Markov Decision Processes

Rupam Mahmood

January 24, 2020

Worksheet question 1

1. Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2, R_2 = -2, R_3 = 0$ followed by an infinite sequence of 7s. What are G_1 and G_0?

Worksheet question 2

2. Assume you have a bandit problem with 4 actions, where the agent can see rewards from the set $\mathcal{R} = \{-3.0, -0.1, 0, 4.2\}$. Assume you have the probabilities for rewards for each action: $p(r|a)$ for $a \in \{1, 2, 3, 4\}$ and $r \in \{-3.0, -0.1, 0, 4.2\}$. How can you write this problem as an MDP? Remember that an MDP consists of $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma)$.

   More abstractly, recall that a Bandit problem consists of a given action space $\mathcal{A} = \{1, ..., k\}$ (the k arms) and the distribution over rewards $p(r|a)$ for each action $a \in \mathcal{A}$. Specify an MDP that corresponds to this Bandit problem.

Worksheet question 3

3. Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $R_{\max} > 0$.

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty \qquad \text{for } \gamma \in [0, 1)$$

Hint: Recall that $|a + b| < |a| + |b|$.

The reward hypothesis

That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward).

The goal of a bandit agent

Maximize expected reward R

$$\pi(a) = P(A = a)$$

$$v_\pi = E_\pi[R] = E_\pi[E[R \mid A]] = E_\pi[q_*(A)]$$

Choose policy π that maximizes v_π

The goal of an agent

Contextual Bandits

Maximize expected reward *R* for all state *S*

$$\pi(a \mid s) = P(A = a \mid S = s)$$

$$v_\pi(s) = E_\pi[R \mid S = s] = E_\pi[E[R \mid S = s, A]] = E_\pi[q_*(s, A)]$$

Choose policy π that maximizes v_π for all state *S*

MDPs

Maximize expected sum of discounted future rewards *R* from all states *S*

Maximize expected return *G* from all states *S*

return: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

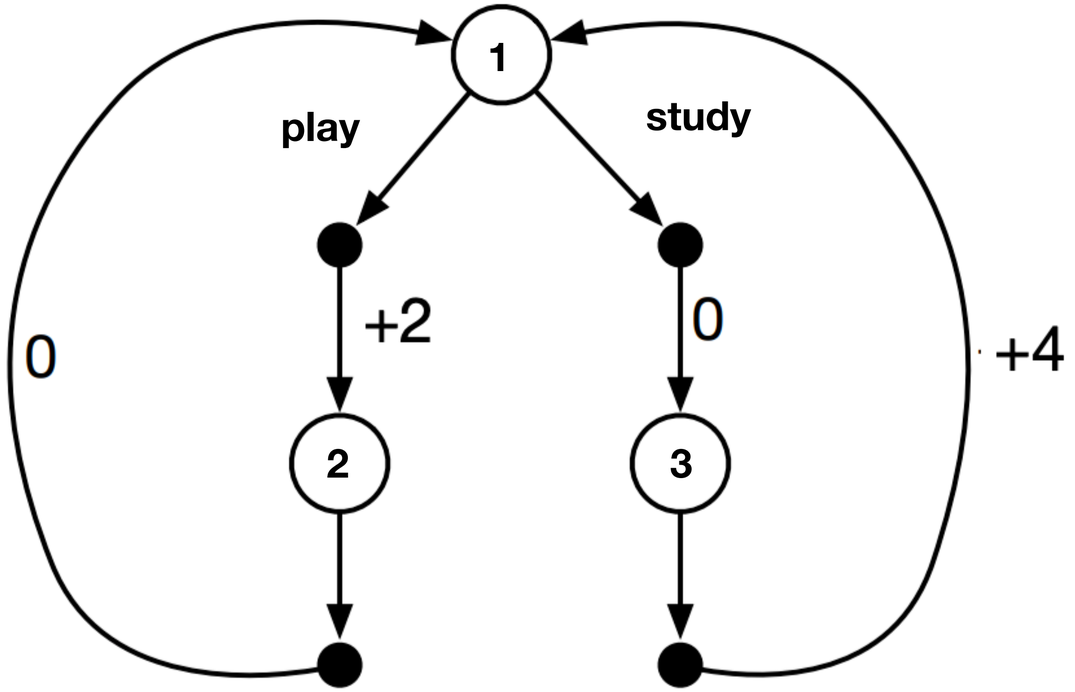
$$= R_{t+1} + \gamma G_{t+1}$$

$$v_\pi(s) = E_\pi[G_t \mid S_t = s]$$

Choose policy π that maximizes v_π for all state *S*

Policies

$$\pi(a \mid s) = P(A = a \mid S = s)$$



Expressing state-value functions & action-value functions

Contextual Bandits

$$v_\pi(s) = E_\pi[R \mid S = s]$$

$$= E_\pi[E[R \mid S = s, A]] = E_\pi[q_*(s, A)]$$

Law of the unconscious statistician: $E[\, g(X)\,] = \sum P(X{=}x)\, g(x)$

$$= \sum_a P(A_t = a \mid S_t = s) q_*(s, a)$$

$$= \sum_a \pi(a \mid s) q_*(s, a)$$

MDPs

state-value function:

$$v_\pi(s) = E_\pi[G_t \mid S_t = s]$$

$$= E_\pi[E_\pi[G_t \mid S_t = s, A_t]] \quad = E_\pi[q_\pi(s, A_t)]$$

$$= \sum_a P(A_t = a \mid S_t = s) q_\pi(s, a)$$

$$= \sum_a \pi(a \mid s) q_\pi(s, a)$$

action-value function:

$$q_\pi(s, a) = E_\pi[G_t \mid S_t = s, A_t = a]$$

The Bellman equation for v_π

return: $G_t = R_{t+1} + \gamma G_{t+1}$

state-value function:

$$v_\pi(s) = E_\pi[G_t \mid S_t = s] = \sum_s \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')]; \quad \textbf{for all } s$$