Reinforcement Learning Homework 1

Date 21st October 2024
Students Sahiti Chebolu, Surabhi S Nath, Xin Sui

Question 1

Assuming the policy is stationary and deterministic, the value for the top state (top) is given by the sum of discounted rewards from top:
$V_{left}(top) = 1 + \gamma \times 0 + \gamma^2 \times 1 + \gamma^3 \times 0 + \gamma^4 \times 1 + ... = \frac{1}{1-\gamma^2}$
$V_{right}(top) = 0 + \gamma \times 2 + \gamma^2 \times 0 + \gamma^3 \times 2 + \gamma^4 \times 0 + ... = \frac{2\times\gamma}{1-\gamma^2}$

The policy at top that maximises expected return is optimal:

For $\gamma = 0$, π_{left} is optimal since
$V_{left}(top) = 1 > V_{right}(s) = 0$

For $\gamma = 0.9$, π_{right} is optimal since
$V_{left}(top) = \frac{1}{0.19} < V_{right}(top) = \frac{1.8}{0.19}$

For $\gamma = 0.5$, both policies are equally good since
$V_{left}(top) = \frac{1}{0.75} = V_{right}(top) = \frac{1}{0.75}$

Question 2

We modified `gridworld.py` to calculate the return of an episode and the estimated value of start state. The return for an episode is simply the sum of discounted rewards from the start state.

```
returns += reward * (discount**timestep)
timestep += 1
```

The estimated value for the start state over multiple episodes is the mean of the returns over the episodes.

```
returns = []
for episode in range(1, opts.episodes + 1):
        ep_return = runEpisode(...) # returns the above computed return
        returns.append(ep_return)
print("Mean returns:", np.mean(returns))
print("Std returns:", np.std(returns))
```

2.2 a

In Figure 2 we plot the estimated value (mean returns) and standard deviation of returns across increasing number of episodes (k):

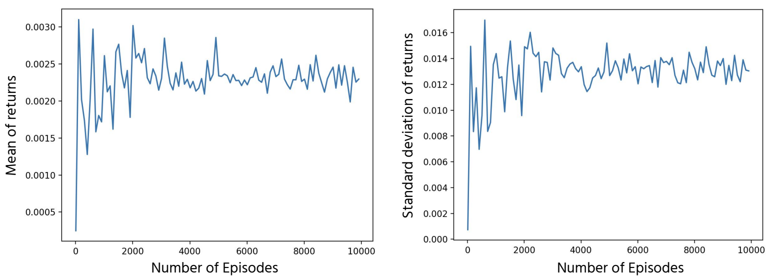

Figure 1: Plot of mean and standard deviation of returns across increasing number of episodes

Initially, the mean and returns are jittery but eventually converge. For $k = 10000$, mean returns ≈ 0.0024 and std deviation ≈ 0.0135

2.2 b

According to the central limit theorem, the distribution of sample means approaches a normal distribution with mean = true mean (μ) of the population and standard deviation = $\frac{\sigma}{\sqrt{n}}$ as the sample size n approaches infinity, where σ is the standard deviation of the population.

We want to find n where the sample mean is within ± 0.0004 of true mean with 95% confidence. That is, $1.96 * \frac{\sigma}{\sqrt{n}} = 0.0004$. Since, $\sigma \approx 0.0135$, we get $n \approx 4375$.

2.2 c

Now we repeat this procedure for the `DiscountGrid`. We plot mean and standard deviation of returns across increasing number of episodes (k):

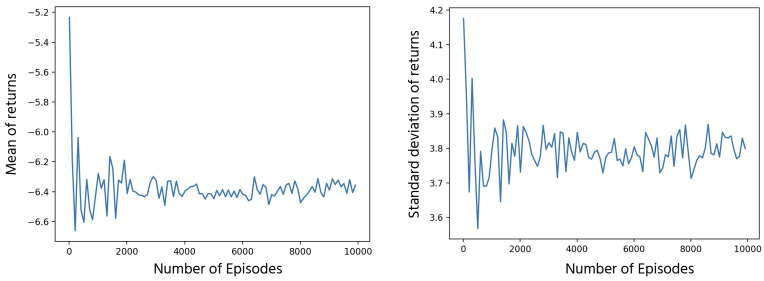

Figure 2: Plot of mean and standard deviation of returns across increasing number of episodes

For $k = 10000$, mean returns ≈ -6.40 and std deviation ≈ 3.8
Sample size n such that mean estimate is ± 0.05 of true mean with 95% confidence is ≈ 22190.

2.2 d

Since we need sample size ≈ 22000 to get sample estimate within ± 0.05 of true mean with 95% confidence, $k = 500$ is not enough.