

# COMPSCI 687 Homework 2

## (100 Points)

**Instructions:** Submit your answers as a single typed PDF file in Gradescope. You may use any software to create the PDF, and any reasonable formatting of your answers will suffice. You may collaborate with other students and may use generative AI to complete this assignment without reporting the collaboration or generative AI usage, but you are responsible for the accuracy of your answers.

### Part One: Written (75 Points)

- Given an MDP  $M = (\mathcal{S}, \mathcal{A}, p, R, d_0, \gamma)$  and a fixed policy,  $\pi$ , the probability that the action at time  $t = 0$  is  $a \in \mathcal{A}$  is:

$$\Pr(A_0 = a) = \sum_{s \in \mathcal{S}} d_0(s) \pi(s, a). \quad (1)$$

Write similar expressions (using only  $\mathcal{S}, \mathcal{A}, p, R, d_0, \gamma$ , and  $\pi$ ) for the following problems. Your final answer should not include Pr terms, e.g.,  $\Pr(S_3 = s)$ , although such terms will show up as you derive your final answer. Assume that every event on which you are asked to condition has positive probability.

**Example problem:** The probability that the state at time  $t = 1$  is  $s \in \mathcal{S}$ .

$$\Pr(S_1 = s) = \sum_{s_0 \in \mathcal{S}} \Pr(S_0 = s_0) \Pr(S_1 = s | S_0 = s_0) \quad (2)$$

$$= \sum_{s_0 \in \mathcal{S}} d_0(s_0) \Pr(S_1 = s | S_0 = s_0) \quad (3)$$

$$= \sum_{s_0 \in \mathcal{S}} d_0(s_0) \sum_{a_0 \in \mathcal{A}} \Pr(A_0 = a_0 | S_0 = s_0) \Pr(S_1 = s | S_0 = s_0, A_0 = a_0) \quad (4)$$

$$= \sum_{s_0 \in \mathcal{S}} d_0(s_0) \sum_{a_0 \in \mathcal{A}} \pi(s_0, a_0) p(s_0, a_0, s). \quad (5)$$

#### Problems:

- The probability that the action at time  $t = 3$  is either  $a \in \mathcal{A}$  or  $a' \in \mathcal{A}$ , with  $a \neq a'$ .
  - The expected reward at time  $t = 6$  given that the action at time  $t = 5$  is  $a \in \mathcal{A}$  and the state at time  $t = 4$  is  $s \in \mathcal{S}$ .
  - The probability that the action at time  $t = 16$  is  $a' \in \mathcal{A}$  given that the action at time  $t = 14$  is  $a \in \mathcal{A}$ , and the state at time  $t = 15$  is  $s \in \mathcal{S}$ .
  - The probability that the initial state was  $s \in \mathcal{S}$  given that the action at time  $t = 1$  is  $a' \in \mathcal{A}$ .
- In 687-Gridworld, if we changed how rewards are generated so that hitting a wall (i.e., when the agent would enter an obstacle state, and is placed back where it started) results in a reward of  $-10$ , then what is  $\mathbf{E}[R_t | S_t = 17, A_t = \text{AL}, S_{t+1} = 17]$ ?
  - How many deterministic policies are there for an MDP with  $|\mathcal{S}| < \infty$  and  $|\mathcal{A}| < \infty$ ? (You may write your answer in terms of  $|\mathcal{S}|$  and  $|\mathcal{A}|$ ).
  - Give an example of an MDP with  $\mathcal{A} = \mathbb{R}$ , bounded rewards (but perhaps not with finite support), and  $\gamma < 1$  that does not have an optimal policy. Prove that it does not have an optimal policy. Hint:  $J$  is well-defined in this case.
  - Give an example of an MDP with  $\mathcal{A} = \mathbb{N}_1$ , bounded rewards (but perhaps not with finite support), and  $\gamma < 1$  that does not have an optimal policy. Prove that it does not have an optimal policy. Hint:  $J$  is well-defined in this case.
  - We refer to the discounted sum of rewards,  $\sum_{t=0}^{\infty} \gamma^t R_t$ , as the *return*. Let an MDP exist such that it has two optimal policies and  $\gamma < 1$ . Can the expected value of their returns differ? If so, give an example. If not, explain why. Can the variance of their returns differ? If so, give an example. If not, explain why.

7. Read the Wikipedia page on Markov chains. A *Markov chain* is *irreducible* if every state can be reached from every other state with positive probability and in finitely many steps. An MDP is *irreducible* if the Markov chain associated with every policy is irreducible.<sup>1</sup> A state in a *Markov chain* has *period*  $k$  if  $k$  is the greatest integer that divides every possible return time. More formally,

$$k = \gcd\{t > 0 : \Pr(S_t = s | S_0 = s) > 0\}.$$

A Markov chain is *aperiodic* if the period of every state is  $k = 1$ . An MDP is aperiodic if the Markov chain associated with every deterministic policy is aperiodic (or equivalently, if the Markov chain associated with every policy is aperiodic). Can an MDP be aperiodic and *not* irreducible? If so, give an example. If not, explain your reasoning.

8. The state of a Markov chain is *positive recurrent* if the expected time until the state recurs is finite. A Markov chain is *positive recurrent* if all states are positive recurrent. An MDP is *positive recurrent* if the Markov chains associated with every policy are all positive recurrent. Give an example of an MDP with  $|\mathcal{S}| > 2$  states that is positive recurrent and aperiodic. For any number of states  $|\mathcal{S}|$ , can you think of a simple way of defining state transitions such that the MDP is positive recurrent and aperiodic? Explain your methodology (a picture might be useful).
9. Describe a real-world problem and how it can be reasonably modeled as an MDP where  $R_t$  is *not* a deterministic function of  $S_t$ ,  $A_t$ , and  $S_{t+1}$ .
10. If you know  $(\mathcal{S}, \mathcal{A}, p, R, d_0, \gamma)$ , can you derive  $d_R$ ? Prove your answer is correct.
11. You are given a program, VI, that takes as input an MDP of the form  $(\mathcal{S}, \mathcal{A}, p, R, d_0, \gamma)$ , and which returns an optimal policy,  $\pi^*$ . The program is highly efficient – significantly more efficient than any other solver. However, its inner workings are secret, and so you are unable to inspect how it works or make changes to it. You are tasked with using this program to solve a large real-world problem. However, for this problem not every action is available in every state. For all  $s \in \mathcal{S}$ , let  $\mathcal{A}'(s)$  be the set of actions available in state  $s$ . So, you need to write a program that takes as input  $(\mathcal{S}, \mathcal{A}', p, R, d_0, \gamma)$ , and which returns an optimal policy  $\pi^*$ , and your program must use the VI program to do the majority of the work (since it is particularly efficient). If needed, you may assume that  $\max_{s \in \mathcal{S}} |\mathcal{A}'(s)|$  is small (roughly 5) and that  $|\bigcup_{s \in \mathcal{S}} \mathcal{A}'(s)|$  is also relatively small (roughly 10), and so the complexity of the large problem comes from having a large state set. Explain in English how your program would work. It does not need to be maximally efficient. Although this will not be as precise as pseudocode (since we haven't specified the format of the inputs to VI or your program), it should explain how the inputs to your program are transformed before being provided to VI, and how the output of VI is transformed before being returned. Assume that  $\mathcal{A}'(s) \neq \emptyset$  for every  $s \in \mathcal{S}$ .

## Part Two: Programming (25 Points)

Implement the 687-Gridworld environment. I recommend creating an object with member variables to store the current state and the following member functions:

- **newEpisode**: Resets to the initial state
- **getState**: Returns the current state.
- **inTerminalState**: Returns true if the current state is the terminal absorbing state. You may alternatively have this function return true if the current state always transitions to the terminal absorbing state.
- **step**: Takes an action as input, updates the state to the next state and returns the resulting reward (not the state – the resulting state must be queried using a separate call to **getState**).

The answers to the questions below should be numerical values and plots. You do not need to submit your source code.

1. Run 10,000 episodes using the policy that selects actions uniformly randomly from  $\mathcal{A}$ . Report the average discounted return, the standard deviation of the discounted returns, and standard error of the discounted returns.
2. Pick one policy that you think is optimal for this MDP. You do not need to be correct – just guess based on your understanding of the MDP. Report what this policy is and run 10,000 episodes using this policy. Report the average discounted return, the standard deviation of the discounted returns, and standard error of the discounted returns.
3. Create histograms of the discounted return distributions under (a) the uniform random policy and (b) your guess of the optimal policy. You may provide two plots, each with one histogram, or one plot showing both histograms.

---

<sup>1</sup>Note that once a policy is fixed, the state transitions of an MDP can be modeled as a Markov chain.

4. Using Monte Carlo simulations, approximate the probability that  $S_{19} = 21$  given that  $S_8 = 18$  (the state above the goal) when running the uniform random policy. Note that sampling episodes as in the previous questions would be inefficient (the condition that  $S_8 = 18$  is exceptionally rare), and so you should devise a more efficient sampling strategy. Explain your methodology (sampling procedure and number of samples) and report your final estimate along with a 95% confidence interval. Report the method that you used to compute the 95% confidence interval. If you're not familiar with such confidence intervals, you may use the confidence interval from Hoeffding's inequality. That is, given  $n$  i.i.d. samples  $X_1, \dots, X_n$ , where each  $X_i \in [a, b]$ ,  $[L, U]$  is a  $1 - \delta$  confidence interval for  $\mathbf{E}[X_1]$ , where:

$$L := \frac{1}{n} \sum_{i=1}^n X_i - (b - a) \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}} \quad (6)$$

$$U := \frac{1}{n} \sum_{i=1}^n X_i + (b - a) \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}. \quad (7)$$