MARKOV DECISION PROCESS NOTES

CARSON JAMES

Contents

1. Introduction

1.1. Setup.

Definition 1.1. *Let $\mathcal{S}$, $\mathcal{A}$ be finite sets, $(S_t)_{t=0}^{\infty}$ a sequence of S-valued random variables and $(A_t)_{t=0}^{\infty}$ a sequence of $\mathcal{A}$-valued random variables. For each $s \in \mathcal{S}$, let $\mathcal{A}_s \subset \mathcal{A}$. For each $s \in \mathcal{S}$ and $a \in \mathcal{A}_s$, let $P_s^a : \mathcal{S} \to [0,1]$ given by $P_s^a(s') \mapsto P_{s,s'}^a$ define a probabilty measure on $\mathcal{S}$, that is:*

$$\sum_{s' \in \mathcal{S}} P_{s,s'}^a = 1$$

Suppose that $(S_t)_{t=0}^{\infty}$ satisfies the following markov property: For each $t \in \mathbb{N}_0$, $s', s, s_{t-1}, \cdots, s_0 \in \mathcal{S}$ and $a \in \mathcal{A}_s, a_{t-1} \in \mathcal{A}_{s_{t-1}}, \cdots, a_0 \in \mathcal{A}_{s_0}$,

$$\begin{aligned}
&\mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a, S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \cdots, S_0 = s_0, A_0 = a_0) \\
&\quad = \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a) \\
&\quad = P_{s,s'}^a
\end{aligned}$$

*For each $t \in \mathbb{N}_0$, let $v_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ and $\pi_t : \mathcal{S} \to \mathcal{A}$ where for each $s \in \mathcal{S}$, $\pi_t(s) \in \mathcal{A}_s$. Define $\pi = (\pi_t)_{t=0}^{\infty}$. Then the tuple $(\mathcal{S}, \mathcal{A}, P, v, \pi)$ is said to be a **Markov decision process** or MDP with states $\mathcal{S}$, actions $\mathcal{A}$, transition probabilites $P_{s,s'}^a$, values $(v_t)_{t=0}^{\infty}$ and policy π.*

Note 1.2. *The values $(v_t)_{t=0}^{\infty}$ are typically called the "reward" or "cost" values which has to do with whether the goal is to maximize or minimize total cost. These notes maintain generality and will simply refer to $(v_t)_{t=0}^{\infty}$ as the values. We do the same with the total cost function defined below.*

2. Finite Horizon

2.1. Values.

Note 2.1. *In the following section, we will consider a MDP over the time horizon of* $t = 0, 1, \cdots, n$. *The time* $t = n$ *is terminal and therefore no action will be chosen at this time. Therefore we may assume that the value at this time will only depend on the state and we denote this value by* $v_n : \mathcal{S} \to \mathbb{R}$. *In addition, we may assume that* $\pi = (\pi_t)_{t=0}^{n-1}$.

Definition 2.2. *Let* π *be a policy. We define the* ***expected total value function*** *for the policy* π, $V^\pi : \mathcal{S} \to \mathbb{R}$, *by*

$$V^\pi(s) = \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{n-1} v_t(S_t, \pi_t(S_t), S_{t+1}) + v_n(S_n)|S_0 = s\right]$$

and we define the ***expected total discounted value function*** *for the policy* π *by*

$$V_\gamma^\pi(s) = \mathop{\mathbb{E}}_{\pi}\left[\sum_{t=0}^{n-1} \gamma^t v_t(S_t, \pi_t(S_t), S_{t+1}) + \gamma^n v_n(S_n)|S_0 = s\right]$$

Theorem 2.3. ***The Bellman Equation****: Define*

$$V_t^\pi(s) = \begin{cases} v_n(s) & t = n \\ \sum\limits_{s' \in \mathcal{S}} P_{s,s'}^{\pi_t(s)}\left[v_t(s, \pi_t(s), s') + V_{t+1}(s')\right] & 0 \le t < n \end{cases}$$

and define

$$V_{\gamma,t}^\pi(s) = \begin{cases} v_n(s) & t = n \\ \sum\limits_{s' \in \mathcal{S}} P_{s,s'}^{\pi_t(s)}\left[v_t(s, \pi_t(s), s') + \gamma V_{t+1}(s')\right] & 0 \le t < n \end{cases}$$

Then $V^\pi(s) = V_0^\pi(s)$ *and* $V_\gamma^\pi(s) = V_{\gamma,0}^\pi(s)$.

2.2. Optimization.

Definition 2.4. *Let* Π *be the set of all policies and* $\pi^* \in \Pi$. *Then* π^* *is said to be a* ***maximal policy*** *if*

$$V^{\pi^*}(s) = \sup_{\pi \in \Pi} V^\pi(s)$$

If we are discounting, then π^* *is said to be a* ***maximal discounted policy*** *if*

$$V_\gamma^{\pi^*}(s) = \sup_{\pi \in \Pi} V_\gamma^\pi(s)$$

Note 2.5. *We can define a "minimal policy" and a "minimal discounted policy" similarly.*

Theorem 2.6. ***Bellman Optimality Equation***

(1) If π^* is a maximal policy, then

$$V_t^{\pi^*}(s) = \begin{cases} v_n(s) & t = n \\ \max\limits_{a \in \mathcal{A}_s} \sum\limits_{s' \in \mathcal{S}} P_{s,s'}^a \left[v_t(s,a,s') + V_{t+1}^{\pi^*}(s') \right] & t = 0, 1, \cdots, n-1 \end{cases}$$

and

$$\pi_t^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}_s} \sum_{s' \in \mathcal{S}} P_{s,s'}^a \left[v_t(s,a,s') + V_{t+1}^{\pi^*}(s') \right]$$

(2) If π^* is a maximal discounted policy, then

$$V_{\gamma,t}^{\pi^*}(s) = \begin{cases} v_n(s) & t = n \\ \max\limits_{a \in \mathcal{A}_s} \sum\limits_{s' \in \mathcal{S}} P_{s,s'}^a \left[v_t(s,a,s') + \gamma V_{t+1}^{\pi^*}(s') \right] & t = 0, 1, \cdots, n-1 \end{cases}$$

and

$$\pi_t^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}_s} \sum_{s' \in \mathcal{S}} P_{s,s'}^a \left[v_t(s,a,s') + \gamma V_{t+1}^{\pi^*}(s') \right]$$

Note 2.7. *A similar result holds for minimal policies and minimal discounted policies.*

Definition 2.8. *The process of computing π^* via the Bellman optimality equation is known as* ***value iteration****. It is a dynamic programming algorithm.*

2.3. **Examples.**

Example 2.9. ***Inventory Control:*** *A company is moving product down the supply chain over the times $t = 0, 1, 2, 3$. At the beginning of each time interval $t = 0, 1, 2$ the company*

(1) has an inventory s between -2 to 2 units of product (negative in the case of a backlog of customer orders)
(2) orders (and immediately receives) a units of product based on the inventory s
(3) sells product to customers according to a random demand d experienced during time t. The maximum demand at time t is 2 units of product.

The demand at time t is given by

$$P(d = 0) = .1$$
$$P(d = 1) = .6$$
$$P(d = 2) = .3$$

The cost to

(1) order one unit of product is 1 .
(2) hold one unit of product after demand is 2.
(3) hold one unit of backlogged product is 3

Suppose that the terminal cost for each possible amount of inventory is 0. Find the minimal expected total cost given that we start with no inventory and a minimal policy for ordering inventory.

Solution 2.10. *We have* $\mathcal{S} = \{-2, -2, 0, 1, 2\}$, $\mathcal{A}_s = \{0, 1, \cdots, 2 - s\}$, $v_t(s, a, s') = a + 2\max(0, s') + 3\max(0, -s')$ *and* $v_3(s) = 0$. *The transition probabilites, organized by action, are given by:*

$$P^0_{-2,-2} = 1, P^0_{-2,-1} = 0, P^0_{-2,0} = 0, P^0_{-2,1} = 0, P^0_{-2,2} = 0$$
$$P^0_{-1,-2} = .9, P^0_{-1,-1} = .1, P^0_{-1,0} = 0, P^0_{-1,1} = 0, P^0_{-1,2} = 0$$
$$P^0_{0,-2} = .3, P^0_{0,-1} = .6, P^0_{0,0} = .1, P^0_{0,1} = 0, P^0_{0,2} = 0$$
$$P^0_{1,-2} = 0, P^0_{1,-1} = .3, P^0_{1,0} = .6, P^0_{1,1} = .1, P^0_{1,2} = 0$$
$$P^0_{2,-2} = 0, P^0_{2,-1} = 0, P^0_{2,0} = .3, P^0_{2,1} = .6, P^0_{2,2} = .1$$

$$P^1_{-2,-2} = .9, P^1_{-2,-1} = .1, P^1_{-2,0} = 0, P^1_{-2,1} = 0, P^1_{-2,2} = 0$$
$$P^1_{-1,-2} = .3, P^1_{-1,-1} = .6, P^1_{-1,0} = .1, P^1_{-1,1} = 0, P^1_{-1,2} = 0$$
$$P^1_{0,-2} = 0, P^1_{0,-1} = .3, P^1_{0,0} = .6, P^1_{0,1} = .1, P^1_{0,2} = 0$$
$$P^1_{1,-2} = 0, P^1_{1,-1} = 0, P^1_{1,0} = .3, P^1_{1,1} = .6, P^1_{1,2} = .1$$

$$P^2_{-2,-2} = .3, P^2_{-2,-1} = .6, P^2_{-2,0} = .1, P^2_{-2,1} = 0, P^2_{-2,2} = 0$$
$$P^2_{-1,-2} = 0, P^2_{-1,-1} = .3, P^2_{-1,0} = .6, P^2_{-1,1} = .1, P^2_{-1,2} = 0$$
$$P^2_{0,-2} = 0, P^2_{0,-1} = 0, P^2_{0,0} = .3, P^2_{0,1} = .6, P^2_{0,2} = .1$$

$$P^3_{-2,-2} = 0, P^3_{-2,-1} = .3, P^3_{-2,0} = .6, P^3_{-2,1} = .1, P^3_{-2,2} = 0$$
$$P^3_{-1,-2} = 0, P^3_{-1,-1} = 0, P^3_{-1,0} = .3, P^3_{-1,1} = .6, P^3_{-1,2} = .1$$

$$P^4_{-2,-2} = 0, P^4_{-2,-1} = 0, P^4_{-2,0} = .3, P^4_{-2,1} = .6, P^4_{-2,2} = .1$$

Since our initial inventory is 0*, we wish to find the minimal policy* π^* *and* $V_0^{\pi^*}(0)$*. Now, using the bellman optimality equation, we know that the maximal policy*

$$\pi_2^*(s) = \underset{a \in \mathcal{A}_s}{\operatorname{argmin}} \sum_{s' \in \mathcal{S}} P^a_{s,s'} \Big[a + 2\max(0, s') + 3\max(0, -s') \Big]$$

and

$$V_2^{\pi^*}(s) = \min_{a \in \mathcal{A}_s} \sum_{s' \in \mathcal{S}} P^a_{s,s'} \Big[a + 2\max(0, s') + 3\max(0, -s') \Big]$$

$$j$$

For $s = -2$*,*