor which is 0.5

We need to determine the optimal policy when the starting policy is
$\pi_0$ (cool) = slow
$\pi_0$ (warm) = slow

Let us formulate the equations for the car domain
V(cool) -> 1 * (1 + 0.5 * V(cool))
V(warm) - > 0.5 * (1 + 0.5 * V(cool)) + 0.5 * ( 1 + 0.5 * V(warm))
V(overheat) = 0

Let us solve for the first Iteration.
Using the initial policy values and the calculated values

V(cool) = 1/0.5 = 2
V(warm) →0.5 * (1 + 0.5 * V(cool)) + 0.5 * ( 1 + 0.5 * V(warm))
= 0.5 * (1 + 0.5 * 2) + 0.5 * ( 1 + 0.5 * V(warm))
= 0.75 * V(warm) = 1.5
V(warm) = 2
V(overheated) = 0

Now, the Policy is as follows:

$\pi_1$ (cool) = max (slow, fast) = 3 → Fast
slow-> 1 * (1 + 0.5 * 2)

fast -> 0.5 * (2 + 0.5 *2) + 0.5 * (2 + 0.5 * 2)

$\pi_1$ (warm) = max (slow, fast) = 2 → Slow
slow-> 0.5 * (1 + 0.5 * 2) + 0.5 * (1 + 0.5 * 2)
fast -> 1 * (-10 + 0.5 * 0)

Let us solve for the second iteration
$\pi_1$ (cool) = fast
$\pi_1$ (warm) = slow
here we have two linear equations for V(cool) and V(warm)
V(cool) = 0.5 * (2 + 0.5 * V(cool)) + 0.5 * (2 + 0.5 * V(warm))
V(warm) = 0.5 * (1 + 0.5 * V(cool)) + 0.5 * (1 +0.5 * (V(warm))

We get V(cool) = 3.5 and V(cool) = 2.5
$\pi_2$ (cool) = max (slow, fast) = 3.5 → Fast
slow-> 1 * (1 + 0.5 * 3.5 )
fast -> 0.5 * (2 + 0.5*3.5) + 0.5 * (2 + 0.5 *2.5)

$\pi_2$ (warm) = max (slow, fast) = 2.5 → Slow
slow-> 0.5 * (1 + 0.5 * 3.5) + 0.5 * (1 + 0.5 * 2.5)
fast -> 1 * (-10 + 0.5 * 0)

Here we see that policy 1 and policy 2 for states WARM and COOL. We see that for both
COOL = Fast and Fast
WARM = Fast and Slow

## Exercise 1.5

a. Estimated values of T and R are as follows:
   For (cool, slow, cool):
   $T(\text{cool, slow, cool}) = 1$
   $R(\text{cool, slow, cool}) = 1$

   For (cool, fast, cool):
   $T(\text{cool, fast, cool}) = 3/6$
   $R(\text{cool, fast, cool}) = 2$

   For (cool, fast, warm):
   $T(\text{cool, fast, warm}) = 3/6$
   $R(\text{cool, fast, warm}) = 2$

For (warm, fast, overheat):
T(warm, fast, overheat) = 1
R(warm, fast, overheat) = -10

For (warm, slow, cool):
T(warm, slow, cool) = 1
R(warm, slow, cool) = 1

b. For(cool, slow) :
Q(cool, slow)
= ((- 2 - 3) + (- 5)) / 3
= - 10 / 3

For(cool, fast):
Q(cool, fast)
= ((- 4 - 6 - 8) + (- 2 - 6 - 8)) / 6
= - 34 / 6

For(warm, slow):
Q(warm, slow)
= ((0) + (- 4)) / 1
= - 4

For(warm, fast):
Q(warm, fast)
= ((- 10) + (- 10)) / 2
= - 10

c. V(S) < -(1 - ) V(S) + Instance
Instance -R (S,(S), ST) + V(ST)

V values of cool, warm, overheat
V(cool) = 0, V(warm) = 0, V(overheated) = 0

For cool:
V(cool)
= ((1-0.5)0)+(0.5(1+0))
=0.5

V(cool)
= ((1-0.5)0.5)+(0.5(1+0.5))
=1

V(cool)
= ((1-0.5)1)+(0.5(2+1))
=2

V(cool)
= ((1-0.5)2)+(0.5(2+2))
=3

V(cool)
= ((1-0.5)3)+(0.5(2+0))
=2.5

For Warm:
V(warm)
= ((1-0.5)0)+(0.5(-10+0))
= -5

V(cool)
= ((1-0.5)2.5)+(0.5(2-5))
= -0.25

V(warm)
= ((1-0.5) (-5))+(0.5(1-0.25))
= -2.125

V(cool)
= ((1-0.5)(-0.25))+(0.5(1-0.25))
=0.25

V(cool)
= ((1-0.5)0.25)+(0.5(2-0.25))
=1.25

V(cool)
= ((1-0.5)1.25)+(0.5(2-2.125))
=0.5625

V(warm)
= ((1-0.5)(-2.125))+(0.5(-10+0))
= - 6.0625.

V values of cool, warm, overheat
V (cool) = 0.5625, V(warm) = -6.0625, V(overheated) = 0

d. $Q^{T+1}(S, A) = (1 - \alpha)\, Q^{T}(S, A)+\alpha(r+\max_a Q^{T}(S', A'))$

Q (cool, slow)
= 0

Q (cool, fast)
= 0

Q (warm, slow)
= 0

Q (warm, fast)
= 0

Q (cool, slow)
= (1-0.5)0+0.5(1+max (0, 0))
=0.5

Q (cool, slow)
= (1-0.5)0.5+0.5(1+max (0, 0.5))
=1

Q (cool, fast)
= (1-0.5)0+0.5(2+max (1, 0))
=1.5

Q (cool, fast)
= (1-0.5)1.5+0.5(2+max (1.5, 1))
=2.5

Q (cool, fast)
= (1 - 0.5) 2.5 + 0.5 (2+max (0, 0))
=2.25

Q (warm, fast)
= (1 - 0.5)0 + 0.5(-10+max (0))
= -5

Q (cool, fast)
= (1 - 0.5)2.25+0.5(2+ max (0,-5))
=2.125

Q (warm, slow)
= (1-0.5)0+0.5(1+max (1, 2.125))
=1.5625

Q (cool, slow)

= (1-0.5)1+0.5(1+max (1, 2.125))
=2.0625

Q (cool, fast)
= (1-0.5)2.125+0.5(2+max (2.0625, 2.125))
=3.125

Q (cool, fast)
= (1 - 0.5) 3.125 + 0.5(2 + max (1.5625,-5))
=3.34375

Q (warm, fast)
= (1-0.5) (-5) + 0.5(-10+ max (0))
= -7.5

## Exercise 1.1

a. Since the reward is gained after taking an action, we add it prior to the values of the states that come before

$Q^*(s, a) = R(s, a) + \gamma \sum T(s, a, s') V^*(S')$

$Q^*(s, a) = \sum T(s, a, s') [R(s') + \gamma V^*(s)]$

b. Let us create a pre-state i.e. P(s, a,s') where an action executing action a from state s leads to it
Therefore we can have the new MDP as

$T'(s, a, \mathrm{P}(s, a, s'))$ when it was previously
$T(s, a, s')$

Now we see that:
$T'(\mathrm{P}(s, a, s'), b, s') = 1$

And also reward function is:
$R'(s, a) = 0$

Therefore the new reward function is
$R'(\mathrm{P}(s, a, s'), b) = \gamma^{-1/2} R(s, a, s')$

Where gamma
$\gamma' = \gamma^{1/2}$

C. $T'(s, a, \mathrm{P}(s, a, s'))$ when it was previously $T(s, a, s')$

The new transition function is
$T'(\mathrm{P}(s, a, s'), b, s') = 1$

Where the reward is
$R'(s) = 0$

The new reward becomes
$R'(\mathrm{P}(s, a, s')) = \gamma^{-1/2} R(s, a, s')$

where gamma
$\gamma' = \gamma^{1/2}$