Review Session

Thursday 24 January 2019

Index

Markov Decision Process

- MDPs are defined by a quintuple $(\mathcal{S}, \mathcal{A}, r, P(\cdot|\cdot,\cdot), \gamma)$
- Objective: Find a stationary policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the sum of cumulative rewards.
- Value of a state given a policy: sum of cumulative rewards, given that the initial state is this state.

$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t), s_{t+1}) | s_0 = s\right] \\
&= \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) \left[r(s, \pi(s), s') + \gamma V^{\pi}(s')\right] \\
&= r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) V^{\pi}(s')
\end{aligned}
$$

- This equality is called bellman equation. It can used to evaluate the value of a state given a policy.
- What happens when $\gamma = 1$?

Optimality in MDPs

Bellman Optimality Theorem

- A policy π is optimal $\iff$ it is greedy with respect to its own value function:

$$V^\star(s) = V^{\pi^\star}(s) = \max_{a \in \mathcal{A}} \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^{\pi^\star}(s') \right]$$

$$\pi^\star(s) = \arg\max_{a \in \mathcal{A}} \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^{\pi^\star}(s') \right]$$

How To Solve MDPs

Policy Iteration

1. Start with a policy $\pi_0(\cdot)$.
2. Evaluate the policy $V^{\pi_0}(\cdot)$.
3. Optimize π as
   $\pi_k = \arg\max_{a \in \mathcal{A}} \left[r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V^{\pi_{k-1}}(s') \right]$.

Value Iteration

1. Start with a value $V_0(\cdot)$.
2. Optimize V as
   $V_k = \max_{a \in \mathcal{A}} \left[r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V_{k-1}(s') \right]$.
3. Recover the optimal policy upon convergence.

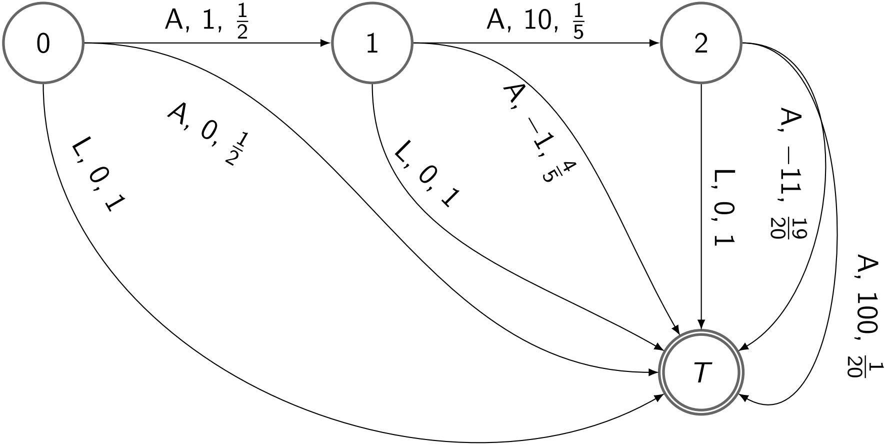
Index

Exercise

- The participant has at the beginning of each question the option to answer the next question, or leave.
- If she decides to leave, the reward is 0 and leaves with the total money she has until know.
- If she decides to answer, she can answer correctly and obtain money that accumulates in her pot, else she leaves the game and loses all money.
- There are three questions in sequence worth 1CHF, 10CHF, and 100CHF. The probability of answering correctly are 0.5, 0.2, and 0.05 respectively.

MDP Scheme

$\mathcal{S} = \{0, 1, 2, T\}$, $\mathcal{A} = \{A, L\}$. Arrows have $(a, r, P(s'|s, a))$.



Value Iteration

Initialization: $V_0(s) = 0$.

$$V_1(0) = \max_{a \in L,A} \left[0; \frac{1}{2}(1 + V_0(1)) + \frac{1}{2}(0+0)\right] = \max_{a \in L,A} \left[0; \frac{1}{2}\right] = 1$$

$$V_1(1) = \max_{a \in L,A} \left[0; \frac{1}{5}(10 + V_0(2)) + \frac{4}{5}(-1+0)\right] = \max_{a \in L,A} \left[0; \frac{6}{5}\right] = \frac{6}{5}$$

$$V_1(2) = \max_{a \in L,A} \left[0; \frac{1}{20}(100 + 0) + \frac{19}{20}(-11+0)\right] = \max_{a \in L,A} \left[0; -\frac{109}{100}\right] = 0$$

$$V_2(0) = \max_{a \in L,A} \left[0; \frac{1}{2}(1 + V_1(1)) + \frac{1}{2}(0+0)\right] = \max_{a \in L,A} \left[0; \frac{11}{10}\right] = \frac{11}{10}$$

$$V_2(1) = \max_{a \in L,A} \left[0; \frac{1}{5}(10 + V_1(2)) + \frac{4}{5}(-1+0)\right] = \max_{a \in L,A} \left[0; \frac{6}{5}\right] = \frac{6}{5}$$

$$V_2(2) = \max_{a \in L,A} \left[0; \frac{1}{20}(100 + 0) + \frac{19}{20}(-11+0)\right] = \max_{a \in L,A} \left[0; -\frac{109}{100}\right] = 0$$

Value Iteration

$$V_3(0) = \max_{a \in L,A} \left[0; \frac{1}{2}(1 + V_2(1)) + \frac{1}{2}(0 + 0)\right] = \max_{a \in L,A} \left[0; \frac{11}{10}\right] = \frac{11}{10}$$

$$V_3(1) = \max_{a \in L,A} \left[0; \frac{1}{5}(10 + V_2(2)) + \frac{4}{5}(-1 + 0)\right] = \max_{a \in L,A} \left[0; \frac{6}{5}\right] = \frac{6}{5}$$

$$V_3(2) = \max_{a \in L,A} \left[0; \frac{1}{20}(100 + 0) + \frac{19}{20}(-11 + 0)\right] = \max_{a \in L,A} \left[0; -\frac{109}{100}\right] = 0$$

Converged (this is rare $\gamma = 1$) as $V_2(s) = V_3(s)$. Policy:

$$\pi(0) = \arg\max_{a \in L,A} \left[0; \frac{1}{2}(1 + V_3(1)) + \frac{1}{2}(0 + 0)\right] = A$$

$$\pi(1) = \arg\max_{a \in L,A} \left[0; \frac{1}{5}(10 + V_3(2)) + \frac{4}{5}(-1 + 0)\right] = A$$

$$\pi(2) = \arg\max_{a \in L,A} \left[0; \frac{1}{20}(100 + 0) + \frac{19}{20}(-11 + 0)\right] = L$$

Policy Iteration

Initialization: $\pi_0(s) = A$. Evaluation:

$$V^{\pi_0}(0) = \frac{1}{2}(1 + V^{\pi_0}(1)) + \frac{1}{2}(0 + 0) = \frac{11}{200}$$
$$V^{\pi_0}(1) = \frac{1}{5}(10 + V^{\pi_0}(2)) + \frac{4}{5}(-1 + 0) = \frac{11}{100}$$
$$V^{\pi_0}(2) = \frac{1}{20}(100 + 0) + \frac{19}{20}(-11 + 0) = \frac{-109}{20}$$

Optimization:

$$\pi_1(0) = \arg\max_{a \in L,A} \left[0, \frac{11}{200}\right] = A$$
$$\pi_1(1) = \arg\max_{a \in L,A} \left[0, \frac{11}{100}\right] = A$$
$$\pi_2(2) = \arg\max_{a \in L,A} \left[0, \frac{-109}{20}\right] = L$$

Policy Iteration

Evaluation:

$$
\begin{aligned}
V^{\pi_1}(0) &= \frac{1}{2}(1 + V^{\pi_1}(1)) + \frac{1}{2}(0 + 0) = \frac{11}{10} \\
V^{\pi_1}(1) &= \frac{1}{5}(10 + V^{\pi_1}(2)) + \frac{4}{5}(-1 + 0) = \frac{6}{5} \\
V^{\pi_1}(2) &= 0
\end{aligned}
$$

Optimization:

$$
\begin{aligned}
\pi_2(0) &= \arg\max_{a \in L,A} \left[0, \frac{11}{10}\right] = A \\
\pi_2(1) &= \arg\max_{a \in L,A} \left[0, \frac{6}{5}\right] = A \\
\pi_2(2) &= \arg\max_{a \in L,A} \left[0, \frac{-109}{20}\right] = L
\end{aligned}
$$

Converged!