COMPSCI 687 Homework 3
(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.

Note: For Questions 1–3, you may assume that whenever you condition on one or more events, those events have positive probability of occurring together. This applies to intermediate steps in your derivation as well as your final answer.

1. One day while working in the engineering department of the Starship Enterprise, your friend Geordi comes to you with an idea. He points out that the warp core (engine) uses a reinforcement learning algorithm to regulate its temperature. He hypothesizes the value function that it uses would be easier to represent and/or faster to approximate in two distinct parts: one that estimates the value of a state given that the next state is safe (within desirable thresholds), and another that estimates the value of a state given that the next state is not safe. Working with Geordi, who of course uses the notation from this class, you decide to define $\mathcal{X}$ to be the set of safe states, and $\mathcal{X}^{\complement}$ to be the set of unsafe states, i.e., $\mathcal{X}^{\complement} = \mathcal{S} \setminus \mathcal{X}$.[1] In order to continue, you and Geordi decide to establish some notation. Specifically, you want to define $v^{\pi}_{\mathcal{Y}}(s)$ to be the expected discounted return given that the agent begins in state s, follows policy π, and the next state (but not necessarily the states after the next state) happens to be in $\mathcal{Y}$. Give a mathematical definition for $v^{\pi}_{\mathcal{Y}}$ like our definition for v^{π}:

$$v^{\pi}_{\mathcal{Y}}(s) \coloneqq \tag{1}$$

2. Having defined $v^{\pi}_{\mathcal{Y}}$, you decide to relate your new value functions, $v^{\pi}_{\mathcal{X}}$ and $v^{\pi}_{\mathcal{X}^{\complement}}$, to the standard value function, v^{π}. Derive an expression for $v^{\pi}(s)$ that *only* uses the following terms: $\pi, \mathcal{X}^{\complement}, p, \mathcal{A}, \mathcal{S}, \mathcal{X}, v^{\pi}_{\mathcal{X}}$ and $v^{\pi}_{\mathcal{X}^{\complement}}$. Note: You may introduce variables when summing over sets, e.g., x in $\sum_{x \in \mathcal{X}}$. Your final answer should not include expectations or any random variables like S_t or R_t. You should begin with the definition of $v^{\pi}(s)$ and end with an expression that only contains the allowed terms. Show your work (show the steps, don't just jump to your final answer). You may want to derive some properties before proceeding with the derivation for $v^{\pi}(s)$—that is allowed.

$$v^{\pi}(s) = \tag{2}$$

3. Having related your new value functions to the standard value function, you now talk to Geordi about what to do next to design a reinforcement learning algorithm using your new value functions. Another friend named Data loads the course notes from COMPSCI 687. He finds that the next step towards developing an algorithm with this value function may be to write out a new Bellman equation for $v^{\pi}_{\mathcal{X}}$. Derive a Bellman-like equation for this new value function. You should begin with the definition of $v^{\pi}_{\mathcal{X}}$ according to your answer to the first question, and should end with a recursive expression for $v^{\pi}_{\mathcal{X}}$ that is written only in terms of $\mathcal{S}, \mathcal{A}, p, R, d_0, \gamma, \pi, \mathcal{X}$, and $\mathcal{X}^{\complement}$ (you may choose not to use some of these terms, but you cannot use others). Your definition may also be in terms of $v^{\pi}_{\mathcal{X}}$ and $v^{\pi}_{\mathcal{X}^{\complement}}$, but should be in the form of a Bellman-like equation (e.g., don't write $v^{\pi}_{\mathcal{X}}(s) = v^{\pi}_{\mathcal{X}}(s)$). **Note:** For this problem you may assume that $\mathbf{E}[R_t | S_t = s, A_t = a] = \mathbf{E}[R_t | S_t = s, A_t = a, S_{t+1} = s']$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, $s' \in \mathcal{S}$, and $t \in \mathbb{N}_0$, even though this property is not necessarily true for all MDPs. **Hint:** Our solution is quite long – using LaTeX font size "tiny" it barely fits on a line.

$$v^{\pi}_{\mathcal{X}}(s) = \tag{3}$$

4. Let M and M' be two MDPs that are identical except for their initial state distributions. Prove that v^{π} is the same for both MDPs. You may write v^{π}_{M} and $v^{\pi}_{M'}$ to denote the state-value functions for π on M and M' respectively.

5. Consider a finite MDP where all rewards are negative (i.e., $R_t < 0$ always) and where every episode ends after precisely L time steps (i.e., $S_L = s_\infty$, and s_∞ does not occur before time L). Let $\gamma = 1$. The MDP has a deterministic transition function and initial state distribution (rewards may be stochastic). Let $h_L = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_L, a_L, r_L)$ be any history of length $L+1$ that can be generated by a deterministic policy, π. Prove that the sequence $v^{\pi}(s_0), v^{\pi}(s_1), \ldots, v^{\pi}(s_L)$ is strictly increasing.

[1] In latex, here we are using the symbols \complement and \setminus for $\complement$ and $\setminus$ respectively.