Reinforcement Learning HW 1

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

October 20, 2025

1 Optimal Policy - Small Example Solution

The deterministic nature transforms the Bellman equation as follows:

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t \mid S_t = s \right]$$

We have the following sequence of rewards for the left policy:

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

with odd steps.

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

Similarly, the right policy has the following sequence of rewards with the even-indexed steps.

$$\text{Sequence: } [0, 2, 0, 2, \dots]$$

$$v_{\pi_{\text{right}}}(s) = G_t \Big|_{\pi=\text{right}} = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \Big|_{\pi=\text{right}} = \sum_{k=0}^{\infty} 2\gamma^{2k} = \frac{2}{1-\gamma^2}$$

$$\text{Since } \frac{2}{1-\gamma^2} > \frac{\gamma}{1-\gamma^2} \text{ for all } \gamma \in [0,1], \text{ policy } \pi_{\text{right}} \text{ is optimal.}$$