

Reinforcement Learning

Ash Bellett

Contents

1	Markov processes	1
1.1	Markov process	1
1.2	Markov reward process	1
1.3	Markov decision process	2
2	Model-Based Policy Optimisation	4
2.1	Policy Evaluation	4
2.2	Policy Iteration	4
2.3	Value Iteration	4
3	Model-Free Policy Evaluation	4
3.1	Monte Carlo Learning	4
3.2	Temporal Difference Learning	4
4	Model-Free Policy Optimisation	4
5	Value Approximation	4
6	Policy Gradient Methods	4
7	Actor-Critic Methods	4

1 Markov processes

1.1 Markov process

A Markov process is described by:

- a finite set of states \mathcal{S}
- a state transition probability matrix $P_{s,s'}$ where $s, s' \in \mathcal{S}$

A sequence of random variables $\{S_t \in \mathcal{S}\}$ follow the Markov property: the next state S_{t+1} depends only on the current state S_t :

$$P(S_{t+1} = s' | S_t = s, S_{t-1} = s_{t-1}, S_{t-2} = s_{t-2}, \dots) = P(S_{t+1} = s' | S_t = s)$$

The state transition probability matrix $P_{s,s'}$ gives the probability of transitioning from state s to state s' :

$$P_{s,s'} = P(S_{t+1} = s' | S_t = s)$$

1.2 Markov reward process

A Markov reward process is described by:

- a finite set of states \mathcal{S}
- a finite set of rewards $\mathcal{R} \subset \mathbb{R}$
- a state transition probability matrix $P_{s,s'}$
- a reward function $r : \mathcal{S} \rightarrow \mathcal{R}$
- a discount factor $\gamma \in [0, 1]$

The reward function r maps a state $s \in \mathcal{S}$ to a reward $R_t \in \mathcal{R}$:

$$R_t = r(s | S_t = s)$$

The next expected reward R_s at state s is defined as:

$$\begin{aligned} R_s &= E(R_{t+1} | S_t = s) \\ &= \sum_{s' \in \mathcal{S}} r(s') P_{s,s'} \end{aligned}$$

The return $G_t \in \mathbb{R}$ is a sum of attenuated future rewards represented as a geometric series:

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \\ &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \end{aligned}$$

The return can be rewritten as:

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \dots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

The state value function $v : S \rightarrow \mathbb{R}$ is the expected return at state s :

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

From the law of total expectation, $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$:

$$\begin{aligned} \mathbb{E}(G_{t+1}) &= \mathbb{E}(\mathbb{E}(G_{t+1} | S_{t+1} = s')) \\ &= \mathbb{E}(v(s')) \end{aligned}$$

The Bellman expectation equation for the state value function $v(s)$ is:

$$\begin{aligned} v(s) &= \mathbb{E}(R_{t+1} + \gamma v(s') | S_t = s) \\ &= \sum_{s' \in S} P_{s,s'} r(s') + \gamma \sum_{s' \in S} P_{s,s'} v(s') \\ &= R_s + \gamma \sum_{s' \in S} P_{s,s'} v(s') \end{aligned}$$

1.3 Markov decision process

A Markov decision process is described by:

- a finite set of states \mathcal{S}
- a finite set of rewards $\mathcal{R} \subset \mathbb{R}$
- a finite set of actions \mathcal{A}
- a state transition probability matrix $P_{s,s'}^a$
- a policy distribution $\pi \in [0, 1]$
- a reward function $r : \mathcal{S} \rightarrow \mathcal{R}$
- a discount factor $\gamma \in [0, 1]$

The state transition probability matrix $P_{s,s'}^a$ gives the probabilities of transitioning from state s to state s' if action $a \in \mathcal{A}$ is taken:

$$P_{s,s'}^a = \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a)$$

The policy π maps a state s to the probability of selecting action a :

$$\pi(a|s) = \mathbb{P}(A_t = a | S_t = s)$$

The action A_t is sampled from the policy:

$$A_t \sim \pi(\cdot | S_t = s)$$

While following policy π , the state transition probability matrix $P_{s,s'}^\pi$ is:

$$P_{s,s'}^\pi = \sum_{a \in \mathcal{A}} \pi(a|s) P_{s,s'}^a$$

The next expected reward R_s^a at state s given action a is taken is defined as:

$$\begin{aligned} R_s^a &= \mathbb{E}(R_{t+1} | S_t = s, A_t = a) \\ &= \sum_{s' \in \mathcal{S}} r(s') P_{s,s'}^a \end{aligned}$$

While following policy π , the next expected reward R_s^π is:

$$R_s^\pi = \sum_{a \in \mathcal{A}} \pi(a|s) R_s^a$$

The state value function $v_\pi : \mathcal{S} \rightarrow \mathbb{R}$ is the expected return at state s under policy π :

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi(G_t | S_t = s) \\ &= \mathbb{E}_\pi(R_{t+1} + \gamma G_{t+1} | S_t = s) \end{aligned}$$

From the law of total expectation, $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$:

$$\begin{aligned} \mathbb{E}_\pi(G_{t+1}) &= \mathbb{E}_\pi(\mathbb{E}_\pi(G_{t+1} | S_{t+1} = s')) \\ &= \mathbb{E}_\pi(v_\pi(s')) \end{aligned}$$

The Bellman expectation equation for the state value function $v_\pi(s)$ under policy π is:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi(R_{t+1} + \gamma v_\pi(s') | S_t = s) \\ &= \mathbb{E}_\pi(R_{t+1} | S_t = s) + \gamma \mathbb{E}_\pi(v_\pi(s') | S_t = s) \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) R_s^a + \gamma \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} P_{s,s'}^a v_\pi(s') \\ &= R_s^\pi + \gamma \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} P_{s,s'}^a v_\pi(s') \end{aligned}$$

The action-value function $q_\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the expected return at state s , taking action a and following policy π :

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}_\pi(G_t | S_t = s, A_t = a) \\ &= \mathbb{E}_\pi(R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a) \end{aligned}$$

From the law of total expectation, $E(X) = E(E(X|Y, Z))$:

$$\begin{aligned} E_\pi(G_{t+1}) &= E_\pi(E_\pi(G_{t+1}|S_{t+1} = s', A_{t+1} = a')) \\ &= E_\pi(q_\pi(s', a')) \end{aligned}$$

The Bellman expectation equation for the action-state value function $q_\pi(s, a)$ under policy π is:

$$\begin{aligned} q_\pi(s, a) &= E_\pi(R_{t+1} + \gamma q_\pi(s', a') | S_t = s, A_t = a) \\ &= E_\pi(R_{t+1} | S_t = s, A_t = a) + \gamma E_\pi(q_\pi(s', a') | S_t = s, A_t = a) \\ &= R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{s,s'}^a \sum_{a' \in \mathcal{A}} \pi(a' | s') q_\pi(s', a') \end{aligned}$$

The state value function v_π can be written in terms of the action-state value function q_π :

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a | s) q_\pi(s, a)$$

Similarly, the action-state value function q_π can be written in terms of the state value function v_π :

$$q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{s,s'}^a v_\pi(s')$$

2 Model-Based Policy Optimisation

2.1 Policy Evaluation

2.2 Policy Iteration

2.3 Value Iteration

3 Model-Free Policy Evaluation

3.1 Monte Carlo Learning

3.2 Temporal Difference Learning

4 Model-Free Policy Optimisation

5 Value Approximation

6 Policy Gradient Methods

7 Actor-Critic Methods