CMPUT 365: Introduction to Reinforcement Learning, Winter 2023
Worksheet #2: Markov Decision Processes

Manuscript version: *#947a9c-dirty* - 2023-04-05 12:30:13-06:00

Question 1. Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2, R_2 = -2, R_3 = 0$ followed by $R_4 = R_5 = \cdots = 7$, an infinite sequence of 7s. What are G_1 and G_0?

Question 2. Assume you have a bandit problem $B = (p_a)_{a\in[k]}$, where for each action $a \in [k]$, p_a is a pmf of rewards incurred when a is chosen.

1. Specify an MDP M by giving its state space, action space, reward space, transition probability function such that the following holds:

   > *Any bandit algorithm $\mathcal{B}$ gives rise to an MDP algorithm $\tilde{\mathcal{B}}$ such that for any $t \geq 0$, the distribution of $\tilde{\mathcal{B}}$'s (undiscounted) return in M after t steps is the same as the distribution of $\mathcal{B}$'s total reward in B after t steps, and the converse also holds: the distributions of the returns of any MDP algorithm is matched by that of a corresponding bandit algorithm.*

   (*)

2. Let the bandit algorithm $\mathcal{B}$ be $\mathcal{B} = (\pi_t)_{t\geq 0}$, where for $t \geq 0$, $\pi_t : ([k] \times \mathcal{R})^t \to \mathcal{M}_1([k])$. In particular, in time step $t \geq 0$, when the past action-reward sequence is $(a_0, r_0, \ldots, a_{t-1}, r_{t-1})$, $\mathcal{B}$ recommends using an action randomly sampled from $\pi_t(a_0, r_0, \ldots, a_{t-1}, r_{t-1})$.

   Describe how $\tilde{\mathcal{B}}$ works. How does $\tilde{\mathcal{B}}$ get action A_t when the history of its interaction with M is $S_0, A_0, R_1, S_1, A_1, \ldots, R_t, S_t$? How does it use π?

3. (hardness: *) Argue that (*) holds for $\tilde{\mathcal{B}}$ that you specified in the previous item.

4. (hardness: *) Show the converse reduction: Given an MDP algorithm for M, describe a corresponding bandit algorithm such that their reward distributions match (which you need to show formally).

5. In the above two-way reduction between bandits and MDPs, we used undiscounted, continuing MDPs. Yet, often people compare bandits to using a horizon of 1, or a discount factor of $\gamma = 0$. What is their argument? When a bandit is viewed as an MDP with discount factor $\gamma = 0$, what is that we can claim for the relationship between the corresponding problems?

Remember that to specify an MDP you need to state what the state space, the action space, the transition function is and how returns are calculated and whether the MDP is continuing or episodic.

Question 3. Fix $t \geq 0$ and let $R_1, R_2, \cdots \in [-\rho, \rho]$ for some $\rho \geq 0$. Let $T \in \{0, 1, \dots\} \cup \{\infty\}$. Recall that $G_t = \sum_{i=0}^{T-t-1} \gamma^i R_{t+1+i}$ where the value of the sum is defined to be zero if $T - t - 1 < 0$ ("empty sum evaluates to zero").

1. Assume that $T < \infty$. Show that $|G_t| < \infty$.
2. Assume now that $T = \infty$ (i.e., in the definition of G_t the upper limit of the sum is ∞). Show that it still holds that $|G_t| < \infty$.

Hint: Recall the triangle inequality.

Question 4 (From discounted to undiscounted problems with coin flips)**.** (**)

Fix $r_1, r_2, \cdots \in [-\rho, \rho]$ and $0 \le \gamma < 1$. Consider flipping biased coins where the probability of head is $1-\gamma$ until a head comes out. Let T be the time when we get the first head. Show that

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \sum_{t=0}^{\infty} \gamma^t r_{t+1}\,.$$

Note: The result of this exercise allows to transform discounted MDPs with a discount factor $0 \le \gamma < 1$ into an undiscounted episodic MDP so that under any causal action-selection method, the total expected discounted in the first MDP is the same as the total *undiscounted* return in the second MDP.

The transformation is as follows: Let M be a discrete MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, reward space $\mathcal{R}$ and transition function P. Without loss of generality assume that $0 \in \mathcal{R}$. Consider an MDP M' with state space $\mathcal{S}' = \mathcal{S} \cup \{\bot\}$ where $\bot \notin \mathcal{S}$, and action and reward spaces identical to the respective spaces of M. Let P' be the transition function for the new MDP so that

$$P'(s', r'|s, a) = \begin{cases} \gamma P(s', r'|s, a)\,, & \text{if } s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}, r' \in \mathcal{R}\,; \\ (1-\gamma) P(r'|s, a)\,, & \text{if } s \in \mathcal{S}, a \in \mathcal{A}, s' = \bot, r' \in \mathcal{R}\,; \\ 1, & \text{if } s = s' = \bot, a \in \mathcal{A}, r' = 0\,; \\ 0, & \text{otherwise}\,. \end{cases}$$

Here, $P(r'|s, a) = \sum_{s' \in \mathcal{S}} P(s', r'|s, a)$ is the probability of seeing r' when a is used in s.

In words, $\bot$ is a new terminal state: Upon reaching $\bot$, the state remains $\bot$ and no further rewards are received ($\bot$ is an *absorbing state*). In any state of the original MDP, together with generating a transition, a biased coin is flipped, independently of previous transitions and coin flips. The bias (probability of head) of the coin is $1-\gamma$. If the outcome is head, the next state is $\bot$, otherwise the next state is whatever was generated from P. The reward incurred is always as generated.

The way terminal states are treated here are a little different from what is in the SB20 book: The book avoids defining transition probabilities for terminal states, while here we make terminal states into absorbing ones. We do this because it is more convenient: We do not need to treat terminal states as special from the perspective of how transitions are defined.

Question 5. Let $r_1, r_2, \ldots$ be such that for $t \geq 1$, $|r_t| \leq \rho^t$ for some $\rho \geq 1$. Let $0 \leq \gamma < 1$. Show that

$$\sum_{t=0}^{\infty} \gamma^t r_{t+1}$$

is absolutely convergent if $\rho\gamma < 1$.

Question 6. Recall that by Eq. (3.5) from the book, $r(s,a) = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s',r|s,a)$ and by Eq. (3.6) of the book, $r(s,a,s') = \sum_{r\in\mathcal{R}} r \frac{p(s',r|s,a)}{p(s'|s,a)}$ for any $s' \in \mathcal{S}$ such that $p(s'|s,a) > 0$ (if $p(s'|s,a)$, we let $r(s,a,s')$ to be an arbitrary value). Show that

$$r(s,a) = \sum_{s'\in\mathcal{S}} p(s'|s,a) r(s,a,s') \,.$$

Question 7. Consider a three state, episodic MDP with state space $\mathcal{S} = \{s_1, s_2, \perp\}$, where the only terminal state of the MDP is $\perp$. The action space $\mathcal{A} = \{a_1, a_2\}$. From both states, action a_1 either leads to the terminal state with probability of β, or it leads to the same state where it was used. From s_2, action a_2 either leads to the terminal state with probability of α, or it leads s_2. From s_1, action a_2 either leads to s_2 with probability α, or it leads to s_1.

Describe $p(s'|s, a)$ for all combinations of (s, a, s') as a table (the columns should be s, a, s' and the corresponding probability $p(s'|s, a)$, the rows should be ordered lexicographically where s_1 precedes s_2, s_2 precedes $\perp$ and a_1 precedes a_2.

While the book avoids defining transition probabilities when s is a terminal state, in this exercise follow the convention that once the terminal state is reached, it cannot be escaped: the terminal state is *absorbing*. Include the full table: Your table should have $3 \times 2 \times 3 = 18$ rows.
