Assignment 1: Gridworld Exploration

Submission by: Apoorv Agnihotri (6604679), Gaurav Niranjan (6599177), Carla López Martínez (6637484)

1. Optimal Policy

Using the following formula for the return of this system we can solve this exercise:
$G_t = R_{t+1} + \gamma \cdot R_{t+2}\$, where\$R_{t+1}$ = 1 for the left action and 0 for the right action, and R_{t+2} = 0 for the left action and 2 for the right action.
Following are different cases with different values of γ and the optimal policy for each case.

- If we consider $\gamma = 0$, the optimal policy maximizes the immediate reward as $G_t = R_{t+1}$. In this case, the optimal policy is π_{left}. This is because the reward for moving to the left in the first state is 1. Meanwhile, the other action (right) leads to an immediate reward of 0. Thus, the optimal policy is to move to the left at each step.
- If we consider $\gamma = 0.9$, $G_t = R_{t+1} + 0.9 \cdot R_{t+2}$. The optimal policy is π_{right}. This is because the reward for moving to the right in the first time step is 0, but the reward in the next timestep is 2, which when multiplied by 0.9 (the discount factor) leads to a total reward of 0 + 1.8. Meanwhile, the other action (left) leads to an immediate reward of 1 and a delayed reward of 0 (multiplied by the discount of 0.9). π_{left} gives a cumulative reward of 1. Thus, we choose the optimal policy as π_{right}.
- If we consider $\gamma = 0.5$, either of the policies is optimal because in either case, we get the same reward $G_t = 1 + 0.5 \cdot 0 = 0 + 0.5 \cdot 2 = 1$. The optimal policy is π_{left} or π_{right}.

2. Value Estimation in Grid Worlds: implement return computation and value estimation

a. We tried the following values of k (num of episodes) to calculate the mean and standard deviation of the long-term average discounted rewards of the start state using MazeGrid environment if we were to use the random agent.

k	mean	std
1	0.185302	0
10	0.073958	0.209786
100	-0.093550	0.244046
1000	-0.058471	0.205473
5000	-0.061496	0.206545
10000	-0.061757	0.207711

b. To get a 95% confidence that our mean is within ± 0.0004 of the true mean, we use the law of large numbers. We can assume that the returns from different episodes are iid. According to CLT, we know that n independent samples from a population with mean μ and standard deviation σ will have the mean X^{bar} distributed approximately with $N(\mu, \frac{\sigma}{\sqrt{n}})$. Now we use the formula that relates confidence intereval of the mean with standard deviation and the number of samples. This is a result of using the formula of a Guassian distribution.

$n = (z_{\frac{\alpha}{2}} * \sigma / E)^2$

Where:
z = 1.96 (for 95% confidence)
σ = standard deviation (from part a)
E = 0.0004 (desired margin of error)

We can only use the above formula when we already have a good estimate of the standard deviation. In this case, we can use the standard deviation from the previous part to calculate the number of samples needed to get a 95% confidence that our mean is within ± 0.0004 of the true mean. Since we observed that the standard deviation kind of stopped changing after 1000 episodes, we can use the standard deviation from 10000 episodes to calculate the number of samples needed.

$n = (1.96 * 0.21 / 0.0004)^2 = 1,058,841$

This means we need around ~1,000,000 episodes to get a 95% confidence that our mean is within ± 0.0004 of the true mean.

c. We here need a margin of error ± 0.05. Below is the table we calculated in part a for the new environment "DiscountGrid" with a discount factor of 0.95.

k	mean	std
1	0.291989	0
10	-0.265131	0.442209
100	-0.144123	0.354744
1000	-0.098549	0.370524
5000	-0.117813	0.363687
10000	-0.106302	0.361857

We can use the same formula as in part b to calculate the number of samples needed to get a 95% confidence that our mean is within ± 0.05 of the true mean. We can use the standard deviation from 10000 episodes to calculate the number of samples needed.

$n = (1.96 * 0.36 / 0.05)^2 = 199.15$

This means we need around ~200 episodes to get a 95% confidence that our mean is within ± 0.05 of the true mean.

d. Getting the mean long-term discounted rewards with just 500 episodes returns with the following values:

```
| k     |  mean     | std       |
|-------| ----------|-----------|
| 500   | -0.108697 | 0.372299 |
```

With 500 episodes, E = 1.96*0.37/sqrt(500) = 0.032 < 0.1 -> the value estimate will be within the specified confidence interval. We can see in the table,
that for 500 episodes the mean is -0.108697, which is in the ± 0.05 range for the long-term mean -0.106302 (for 10000 episodes).