EE473 Deep Reinforcement Learning Homework 2

Author: Jingkun (Allen) Liu

Excercise 3.7

Since that no *discount rate* γ is used in this problem, so that the maximum return $G_t = \sum_{k=0}^{\infty} R_{t+k+1}$ would always be $+1$. In order to solve this problem, this system need to add a discount γ.

Excercise 3.10

Proof:

The equation for G_t can be rewritten as:

$$\begin{aligned} G_t &= \lim_{n\to\infty} \sum_{k=0}^{n} \gamma^k \\ &= \lim_{n\to\infty} S_n \\ S_n &= \sum_{k=0}^{n} \gamma^k \\ &= 1 + \gamma + \gamma^2 + \cdots + \gamma^n \end{aligned}$$

Then γS_n can be calculated as:

$$\begin{aligned} \gamma S_n &= \gamma \left(1 + \gamma + \gamma^2 + \cdots + \gamma^n\right) \\ &= \gamma + \gamma^2 + \gamma^3 + \cdots + \gamma^{n+1} \end{aligned}$$

By substracting from (1) to (3) we get

$$\begin{aligned} (1-\gamma)\, S_n &= \left(1 + \gamma + \gamma^2 + \cdots + \gamma^n\right) - \left(\gamma + \gamma^2 + \cdots + \gamma^{n+1}\right) \\ &= 1 - \gamma^{n+1} \\ S_n &= \frac{1-\gamma^{n+1}}{1-\gamma} \end{aligned}$$

Since we know that $|\gamma| < 1$

$$\begin{aligned} G_t &= \lim_{n\to\infty} S_n \\ &= \lim_{n\to\infty} \frac{1-\gamma^{n+1}}{1-\gamma} \\ &= \frac{1}{1-\gamma} \end{aligned}$$

Excercise 3.20

The optimal state-value function would be a function that combines $v_{\texttt{putt}}(s)$ and $q_*\left(s, \texttt{driver}\right)$ that

$$v_*(s) = \begin{cases} q_*(s, \texttt{driver}) & \text{Outside green and sand} \\ v_{\texttt{putt}}(s) & \text{On green or sand} \end{cases}$$

Excercise 4.3

$$\begin{aligned}
q_\pi\left(s,a\right) &= \mathbb{E}_\pi\left[G_t|S_t=s,A_t=a\right]\\
&= \mathbb{E}_\pi\left[R_{t+1}+\gamma G_{t+1}|S_t=s,A_t=a\right]\\
&= \mathbb{E}_\pi\left[R_{t+1}+\gamma\sum_{s',a'}q_\pi\left(s',a'\right)|S_t=s,A_t=a\right]\\
&= \sum_{s',r}p\left(s',r|s,a\right)\left[r+\gamma\sum_{a'}\pi(a'|s')q_\pi(s',a')\right]
\end{aligned}$$

$$\begin{aligned}
q_{k+1}\left(s,a\right) &= \mathbb{E}_\pi\left[R_{t+1}+\gamma G_{t+1}|S_t=s,A_t=a\right]\\
&= \sum_{s',r}p\left(s',r|s,a\right)\left[r+\gamma\sum_{a'}\pi\left(a'|s'\right)q_k\left(s',a'\right)\right]
\end{aligned}$$

Risk Aversion and Loss Aversion

Risk Aversion

The term **Risk Aversion** is defined as tendency of people to prefer outcome with low uncertainty to one with high uncertainty even the monetary value is higher in more uncertain outcome. This would change the way in optimization since traditional optimization method it to find the optimal outcome from all possible approaches. But in this situation, people favor the outcome with less variation rather than the one with the most significant value. So the optimization process would take the variation into account rather than only looking for the final outcome.

Loss Aversion

The term **Loss Aversion** is defined as losses are more sensitive to people's response rather than the gains acquired. This would impact the optimization method by changing the reward function in taking account more about people's response rather than purely looking for the gain acquired.