Reinforcement Learning HW 1

Mert Bilgin (7034879)
mert.bilgin@student.uni-tuebingen.de

Lalitha Sivakumar (6300674)
lalitha.sivakumar@student.uni-tuebingen.de

Kevin Van Le (7314700)
kevin-van.le@student.uni-tuebingen.de

October 21, 2025

1 Optimal Policy

The deterministic nature of the system transforms the Bellman equation as follows:

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t \mid S_t = s \right], \qquad G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}.$$

Left Policy

The sequence of rewards for the **left policy** is:

$$[1, 0, 1, 0, \dots]$$

with nonzero rewards occurring at every second (odd) time step.

$$\begin{aligned} v_{\pi_{\text{left}}}(s) &= G_t \Big|_{\pi=\text{left}} \\ &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \Big|_{\pi=\text{left}} \\ &= \sum_{k=0}^{\infty} \gamma^{2k} = 1 + \gamma^2 + \gamma^4 + \cdots = \frac{1}{1-\gamma^2}. \end{aligned}$$

Right Policy

For the **right policy**, the reward sequence is:

$$[0, 2, 0, 2, \dots]$$

with nonzero rewards at even-indexed steps.

$$
\begin{aligned}
v_{\pi_{\text{right}}}(s) &= G_t\Big|_{\pi=\text{right}} \\
&= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}\Big|_{\pi=\text{right}} \\
&= \sum_{k=0}^{\infty} 2\gamma^{2k+1} = 2\gamma + 2\gamma^3 + 2\gamma^5 + \cdots = \frac{2\gamma}{1-\gamma^2}.
\end{aligned}
$$

$$
v_{\pi_{\text{right}}}(s) > v_{\pi_{\text{left}}}(s) \quad \Longleftrightarrow \quad \frac{2\gamma}{1-\gamma^2} > \frac{1}{1-\gamma^2} \quad \Longleftrightarrow \quad 2\gamma > 1 \quad \Longleftrightarrow \quad \gamma > \tfrac{1}{2}.
$$

Thus:

$$
\begin{cases}
\gamma < \frac{1}{2}: & \pi_{\text{left}} \text{ yields a higher return (move left each step)}, \\
\gamma = \frac{1}{2}: & v_{\pi_{\text{left}}} = v_{\pi_{\text{right}}} \text{ (both equal)}, \\
\gamma > \frac{1}{2}: & \pi_{\text{right}} \text{ yields a higher return (move right each step)}.
\end{cases}
$$

$$
\boxed{
\begin{aligned}
&\gamma = 0 && \Rightarrow \pi_{\text{left}}, \\
&\gamma = 0.5 && \Rightarrow \text{either } \pi_{\text{left}} \text{ or } \pi_{\text{right}}, \\
&\gamma = 0.9 && \Rightarrow \pi_{\text{right}}.
\end{aligned}
}
$$

When the discount factor γ is small, only immediate rewards matter, so moving left (which immediately gives 1) is optimal. As γ increases, future rewards become more important. Moving right initially gives 0 but leads to a delayed reward of 2, and as this future reward is discounted less, the right policy becomes better. At $\gamma = 0.5$, both policies yield the same total discounted return.

2 Value Estimation in Grid Worlds

2.1 Gridworld: Getting started

Code is in gridworld.py

2.2 Implement return computation and value estimation

(a) We tried the following values of k (number of episodes) with default discount factor $\gamma = 0.9$ to estimate the value of the start state under the random policy.

k	**Mean Return**	**Std. Dev.**
1	0.000003	0.000000
100	0.003590	0.018780
500	0.002594	0.013022
1000	0.002057	0.011245
5000	0.002195	0.012753
10000	0.002214	0.013063

(b) Using the sample size formula with 95% confidence and margin of error $E = \pm 0.0004$:

$$n = \left(\frac{z \times \sigma}{E} \right)^2$$

For 95% confidence, $z = 1.96$ which is obtained from the standard normal distribution table where the cumulative probability is 0.975 (leaving 2.5% in each tail) With $\sigma = 0.01306$ and $E = 0.0004$:

$$n = \left(\frac{1.96 \times 0.01306}{0.0004} \right)^2 = 4093$$

$n = 4093$ episodes are required.

(c) For DiscountGrid with $\gamma = 0.95$, running $10,000$ episodes with a random agent produced mean return$= -6.394126$ and standard deviation: 3.783941

To estimate the mean within ± 0.05 with 95% confidence, the required sample size is:

$$n = \left(\frac{1.96 \times 3.783941}{0.05} \right)^2 \approx 21995$$

At least $n = 21,995$ episodes are required.

(d) For $n = 500$ episodes, the 95% confidence interval is:

$$CI = 1.96 \times \frac{3.783941}{\sqrt{500}} \approx 1.96 \times 0.1693 \approx 0.33$$

This is much larger than the required interval of ± 0.1.

Answer: No, with 500 episodes, the confidence interval is approximately ± 0.33, so the value estimate does not predict the long-term average within ± 0.1.