Nabin Chapagain 1001551151

Task 1

value_iteration.py environment2.txt -0.04 1 20

```
 utilities:
 0.699  0.764  0.910  1.000
 0.577  0.000  0.588 -1.000
 0.514  0.413  0.500  0.289

 policy:
> > > o
^ X ^ o
^ < ^ <
```

value_iteration.py environment2.txt -0.04 0.9 20

```
 utilities:
 0.447  0.582  0.791  1.000
 0.310  0.000  0.439 -1.000
 0.220  0.195  0.303  0.097

 policy:
> > > o
^ X ^ o
^ > ^ <
```

Task 2

a) What value would you assign for the reward of the non-terminal states? Why?
→ For the non-terminal states, I would assign value of zero as in the game of chess, there is either win or loss and the reward is +1 waiting at the end of the board as a result of checkmate. We do not care about intermediate rewards and only care about the reward (or punishment) at the end, 0 will be assigned for the reward of non-terminal states.

b) What value would you use for the discount factor γ? Why?
→ Since the use of bigger value of γ would result in under fitting and the use of very very small γ would result in over fitting of the model, it is better to use a small γ but not so small that it leads to over-fit model.

Task 3

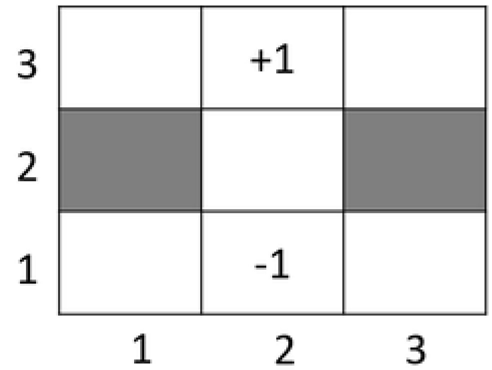

Figure 3: The environment to be considered in this task.

a) Suppose that the reward for non-terminal states is -0.04, and that $\gamma=0.9$. What is the utility for state (2,2)? Show how you compute this utility.

→ We start from (2,2) and we want to end up at (3,2). So,
$U((2,2)|(3,2)) = -0.04+0.8*1$ (*using the formula*)
$=0.76$

And, now for U(2,2), we calculate:
E(U((2,2),……….,s)) and as we know for optimal policy:

$U(2,2) = 0.8*(0.76) + 0.2(-0.04+0.9*U(2,2))$

This is because we are climbing upwards and with 80% chance that we reach the destination, there remains 20% chance that we take a right or left which just leaves us at the original place hence, 0.2 times utility of itself.

Now,
$U(2,2) = 0.8*0.76 - 0.2*(0.04+0.9*U(2,2))$
$U(2,2) = 0.608-0.008-0.18\ U(2,2)$
$1.18U(2,2) = 0.6$

U(2,2) = 0.51

Nabin Chapagain 1001551151

b) Suppose that $\gamma=0.9$, and that the reward for non-terminal states is an unspecified real number r (that can be positive or negative). For state (2,2), give the precise range of values for r for which the "up" action is not optimal. Show how you compute that range.

Now, for ‘up’ action:
U(3,2)=r+0.9(0.8*1)
and for us to stay on (2,2) with ‘up’ action:
U(2,2) = 0.8*(r+0.9(0.8*1)-0.2*(r+0.9*0.51)) (*we calculated 0.51 in last question*)
now, for ‘up’ action to not be optimal, clearly,
U(2,2)>U(3,2)
I.e;

0.8(r+0.72-0.2(r+0.45)) > r+0.72
0.64r+0.504 > r+0.72
0.504 – 0.72 > r – 0.64r
-0.216 > 0.36r

r < -0.6

So, as we can see for r less than negative 0.6, the ‘up’ action will not be optimal.