Markov Decision Processes

Guillaume Barnier

Academic Year 2020-2021

Contents

1 Markov Process

Definition 1.1 (Markov Property)**.** Let $S_t = (s_0, s_1, ...)$ be a stochastic process evolving according to a transition dynamic P. This stochastic process satisfies the Markov property if

$$P(s_t|s_0, s_1, ..., s_{t-1}) = P(s_t|s_{t-1}),\ \forall t \in \mathbb{N} \tag{1}$$

Definition 1.2 (Markov Process)**.** A Markov Process (MP) is a stochastic process that satifies the Markov property

In a RL setting, we often make two additional assumptions:

- **Finite state space**. The state space of the Markov process is finite. This means that for the Markov process $(s_0, s_1, ...)$, there is a state space S with $|S| < \infty$, such that for all realizations of the Markov process, we have $s_t \in S$ for all t.
- **Stationary transition probability**. The transition probabilities are time independent:

$$P(s_p = s'|s_{p-1} = s) = P(s_q = s'|s_{q-1} = s),\ \forall (p, q) \tag{2}$$

A Markov process satisfying these assumptions is also sometimes called a Markov chain, although the precise definition of a Markov chain varies. With these assumptions, we can define characterize a Markov process with the following definition.

Definition 1.3 (Markov Process/Markov Chain)**.** A Markov Process is a tuple $\big(S, P\big)$, where

- S is the finite state-space of the Markov process, $|S| < \infty$
- P is the state transition probability model where $P_{ss'} = P[s_{t+1} = s'|s_t = s]$

Lemma 1.1. $P(s_{t+n}|s_t = s) = P(s_n|s_0 = s)$ for all t and n

Proof. I show this property by recursion on n:

- For $n = 1$, $P(s_{t+1}|s_t = s) = P(s_1|s_0 = s)$ is true by the stationarity assumption
- I assume that $P(s_{t+n}|s_t = s) = P(s_n|s_0 = s)$ is true for n

- I show that $P(s_{t+n+1}|s_t = s) = P(s_{n+1}|s_0 = s)$:

$$\begin{aligned} P(s_{t+n+1}|s_t = s) &= \sum_{s'} P(s_{t+n+1}, s_{t+n} = s'|s_t = s) && (3)\\ &= \sum_{s'} P(s_{t+n+1}|s_t = s, s_{t+n} = s')P(s_{t+n} = s'|s_t = s) && (4)\\ &= \sum_{s'} P(s_{t+n+1}|s_{t+n} = s')P(s_n = s'|s_0 = s) && (5)\\ &= \sum_{s'} P(s_{n+1}|s_n = s')P(s_n = s'|s_0 = s) && (6)\\ &= \sum_{s'} P(s_{n+1}, s_n = s'|s_0 = s) && (7)\\ &= P(s_{n+1}|s_0 = s) && (8) \end{aligned}$$

2 Markov Reward Process

Definition 2.1 (Markov Reward Process)**.** A Markov Reward Process (MRP) is a tuple $\big(S, P, R, \gamma\big)$, where

- S is the finite state-space of the Markov process, (assume $|S| < \infty$)
- P is the state transition probability model where $P_{ss'} = P[s_{t+1} = s'|s_t = s]$
- $R : S \mapsto \mathbb{R}$ is a reward function that maps states to rewards, $R(s) = E[r_t|s_t = s]$
- $\gamma \in [0, 1]$ is a discount factor

In a Markov reward process, whenever a transition happens from a current state s to a successor state s', a reward is obtained depending on the current state s. Thus for the Markov process $(s_0, s_1, ...)$, each transition $s_t \to s_{t+1}$ is accompanied by a reward r_t for all $i = 0, 1, ...$, and so a particular episode of the Markov reward process is represented as $(s_0, r_0, s_1, r_1, s_2, r_2, ...)$. We should note that these rewards can be either deterministic or stochastic.

Definition 2.2 (Expected reward)**.** For a state $s \in S$, we define the expected reward $R(s)$ by

$$R(s) = E[r_0|s = s_0] \tag{9}$$

Just like the assumption of stationary transition probabilities, going forward we will also assume *stationarity of the rewards*. In the deterministic case, this implies that $r_i = r_j$ wherever $s_i = s_j$. In the stochastic case, we require that the cumulative distribution functions (CDF) of the rewards conditioned on the current state be time independent:

$$F(r_i|s_i = s) = F(r_j|s_j = s), \tag{10}$$

where F denotes the cumulative distribution function of r_i conditioned on s_i. Therefore, the reward function $R(s)$ is independent of t,

$$R(s) = E(r_t|s_t = s) \tag{11}$$

Definition 2.3 (Horizon)**.** The horizon H of a Markov reward process is defined as the number of time steps in each episode (realization) of the process. The horizon can be finite or infinite. If the horizon is finite, then the process is also called a finite Markov reward process.

Definition 2.4 (Return)**.** The return G_t of a Markov reward process is defined as the discounted sum of rewards starting at time t up to the horizon H, and is given by

$$G_t = \sum_{k=t}^{H-1} \gamma^{k-t} r_k, \ for\ t \in [0, H-1] \tag{12}$$

For example, $G_0 = r_0 + \gamma r_1 + \gamma^2 r_2 + ... + \gamma^{H-1} r_{H-1}$

Definition 2.5 (State value function)**.** The state value function $V_t(s)$ for a Markov reward process and a state $s \in S$ is defined as the expected return starting from state s at time t, and is given by the following expression:

$$V_t(s) = E[G_t | s_t = s] \tag{13}$$

Lemma 2.1. Let us assume that

- Transition probability is stationary
- Rewards are stationary
- H is infinite.

Then $V_t(s)$ is independent of t. That is,

$$V_i(s) = V_j(s),\ \forall i, j \tag{14}$$

Proof. Even though this property seems obvious and intuitive, the proof is not totally straightforward (at least to me). I conduct the demonstration in two steps:

(1) I prove that $E[r_{t+n}|s_t = s] = E[r_n|s_0 = s]$ by recursion on n:

- For $n = 0$, I use the result proved in lemma 1.1 to show that

$$\begin{aligned} E(r_t|s_t = s) &= \sum_{r\in\mathbb{R}} r\ p(r_t = r|s_t = s) && (15)\\ &= \sum_{r\in\mathbb{R}} r\ p(r_0 = r|s_0 = s) && (16)\\ &= E(r_0|s_0 = s) && (17) \end{aligned}$$

- I assume $E[r_{t+n}|s_t = s] = E[r_n|s_0 = s]$ for all n. Then, I show that $E[r_{t+n+1}|s_t = s] = E[r_{n+1}|s_0 = s]$.

$$\begin{aligned} E\big[r_{t+n+1}|s_t = s\big] &= E\Big[E\big[r_{t+n+1}|s_t = s, s_{t+n+1} = s'\big]|s_t = s\Big] && (18)\\ &= E\Big[E\big[r_{t+n+1}|s_{t+n+1} = s'\big]|s_t = s\Big] && (19)\\ &= \sum_{s'} E\big[r_{t+n+1}|s_{t+n+1} = s'\big] P(s_{t+n+1} = s'|s_t = s) && (20)\\ &= \sum_{s'} E\big[r_{t+n+1}|s_{t+n+1} = s'\big] P(s_{n+1} = s'|s_0 = s) && (21)\\ &= \sum_{s'} E\big[r_{n+1}|s_{n+1} = s'\big] P(s_{n+1} = s'|s_0 = s) && (22)\\ &= E\big[r_{n+1}|s_0 = s\big] && (23) \end{aligned}$$

However, I think equation 18 does not come from the law of iterated expectation. That is what was confusing me (I saw this written everywhere). In fact, I think it comes from the tower property for two random variables, which states that

$$E\Big[E\big[X|Y, Z\big]|Y\Big] = E\big[X|Y\big], \tag{24}$$

whereas the law of iterated expectation is

$$E\Big[E\big[X|Z\big]\Big] = E\big[X\big]. \tag{25}$$

(2) Finally, I conclude that $V_{t+n}(s) = V_t(s)$

$$\begin{aligned} V_{t+n}(s) &= E\Big[G_{t+n}|s_{t+n} = s\Big] && (26)\\ &= E\Big[\sum_{k=0}^{\infty}\gamma^k r_{t+n+k}\Big|s_{t+n} = s\Big] && (27)\\ &= \sum_{k=0}^{\infty}\gamma^k E\big[r_{t+n+k}\Big|s_{t+n} = s\big] && (28)\\ &= \sum_{k=0}^{\infty}\gamma^k E\big[r_{t+k}\Big|s_t = s\big] && (29)\\ &= V_t(s) && (30) \end{aligned}$$

3 Other useful tricks

if $X \geq 0$

$$E[X] = \int_0^{\infty} 1 - F_X(x)dx \tag{31}$$

4 Acknowledgments

I would like to specifically thank former CS 234 Teacher Assistant Rahul Sarkar for providing his quality notes that greatly helped me understand the material, and also inspired me to write my own notes. I used his work as a reference, and most of the changes made focused on deriving detailed proofs that were not always provided in his notes.