Reinforcement Learning, WS 24/25 – Exercise 01

Stephan Amann, Anna Schäfer, Tina Truong

# 1 Optimal Policy – small Example

Consider the system shown in Figure 1. The only decision to be made is that in the top state, where two actions are available, *left* and *right*. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, $\pi_{\text{left}}$ and $\pi_{\text{right}}$. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$?. *Hint:*[1]

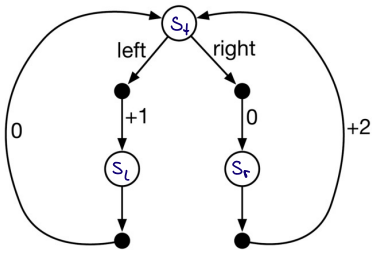


Figure 1: Simple system with one decision point.

There are 3 possible initial states per policy: $s_t$, $s_r$, $s_l$

For $\gamma = 0$ the action of the initial state is the final reward: $G_\pi(s) = R_s$

Case I: $G_{\pi_{left}}(s_{top}) = 1$ $\quad$ $G_{\pi_{right}}(s_{top}) = 0$

Case II: $G_{\pi_{left}}(s_{left}) = 0$ $\quad$ $G_{\pi_{right}}(s_{right}) = 2$

Case III: $G_{\pi_{left}}(s_{right}) = 2$ $\quad$ $G_{\pi_{right}}(s_{left}) = 0$

=> Depending on the initial state both policies can be optimal for $\gamma = 0$.

To get the reward for $\gamma = 0.9$ and $\gamma = 0.5$ we have to derive the geometric series, depending on the reward $R_0$ initial state and the second reward $R_1$ of the following state.

For Case I/II: $G_{cu/2} = R_0 + \gamma R_1 + \gamma^2 R_0 + \gamma^3 R_1 + \dots$

$$= \sum_{i=0}^{\infty} \left[ \gamma^{2i} R_0 + \gamma^{2i+1} R_1 \right]$$

$$= \sum_{i=0}^{\infty} \gamma^{2i} R_0 + \sum_{i=0}^{\infty} \gamma^{2i+1} R_1$$

$$\overset{i \to \infty}{=} \frac{R_0}{1-\gamma^2} + \frac{\gamma R_1}{1-\gamma^2}$$

$$\overset{i \to \infty}{=} \frac{R_0 + \gamma R_1}{1-\gamma^2}$$

Case III is special, as the initial state is the "opposite" state to the ones covered by the policy, i.e. for $\pi_{left}$ the start is at $s_{right}$ and vice versa. This means $R_{opp}$ occurs just once at the beginning followed by $R_{top}$ and the continuous alternation between $R_{ps}$ (reward for side the policy defines) and $R_{top}$:

$$
\begin{aligned}
G_{C3} &= R_{opp} + \gamma R_{top} + \gamma^2 R_{ps} + \gamma^3 R_{top} + \dots \\
&= R_{opp} + \sum_{i=0}^{\infty} \gamma^{2i+1} R_{top} + \sum_{i=0}^{\infty} \gamma^{2i+2} R_{ps} \\
&\overset{i\to\infty}{=} R_{opp} + \frac{\gamma R_{top}}{1-\gamma^2} + \frac{\gamma^2 R_{ps}}{1-\gamma^2} \\
&\overset{i\to\infty}{=} R_{opp} + \frac{\gamma (R_{top} + \gamma R_{ps})}{1-\gamma^2}
\end{aligned}
$$

Plug in the numbers for $\gamma = 0.9$:

**Case I:**

$$G_{\pi_{left}}(s_{top}) = \frac{1 + 0.9 \cdot 0}{1 - 0.9^2} = \frac{1}{0.19} = 5.263$$

$$G_{\pi_{right}}(s_{top}) = \frac{0 + 0.9 \cdot 2}{1 - 0.9^2} = \frac{1.8}{0.19} = 9.473$$

**Case II:**

$$G_{\pi_{left}}(s_{left}) = \frac{0 + 0.9 \cdot 1}{1 - 0.9^2} = \frac{0.9}{0.19} = 4.736$$

$$G_{\pi_{right}}(s_{right}) = \frac{2 + 0.9 \cdot 0}{1 - 0.9^2} = \frac{2}{0.19} = 10.526$$

**Case III:**

$$
\begin{aligned}
G_{\pi_{left}}(s_{right}) &= R_{right} + \frac{\gamma (R_{top} + \gamma R_{left})}{1-\gamma^2} \\
&= 2 + \frac{0.9(1 + 0.9 \cdot 0)}{1 - 0.9^2} = 2 + \frac{0.9}{0.19} = 6.736
\end{aligned}
$$

$$
\begin{aligned}
G_{\pi_{right}}(s_{left}) &= R_{left} + \frac{\gamma (R_{top} + \gamma R_{right})}{1-\gamma^2} \\
&= 0 + \frac{0.9(0 + 0.9 \cdot 2)}{1 - 0.9^2} = \frac{0.9 \cdot 1.8}{0.19} = 8.526
\end{aligned}
$$

$\Rightarrow$ In all cases $\pi_{right}$ is the best policy for $\gamma = 0.9$.

And for $\gamma = 0.5$:

**Case I:**

$$G_{\pi_{left}}(s_{top}) = \frac{1 + 0.5 \cdot 0}{1 - 0.5^2} = \frac{1}{0.75} = 1.33$$

$$G_{\pi_{right}}(s_{top}) = \frac{0 + 0.5 \cdot 2}{1 - 0.5^2} = \frac{1}{0.75} = 1.33$$

**Case II:**

$$G_{\pi_{left}}(s_{left}) = \frac{0 + 0.5 \cdot 1}{1 - 0.5^2} = \frac{0.5}{0.75} = 0.66$$

$$G_{\pi_{right}}(s_{right}) = \frac{2 + 0.5 \cdot 0}{1 - 0.5^2} = \frac{2}{0.75} = 2.66$$

**Case III:**

$$G_{\pi_{left}}(s_{right}) = R_{right} + \frac{\gamma (R_{top} + \gamma R_{left})}{1 - \gamma^2}$$

$$= 2 + \frac{0.5 (1 + 0.5 \cdot 0)}{1 - 0.5^2} = 2 + \frac{0.5}{0.75} = 2.66$$

$$G_{\pi_{right}}(s_{left}) = R_{left} + \frac{\gamma (R_{top} + \gamma R_{right})}{1 - \gamma^2}$$

$$= 0 + \frac{0.5 (0 + 0.5 \cdot 2)}{1 - 0.5^2} = \frac{0.5}{0.75} = 0.66$$

=> Depending on the case both policies can be optimal. Most interesting is case I, where no optimal policy can be determined, as the reward is the same.

# 2 Value Estimation in Grid Worlds

## 2.2 Implement return computation and value estimation

### (a)

These findings were found by using a discount factor of $\gamma = 0.9$ and varying episode size ranging from 10, 100, 1000, 10000 and 50000.

The average of rewards over all episodes (sample mean) helps us understand the long-term level of rewards given a particular start state.

The standard deviation helps us assess the confidence we can place in the sampled mean value. Typically, the larger the standard deviation, the more episodes are needed to be confident about the sampled mean.

| Episodes | Mean | Standard Deviation |
|---|---|---|
| 10 | 0.00047327 | 0.00119203 |
| 100 | 0.00178144 | 0.01251740 |
| 1000 | 0.00177891 | 0.00934980 |
| 10000 | 0.00253486 | 0.01380255 |
| 50000 | 0.00230841 | 0.01323271 |

### (b)

Since the true distribution is unknown, we estimate the population standard deviation using a sample standard deviation. When the sample size is sufficiently large, the sample mean and sample standard deviation can closely approximate the population mean and population standard deviation, respectively, due to the Central Limit Theorem.

For this purpose, we computed the standard deviation for a large amount of episodes ($n = 50,000$) and used this as an estimate for the population standard deviation, $\sigma'$. Given the allowed margin of error $E = 0.004$ and a z-score of $z = 1.96$ (for 95% confidence), we can use the following formula to determine the number of episodes needed:

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

Using this formula, we find that the number of episodes needed to estimate the population mean return is approximately 4410, with a margin of error of 0.0004. Running 4410 episodes yielded the sampled mean of 0.00232128. Which is within the allowed margin of error.

$$|\overline{X}_{50,000} - \overline{X}_{4410}| = |0.00230841 - 0.00232128| \approx 0 < 0.004$$

**(c)**

As in (b). Only now we use a discount factor of $\gamma = 0.95$ and a margin of error of $E = 0.05$.

For 50,000 episodes, the mean value was found to be -6.40656438. Each episode deviates from the mean by an average of 3.79408146.
To estimate the population mean return with a margin of error of 0.05, approximately 22,120 episodes are needed. For 22,120 episodes, the sampled mean was found to be -6.38895148. Which is within the allowed margin of error.

$$|\overline{X}_{50,000} - \overline{X}_{22,120}| = |-6.40656438 + 6.38895148| \approx 0.018 < 0.05$$

**(d)**

???TODO So, for the DiscountedGrid with a discount factor of $\gamma = 0.95$, we need at least 22,120 episodes to be confident that the sampled mean is accurate within an interval of 0.1. Therefore, 500 episodes are not sufficient to ensure this level of confidence. ???TODO