Markov Decision Processes

Guillaume Barnier

Academic Year 2020-2021

Contents

1 Markov Process

1.1 Definitions

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 satisfies 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 induction 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 $P(s_{t+n+1}|s_t = s) = P(s_{n+1}|s_0 = s)$:

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

Remark:

- Equation 3 is obtained by using the fact that for X, Y, Z random variables,

$$P(X|Y) = \sum_z z\ P(X, Z = z|Y) \tag{9}$$

- Equation 5 is obtained by using the Markov property and the induction assumption
- Equation 7 is obtained by using the Markov property again followed by Bayes rule

$$\begin{aligned} P(s_{n+1}|s_n = s')P(s_n = s'|s_0 = s) &= P(s_{n+1}|s_n = s', s_0 = s)P(s_n = s'|s_0 = s) && (10)\\ &= P(s_{n+1}, s_n = s'|s_0 = s) && (11) \end{aligned}$$

2 Markov Reward Process

2.1 Definitions

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 $n = |S| < \infty$)
- 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

$$R(s) = E[r_0|s = s_0] \tag{12}$$

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:

$$F(r_i|s_i = s) = F(r_j|s_j = s), \tag{13}$$

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

$$P(r_{t+p}|s_{t+p} = s) = P(r_t|s_t = s) \quad (14)$$

$$R(s) = E(r_t|s_t = s) \quad (15)$$

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

$$G_t = \sum_{k=t}^{H-1} \gamma^{k-t} r_k, \ for\ t \in [0, H-1] \quad (16)$$

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:

$$V_t(s) = E[G_t|s_t = s], \quad (17)$$

and can be interpreted as the long-term value of state s.

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,

$$V_t(s) = V(s) \quad (18)$$

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 the rewards' stationarity assumption to show that

$$E(r_t|s_t = s) = \sum_r r\ p(r_t = r|s_t = s) \quad (19)$$

$$= \sum_r r\ p(r_0 = r|s_0 = s) \quad (20)$$

$$= E(r_0|s_0 = s) \quad (21)$$

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

$$E\big[r_{t+n+1}|s_t = s\big] = E\Big[E\big[r_{t+n+1}|s_t = s, s_{t+n+1} = s'\big]|s_t = s\Big] \quad (22)$$

$$= E\Big[E\big[r_{t+n+1}|s_{t+n+1} = s'\big]|s_t = s\Big] \quad (23)$$

$$= \sum_{s'} E\big[r_{t+n+1}|s_{t+n+1} = s'\big] P(s_{t+n+1} = s'|s_t = s) \quad (24)$$

$$= \sum_{s'} E\big[r_{t+n+1}|s_{t+n+1} = s'\big] P(s_{n+1} = s'|s_0 = s) \quad (25)$$

$$= \sum_{s'} E\big[r_{n+1}|s_{n+1} = s'\big] P(s_{n+1} = s'|s_0 = s) \quad (26)$$

$$= E\big[r_{n+1}|s_0 = s\big] \quad (27)$$

Equation 22 does not come from the law of iterated expectation (as most of the papers/proofs I have seen), but rather from one form of the **tower property** for random variables, which states that

$$E\Big[E\big[X|Y,Z\big]|Y\Big] = E\big[X|Y\big], \tag{28}$$

whereas the law of iterated expectation is

$$E\Big[E\big[X|Z\big]\Big] = E\big[X\big]. \tag{29}$$

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

$$\begin{aligned}
V_{t+n}(s) &= E\Big[G_{t+n}|s_{t+n}=s\Big] && (30)\\
&= E\Big[\sum_{k=0}^{\infty}\gamma^k r_{t+n+k}\Big|s_{t+n}=s\Big] && (31)\\
&= \sum_{k=0}^{\infty}\gamma^k E\big[r_{t+n+k}\Big|s_{t+n}=s\big] && (32)\\
&= \sum_{k=0}^{\infty}\gamma^k E\big[r_{t+k}\Big|s_t=s\big] && (33)\\
&= V_t(s) && (34)
\end{aligned}$$

2.2 Bellman Equations

For an **infinite horizon** MRP, the value function $V_t(s)$ can be decomposed into two parts: (1) an immediate reward r_t, and (2) a discounted value of successor state $\gamma V_{t+1}(s')$:

$$V(s) = E[r_t + \gamma V_{t+1}(s')|s_t = s] \tag{35}$$

$$V(s) = R(s) + \gamma\sum_{s'} V(s')P_{ss'} \tag{36}$$

Proof.

$$\begin{aligned}
V(s) = V_t(s) &= E[G_t|s_t=s] && (37)\\
&= E[r_t + \gamma G_{t+1}|s_t=s] && (38)\\
&= R(s) + \gamma E[G_{t+1}|s_t=s] && (39)\\
&= R(s) + \gamma E\Big[E\big[G_{t+1}|s_t=s, s_{t+1}=s'\big]|s_t=s\Big] && (40)\\
&= R(s) + \gamma E\Big[E\big[G_{t+1}|s_{t+1}=s'\big]|s_t=s\Big] && (41)\\
&= R(s) + E\Big[\gamma V_{t+1}(s')|s_t=s\Big] && (42)\\
&= R(s) + \gamma\sum_{s'} V_{t+1}(s')P(s_{t+1}=s'|s_t=s) && (43)\\
&= R(s) + \gamma\sum_{s'} V(s')P_{ss'} && (44)
\end{aligned}$$

Remarks on the proof:

- We used the fact that $G_t = r_t + \gamma G_{t+1}$

- Equation 40 is obtained by using the tower property (equation 28)
- We used the fact that $V_t(s) = V_{t+1}(s) = V(s)$ (Lemma 2.1)

Additionally, for $n = |S| < \infty$, equation 36 can be written as linear system of equations,

$$\mathbf{V} \quad = \quad \mathbf{R} + \gamma \mathbf{V}\mathbf{P}, \tag{45}$$

where $\mathbf{V} \in \mathbb{R}^n$ and $\mathbf{R} \in \mathbb{R}^n$ are the value-function and expected rewards vectors, respectively. $\mathbf{P} \in \mathbb{R}^{n \times n}$ is the transition probability matrix, where $\mathbf{P}_{i,j} = P(s_{t+1} = s_j | s_t = s_i)$. Equation can also be written as
which can also be written as

$$\begin{bmatrix} v(s_1) \\ \vdots \\ v(s_n) \end{bmatrix} = \begin{bmatrix} R_{s_1} \\ \vdots \\ R_{s_n} \end{bmatrix} + \gamma \begin{bmatrix} P_{s_1 s_1} & \dots & P_{s_1 s_n} \\ \vdots & \ddots & \vdots \\ P_{s_n s_1} & \dots & P_{s_n s_n} \end{bmatrix} \begin{bmatrix} v(s_1) \\ \vdots \\ v(s_n) \end{bmatrix}. \tag{46}$$

For $\gamma < 1$, $(\mathbf{I} - \gamma \mathbf{P})$ is invertible and Equation 45 yields the following analytical solution

$$\mathbf{V} \quad = \quad (\mathbf{I} - \gamma \mathbf{P})^{-1} \mathbf{R}. \tag{47}$$

Proof. We show that for $0 \leq \gamma < 1$, $(\mathbf{I} - \gamma \mathbf{P})$ is invertible.

1. We first prove that $\mathbf{P}$ has an eigenvalue equal to 1
2. We show that any eigenvalue λ of $\mathbf{P}$ is such that $|\lambda| < 1$
3. We conclude that $\mathbf{I} - \gamma \mathbf{P}$ is invertible

1. $\mathbf{P}$ is a row stochastic matrix: $\sum_{j=1}^{|S|} P_{ij} = 1$, and $\lambda = 1$ is an eigenvalue of $\mathbf{P}$ with a corresponding eigenvector $\mathbf{v} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

$$\mathbf{P} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{48}$$

2. Let λ be an eigenvalue of $\mathbf{P}$ with a corresponding eigenvector $\mathbf{v}$. Then, $\mathbf{P}\mathbf{v} = \lambda \mathbf{v}$. By examining the i^{th} row, we can write

$$\sum_{j=1}^{|S|} P_{ij} v_j = \lambda v_i \tag{49}$$

Let $|v_k| = \max_q |v_q|$. Since $\mathbf{v}$ is an eigenvector, we have $|v_k| > 0$. Therefore

$$\begin{aligned} |\lambda||v_k| &= |\sum_{j=1}^{|S|} P_{ij} v_j| && (50) \\ &\leq \sum_{j=1}^{|S|} P_{ij} |v_j| && (51) \\ &\leq |v_k| \sum_{j=1}^{|S|} P_{ij} && (52) \\ &= |v_k| && (53) \end{aligned}$$

Therefore, $|\lambda| \leq 1$ 3. Let us assume that $\lambda = 0$ is an eigenvalue of $\mathbf{I} - \gamma\mathbf{P}$. Then, there exists $\mathbf{v} \neq \mathbf{0}$ such that

$$(\mathbf{I} - \gamma\mathbf{P})\mathbf{v} = \mathbf{0} \tag{54}$$
$$\gamma\mathbf{P}\mathbf{v} = \mathbf{v} \tag{55}$$

and thus $\frac{1}{\gamma} > 1$ is an eigenvalue of $\mathbf{P}$, which contradicts our previous result.

3 Markov Decision Process

3.1 Definitions

Definition 3.1 (**Markov Decision Process**). A Markov Decision Process (MDP) is a tuple $\big(S, A, P, R, \gamma\big)$, where

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

The basic model of the dynamics is that there is a state space S, and an action space A, both of which we will consider to be finite. The agent starts from a state s_t at time t, chooses an action a_t from the action space, obtains a reward r_t and then reaches a successor state s_{t+1}. An episode of a MDP is thus represented as $(s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, ...)$.

Unlike in the case of a Markov Process or a Markov Reward Process where the transition probability was only a function of the successor state and the current state, the transition probabilities for a MDP at time t are a function of the successor state s_{t+1} along with both the current state s_t and the action a_t, written as $P^a_{ss'} = P(s_{t+1} = s' | s_t = s, a_t = a)$. We still assume the principle of stationary transition probabilities which in the context of a MDP is written mathematically as

$$P^a_{ss'} = P(s_{i+1} = s' | s_i = s, a_i = a) = P(s_{j+1} = s' | s_j = s, a_j = a) \tag{56}$$

Additionally, the reward r_t at time t depends on both s_t and s_t, in contrast to a Markov Reward Process where it depended only on the current state. These rewards can be stochastic or deterministic, but just like in the case of a Markov reward process, we will assume that the rewards are stationary and the only relevant quantity will be the expected reward which we will denote by $R(s, a)$ for a fixed state s and action a, and defined below:

$$R(s, a) = E[r_t | s_t = s, a_t = a] \tag{57}$$

Definition 3.2 (**Policy for MDPs**). A policy specifies what action to take in each state of a MDP and fully defines the behavior of an agent. Policies can either be deterministic or stochastic. To cover both these cases, we will consider a policy to be a probability distribution over actions given the current state:

$$\pi_t(a|s) = P(a_t = a | s_t = s) \tag{58}$$

The policy may be varying with time, which is especially true in the case of finite horizon MDPs. We will denote a generic policy by π, defined as the infinite dimensional tuple $\pi = (\pi_0, \pi_1, ...)$, where π_t refers to the policy at time t. We will call policies that do not vary with time "stationary policies", and indicate them as π, i.e. in this case $\pi = (\pi, \pi, ...)$.

Given a MDP $M = \big(S, A, P, R, \gamma\big)$ and a policy π:

- The state sequence $(s_0, s_1, ...)$ is a Markov Process (S, P^π)

- The state/reward sequence $(s_0, r_0, s_1, r_1, ...)$ is a Markov reward process $(S, R^\pi, P^\pi, \gamma)$

Additionally, the transition probability matrix and the reward functions are given by

$$P^\pi_{ss'} = \sum_{a \in A} P(s_{t+1} = s' | s_t = s, a_t = a) \pi(a_t = a | s_t = s) \tag{59}$$

$$P^\pi_{ss'} = \sum_{a \in A} P(s_{t+1} = s' | s_t = s, a_t = a) \pi(a_t = a | s_t = s) \tag{60}$$

Definition 3.3 (**State value function for a MDP**)**.** The state-value function $V^\pi(s)$ of a MDP is the expected return starting from state s, and then following policy π

$$V^\pi_t(s) = E_\pi[G_t | s_t = s], \tag{61}$$

where E_π denotes the expected value of a random variable given that the agent follows policy π. The value of the terminal state (if any) is always zero.

Definition 3.4 (**State-action value function for a MDP**)**.** The action-value function Q^π of an MDP is the expected return starting from state s, taking action a, and then following policy π

$$Q^\pi_t(s) = E_\pi[G_t | s_t = s, a_t = a], \tag{62}$$