# Practice Midterm Solutions

EC 400

October 27, 2021

1. Imagine throwing a fair coin $n$ times to obtain a sequence of heads and tails. Let $p_n$ be the probability that the sequence $HTHTHTHT$ appears in the result. Argue that $\lim_{n\to\infty} p_n = 1$.

   **Solution:** The sequence $HTHTHTHT$ has length eight. The probability of generating this sequence on tosses $1, \ldots, 8$ is positive: it is $(1/2)^8$. The probability of generating it on tosses $9, \ldots, 16$ is the same. More generally, there is a positive probability $p = (1/2)^8$ of generating this sequence on any sequence of tosses of length of 8.

   The probability of failing to generate after $8k$ tosses is $(1-p)^k$. This goes to zero as $k \to +\infty$.

2. Consider an MDP with a single state and two actions. The first action gives you a reward of 2, while the second gives you a reward of 3. Compute the $Q$-values associated both actions under the optimal policy. The discount factor $\gamma$ should appear in your answer.

   **Solution:** First, observe that $J^*(1) = 3/(1-\gamma)$. Thus the Q-value of the first action is $2 + \gamma\frac{3}{1-\gamma}$ while the Q-value of the second action is $\frac{3}{1-\gamma}$.

3. Consider an MDP with three states, labeled $1, 2, 3$. In state $i$, you have the option to "stay" or "move right." In each state, staying leaves you at that state with probability 1. At one or two, "move right" brings you to, respectively, two and three, with probability $1/2$; and with probability $1/2$ you stay where you are. At three, "move right" keeps you at 3 with probability 1. The rewards to "move right" are always zero, while the rewards to "stay" are one, two, and four respectively. The terminal reward is zero.

   Compute the optimal policy using dynamic programming with $K = 4$ steps and $\gamma = 1$.

   **Solution:** We have

$$J_3 = \begin{pmatrix} \max(1 + 1\cdot 0, 0 + 1\cdot\left(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0\right) \\ \max(2 + 1\cdot 0, 0 + 1\cdot\frac{1}{2}0) \\ \max(4 + 1\cdot 0, 0 + 1\cdot 0) \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}$$

   Looking at which option achieves the max, we see that at step 3, the optimal policy is to "stay" at every node. Next,

$$J_2 = \begin{pmatrix} \max(1 + 1\cdot 1, 0 + 1\cdot\left(\frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 1\right) \\ \max(2 + 1\cdot 2, 0 + 1\cdot\left(\frac{1}{2}2 + \frac{1}{2}4\right) \\ \max(4 + 1\cdot 4, 0 + 1\cdot 4) \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 8 \end{pmatrix}$$

   This means that, at time 2, the optimal policy is to stay in every node. Finally,

$$J_1 = \begin{pmatrix} \max(1 + 1\cdot 2, 0 + 1\cdot\left(\frac{1}{2}2 + \frac{1}{2}\cdot 4\right) \\ \max(2 + 1\cdot 4, 0 + 1\cdot\left(\frac{1}{2}4 + \frac{1}{2}8\right) \\ \max(4 + 1\cdot 8, 0 + 1\cdot 8) \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \\ 12 \end{pmatrix}$$

   So, at time 1, you want to stay in state 3, but at states 1 and 2 you are indifferent between staying in moving: either choice is optimal.

4. Perform two steps step of policy evaluation on the policy which takes actions uniformly at random at every state on the same MDP. Initialize all the values to be zero. Leave the discount factor $\gamma$ in your final answer. You do not need to simplify: any expression equal to the correct answer will get full credit.

   We initialize at

$$J_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

   and obtain

$$J_1 = \begin{pmatrix} (1/2) \\ 1 \\ 2 \end{pmatrix}$$

   and

$$J_2 = \begin{pmatrix} (1/2) + \gamma((3/4)(1/2) + (1/4) \cdot 1) \\ 1 + \gamma((3/4)1 + (1/4)2) \\ 2 + \gamma 2 \end{pmatrix} = \begin{pmatrix} (1/2) + \gamma(5/8) \\ 1 + \gamma(5/4) \\ 2 + 2\gamma \end{pmatrix}$$

5. Perform three steps of Q-learning on the same MDP, assuming the state-action pairs generated are $s_1 = 1$,$a_1 =$"move right",$s_2 = 2$, $a_2$="stay,", $s_3 = 2$,$a_3$="stay," $s_4 = 2$. Initialize the Q-values at zero. Leave the discount factor $\gamma$ in your final answer.

   **Solution:**

$$Q_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$Q_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$Q_3 = \begin{pmatrix} 0 \\ 0 \\ 0 + \frac{1}{2}(2 + \gamma \cdot 0 - 0) \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$Q_4 = \begin{pmatrix} 0 \\ 0 \\ 1 + \frac{1}{3}(2 + \gamma 1 - 1) \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \frac{4+\gamma}{3} \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

6. Suppose $\gamma = 1$. Give an example of an MDP where the value of a node is infinite.

   **Solution:** A single MDP with a single action which gives a reward of one. With discount $\gamma = 1$, the value function is $1 + 1 + 1 + \cdots$.

7. True or false: if $J \neq T_\pi J$, then $J \neq J_\pi$.

   **Solution:** True. This is because $T_\pi$ is a $\gamma$-contraction, and has a unique fixed point (see lecture notes).

8. True or false: the optimal policy in a continuing reinforcement learning problem is unique.

   **Solution:** False. Indeed, consider an MDP with a single state and two actions, both of which give the same reward.

9. Suppose $T$ is an $\alpha$-contraction in the infinity norm. Argue that

$$||T^{k+1}x - T^{k+1}y||_\infty \leq \alpha ||T^k x - T^k y||_\infty$$

   **Solution:**

$$T^{k+1}x - T^{k+1}y = T(T^k x) - T(T^k y)$$

   so that, applying the contraction property,

$$||T^{k+1}x - T^{k+1}y||_\infty = ||T(T^k x) - T(T^k y)||_\infty \leq \alpha ||T^k x - T^k y||_\infty$$