UCLRL Lecture 2 Notes

1 Markov Decision Processes

1.1 Markov Reward Processes

Definition 1. *A Mariv Process is a tuple* $\langle \mathcal{S}, \mathcal{P} \rangle$

For a Markov state s and successor state s', the state transition probability is defined by

$$\mathcal{P}_{ss'} = \mathbb{P}\left[S_{t+1} = s' | S_t = s\right]$$

and we can characterize the transition from all states by the transition matrix $\mathcal{P}$ where

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_{11} & \dots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{n1} & \dots & \mathcal{P}_{nn} \end{bmatrix}$$

Definition 2. *A Markov reward process is a Markov process but with a tuple* $\langle \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ *such that*

- $\mathcal{R}$ *is a reward function such that* $\mathcal{R}_s = \mathbb{E}\left[R_{t+1} | S_t = s\right]$
- γ *is a discount factor with* $\gamma \in [0, 1]$

Definition 3. *The **return** G_t is the total discounted reward from time-step t*

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Definition 4 (State Value Function)**.** *The state value function $v(s)$ of an MRP is the expected return starting from state s*

$$v(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

The value function be decomposed into

- immediate reward R_{t+1}
- discounted value of successor state $\gamma v(S_{t+1}$

$$\begin{aligned} v(s) &= \mathbb{E}\left[G_t | S_t = s\right] \\ &= \mathbb{E}\left[R_{t+1} + \gamma G_{t+1}\right] \\ &= \mathbb{E}\left[R_{t+1} + \gamma v(S_{t+1}\right] \end{aligned}$$

Definition 5 (Bellman Equation for MRP).

$$v(s) = \mathbb{E}\left[R_{t+1} + \gamma v(S_{t+1}|S_t = s\right]$$

which can be rewritten as

$$v(s) = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} P_{ss'} v(s')$$

or in matrix notation

$$v = \mathcal{R} + \gamma \mathcal{P} v$$

Bellman equation can be solved direcly as

$$\begin{aligned} v &= \mathcal{R} + \gamma \mathcal{P} v \\ &= (I - \gamma \mathcal{P})^{-1} \mathcal{R} \end{aligned}$$

1.2 Markov Decision Process

Definition 6 (Markov Decision Process). *A Markov Decision Process is a Markov Reward Process with a finite set of actions $\mathcal{A}$. Thus, the Markov Decision process is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ such that*

1. *$\mathcal{A}$ is a finite set of actions*
2. *$\mathcal{P}^a_{ss'} = \mathbb{P}\left[S_{t+1} = s' | S_t = s, A_t = a\right]$*
3. *$\mathcal{R}$ is a reward function, $\mathcal{R}^a_s = \mathbb{E}\left[R_{t+1} | S_t = s, A_t = a\right]$*

Definition 7 (Policy). *A policy π is a distribution over actions given states,*

$$\pi(a|s) = \mathbb{P}\left[A_t = a | S_t = s\right]$$

Definition 8 (State Value Function). *The state value function $v_\pi(s)$ of an MMDFP is the expected return starting from state s and then following policy π*

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

Definition 9 (Action Value Function). *The action value function $q_\pi(s, a)$ is the expected return staring from state s, taking action a, and then following policy π*

$$q_\pi(s, a) = \mathbb{E}_\pi\left[G_t | S_t = s, A_t = a\right]$$

Definition 10 (Bellman Expectation Equation). *Bellman expectation eqiation be expressed for the Markov reward process as*

$$v_\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v_\pi$$

which gives us the solution after solving for v_π,

$$v_\pi = \left(I - \gamma \mathcal{P}^\pi\right)^{-1} \mathcal{R}^\pi$$

Definition 11. *The optimal state value function $v_o(s)$ is the maximum value function over all policies*

$$v_\star(s) = \max_\pi v_\pi(s)$$

Definition 12. *The optimal action-value function $q_o(s,a)$ is the maximum action-value function over all policies*

$$q_\star(s,a) = \max_\pi q_\pi(s,a)$$

Definition 13 (Partial Ordering of Policies)**.**

$$\pi \geq \pi' v_\pi(s) \geq v'_\pi(s), \forall s$$

Theorem 1. *For any Markov decision process*

1. *There exists an optimal policy $\pi_\star$ that is better than or equal to all other policies $\pi_\star \geq \pi, \forall \pi$*
2. *All optimal policies achieve the optimal value function*

$$v_{\pi_\star}(s) = v_\star(s)$$

3. *All optimal policies achive the optimal action-value function*

$$q_{\pi_\star}(s,a) = q_\star(s,a)$$

An optimal policy can be found by maximizing over $q_\star(s,a)$,

$$\pi_\star(a|s) = \begin{cases} 1 & \text{if} a = \arg\max_{a \in \mathcal{A}} q_\star(s,a) \\ 0 & \text{otherwise} \end{cases}$$

2 Bellman Equations

For $v_\star$ we look at the action that gives us the most value,

$$v_\star(s) = \max_a q_\star(s,a)$$

and for $q_\star$, we have the immediate reward and the average of all the states and their values,

$$q_\star(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' in \mathcal{S}} \mathcal{P}_{ss'}^a v_\star(s')$$

and combining them together,

$$v_\star(s) = \max_a \left[\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v_\star(s') \right]$$

and similarly for $q_\star$,

$$q_\star(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' in \mathcal{S}} \mathcal{P}_{ss'}^a \max_{a'} q_\star(s',a')$$