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}:

   Answers should be equivalent to

$$v^{\pi}_{\mathcal{Y}}(s) \coloneqq \mathbf{E}\left[G_t | S_t = s, S_{t+1} \in \mathcal{Y}; \pi\right]. \tag{1}$$

   You might expand the definition of G_t giving the equally acceptable answer:

$$v^{\pi}_{\mathcal{Y}}(s) \coloneqq \mathbf{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s, S_{t+1} \in \mathcal{Y}; \pi\right]. \tag{2}$$

   For an MDP, the value of the right hand side does not depend on t, and so you could also substitute in a specific value of t, typically $t = 0$ giving:

$$v^{\pi}_{\mathcal{Y}}(s) \coloneqq \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| S_0 = s, S_1 \in \mathcal{Y}; \pi\right]. \tag{3}$$

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

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.

First notice that for any set $\mathcal{Y} \subseteq \mathcal{S}$:

$$\Pr(S_{t+1} \in \mathcal{Y} | S_t = s; \pi) = \sum_{s' \in \mathcal{S}} \Pr(S_{t+1} = s' | S_t = s; \pi) \Pr(S_{t+1} \in \mathcal{Y} | S_t = s, S_{t+1} = s'; \pi) \tag{4}$$

$$\overset{(a)}{=} \sum_{s' \in \mathcal{Y}} \Pr(S_{t+1} = s' | S_t = s; \pi) \underbrace{\Pr(S_{t+1} \in \mathcal{Y} | S_t = s, S_{t+1} = s'; \pi)}_{=1} \tag{5}$$

$$+ \sum_{s' \in \mathcal{S} \setminus \mathcal{Y}} \Pr(S_{t+1} = s' | S_t = s; \pi) \underbrace{\Pr(S_{t+1} \in \mathcal{Y} | S_t = s, S_{t+1} = s'; \pi)}_{=0} \tag{6}$$

$$= \sum_{s' \in \mathcal{Y}} \Pr(S_{t+1} = s' | S_t = s; \pi) \tag{7}$$

$$= \sum_{s' \in \mathcal{Y}} \sum_{a \in \mathcal{A}} \Pr(A_t = a | S_t = s; \pi) \Pr(S_{t+1} = s' | S_t = s, A_t = a; \pi) \tag{8}$$

$$= \sum_{s' \in \mathcal{Y}} \sum_{a \in \mathcal{A}} \pi(s, a) p(s, a, s'), \tag{9}$$

where **(a)** follows from splitting the summation over $\mathcal{S}$ into separate sums over $\mathcal{Y}$ and $\mathcal{S} \setminus \mathcal{Y}$ and recognizing that two probability terms are necessarily 0 or 1 due to the conditioning of $S_{t+1} = s'$ for $s' \in \mathcal{Y}$ or $s' \in \mathcal{S} \setminus \mathcal{Y}$. We now answer the question:

$$v^\pi(s) = \mathbf{E}\left[G_t | S_t = s; \pi\right] \tag{10}$$

$$\overset{(a)}{=} \Pr(S_{t+1} \in \mathcal{X} | S_t = s; \pi) \mathbf{E}\left[G_t | S_t = s, S_{t+1} \in \mathcal{X}; \pi\right] \tag{11}$$

$$+ \Pr(S_{t+1} \in \mathcal{X}^\complement | S_t = s; \pi) \mathbf{E}\left[G_t \middle| S_t = s, S_{t+1} \in \mathcal{X}^\complement; \pi\right] \tag{12}$$

$$\overset{(b)}{=} \Pr(S_{t+1} \in \mathcal{X} | S_t = s; \pi) v^\pi_{\mathcal{X}}(s) + \Pr(S_{t+1} \in \mathcal{X}^\complement | S_t = s; \pi) v^\pi_{\mathcal{X}^\complement}(s) \tag{13}$$

$$\overset{(c)}{=} \sum_{s' \in \mathcal{X}} \sum_{a \in \mathcal{A}} \pi(s, a) p(s, a, s') v^\pi_{\mathcal{X}}(s) + \sum_{s' \in \mathcal{X}^\complement} \sum_{a \in \mathcal{A}} \pi(s, a) p(s, a, s') v^\pi_{\mathcal{X}^\complement}(s), \tag{14}$$

where **(a)** follows from the law of total expectation, **(b)** follows from the definition of $v^\pi_{\mathcal{Y}}$, and **(c)** follows from the property we derived first. Note that the terms $v^\pi_{\mathcal{X}}(s)$ and $v^\pi_{\mathcal{X}^\complement}(s)$ could be moved before the summations:

$$v^\pi(s) = v^\pi_{\mathcal{X}}(s) \sum_{s' \in \mathcal{X}} \sum_{a \in \mathcal{A}} \pi(s, a) p(s, a, s') + v^\pi_{\mathcal{X}^\complement}(s) \sum_{s' \in \mathcal{X}^\complement} \sum_{a \in \mathcal{A}} \pi(s, a) p(s, a, s'). \tag{15}$$

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.

We begin by expanding the definition of the expected value, and then applying Bayes' theorem so that the events being conditioned on are events from time steps prior to the event whose probability is being measured:

$$v^\pi_{\mathcal{X}}(s) = \mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}; \pi\right] \tag{16}$$

$$= \sum_{a\in\mathcal{A}} \Pr(A_t = a|S_t = s, S_{t+1} \in \mathcal{X}; \pi)\mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a; \pi\right] \tag{17}$$

$$\overset{\text{(a)}}{=} \sum_{a\in\mathcal{A}} \frac{\Pr(S_{t+1} \in \mathcal{X}|S_t = s, A_t = a; \pi)\Pr(A_t = a|S_t = s; \pi)}{\Pr(S_{t+1} \in \mathcal{X}|S_t = s; \pi)}\mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a; \pi\right] \tag{18}$$

$$= \sum_{a\in\mathcal{A}} \frac{\pi(s,a)\sum_{s'\in\mathcal{X}} p(s,a,s')}{\sum_{b\in\mathcal{A}}\pi(s,b)\sum_{s'\in\mathcal{X}} p(s,b,s')}\mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a; \pi\right], \tag{19}$$

where **(a)** follows from Bayes' theorem. Next we focus on the expected value term, applying the law of total expectation to sum over possible values of S_{t+1} given that $S_{t+1} \in \mathcal{X}$:

$$\mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a; \pi\right] \tag{20}$$

$$= \sum_{s'\in\mathcal{S}} \Pr(S_{t+1} = s'|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a; \pi)\mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a, S_{t+1} = s'; \pi\right] \tag{21}$$

$$= \sum_{s'\in\mathcal{X}} \Pr(S_{t+1} = s'|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a; \pi)\mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a, S_{t+1} = s'; \pi\right] \tag{22}$$

$$+ \sum_{s'\in\mathcal{X}^\complement} \underbrace{\Pr(S_{t+1} = s'|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a; \pi)}_{=0}\mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a, S_{t+1} = s'; \pi\right] \tag{23}$$

$$= \sum_{s'\in\mathcal{X}} \Pr(S_{t+1} = s'|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a; \pi)\mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a, S_{t+1} = s'; \pi\right] \tag{24}$$

$$\overset{\text{(a)}}{=} \sum_{s'\in\mathcal{X}} \frac{\Pr(S_{t+1} = s', S_{t+1} \in \mathcal{X}|S_t = s, A_t = a; \pi)}{\Pr(S_{t+1} \in \mathcal{X}|S_t = s, A_t = a; \pi)}\mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a, S_{t+1} = s'; \pi\right] \tag{25}$$

$$= \sum_{s'\in\mathcal{X}} \frac{p(s,a,s')}{\sum_{x'\in\mathcal{X}} p(s,a,x')}\mathbf{E}\left[G_t|S_t = s, S_{t+1} \in \mathcal{X}, A_t = a, S_{t+1} = s'; \pi\right], \tag{26}$$

where **(a)** follows from Bayes' theorem. Next, focusing on the new remaining expected value we have that

$$\mathbf{E}\left[G_t|S_t=s,S_{t+1}\in\mathcal{X},A_t=a,S_{t+1}=s';\pi\right] \tag{27}$$

$$\overset{(a)}{=}\mathbf{E}\left[G_t|S_t=s,A_t=a,S_{t+1}=s';\pi\right] \tag{28}$$

$$=\mathbf{E}\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k}\middle|S_t=s,A_t=a,S_{t+1}=s';\pi\right] \tag{29}$$

$$=\left(\mathbf{E}\left[R_t|S_t=s,A_t=a,S_{t+1}=s';\pi\right]+\mathbf{E}\left[\sum_{k=1}^{\infty}\gamma^k R_{t+k}\middle|S_t=s,A_t=a,S_{t+1}=s';\pi\right]\right) \tag{30}$$

$$\overset{(b)}{=}\left(R(s,a)+\gamma\mathbf{E}\left[G_{t+1}|S_t=s,A_t=a,S_{t+1}=s';\pi\right]\right) \tag{31}$$

$$\overset{(c)}{=}\left(R(s,a)+\gamma\mathbf{E}\left[G_{t+1}|S_{t+1}=s';\pi\right]\right) \tag{32}$$

$$=\left(R(s,a)+\gamma v^{\pi}(s')\right). \tag{33}$$

where **(a)** follows because for $s'\in\mathcal{X}$, conditioning on $S_{t+1}=s'$ implies the condition on $S_{t+1}\in\mathcal{X}$, **(b)** relies on the restriction from the "Note" in the problem to obtain the $R(s,a)$ term, and **(c)** follows due to the Markov property. Substituting the expression for $v^{\pi}(s')$ from the previous question into (33) we obtain:

$$\mathbf{E}\left[G_t|S_t=s,S_{t+1}\in\mathcal{X},A_t=a,S_{t+1}=s';\pi\right] \tag{34}$$

$$=R(s,a)+\gamma\left(v^{\pi}_{\mathcal{X}}(s')\sum_{s''\in\mathcal{X}}\sum_{a'\in\mathcal{A}}\pi(s',a')p(s',a',s'')+v^{\pi}_{\mathcal{X}^{\complement}}(s')\sum_{s''\in\mathcal{X}^{\complement}}\sum_{a'\in\mathcal{A}}\pi(s',a')p(s',a',s'')\right). \tag{35}$$

Substituting this expression into (26) we obtain

$$\mathbf{E}\left[G_t|S_t=s,S_{t+1}\in\mathcal{X},A_t=a;\pi\right] \tag{36}$$

$$=\sum_{s'\in\mathcal{X}}\frac{p(s,a,s')}{\sum_{x'\in\mathcal{X}}p(s,a,x')}\left(R(s,a)+\gamma\left(v^{\pi}_{\mathcal{X}}(s')\sum_{s''\in\mathcal{X}}\sum_{a'\in\mathcal{A}}\pi(s',a')p(s',a',s'')+v^{\pi}_{\mathcal{X}^{\complement}}(s')\sum_{s''\in\mathcal{X}^{\complement}}\sum_{a'\in\mathcal{A}}\pi(s',a')p(s',a',s'')\right)\right). \tag{37}$$

Finally, substituting this expression for the expected value into (19) we obtain:

$$v^{\pi}_{\mathcal{X}}(s)=\sum_{a\in\mathcal{A}}\frac{\pi(s,a)\sum_{s'\in\mathcal{X}}p(s,a,s')}{\sum_{b\in\mathcal{A}}\pi(s,b)\sum_{s'\in\mathcal{X}}p(s,b,s')}\sum_{s'\in\mathcal{X}}\frac{p(s,a,s')}{\sum_{x'\in\mathcal{X}}p(s,a,x')}\left(R(s,a)+\gamma\left(v^{\pi}_{\mathcal{X}}(s')\sum_{s''\in\mathcal{X}}\sum_{a'\in\mathcal{A}}\pi(s',a')p(s',a',s'')+v^{\pi}_{\mathcal{X}^{\complement}}(s')\sum_{s''\in\mathcal{X}^{\complement}}\sum_{a'\in\mathcal{A}}\pi(s',a')p(s',a',s'')\right)\right). \tag{38}$$

While this answer is sufficient, we can simplify it by canceling the two red terms and moving the summation over $s'\in\mathcal{X}$ earlier to obtain:

$$v^{\pi}_{\mathcal{X}}(s)=\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{X}}\frac{\pi(s,a)p(s,a,s')}{\sum_{b\in\mathcal{A}}\pi(s,b)\sum_{x'\in\mathcal{X}}p(s,b,x')}\left(R(s,a)+\gamma\left(v^{\pi}_{\mathcal{X}}(s')\sum_{s''\in\mathcal{X}}\sum_{a'\in\mathcal{A}}\pi(s',a')p(s',a',s'')+v^{\pi}_{\mathcal{X}^{\complement}}(s')\sum_{s''\in\mathcal{X}^{\complement}}\sum_{a'\in\mathcal{A}}\pi(s',a')p(s',a',s'')\right)\right). \tag{39}$$

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.

Let us write the first MDP's transition and reward functions as p and R, and the second's as p' and R'. Note that $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \forall s' \in \mathcal{S}, p(s,a,s') = p'(s,a,s')$ and $R(s,a) = R'(s,a)$. We proceed by writing $v^\pi_M(s)$ as an expected value, and then applying the law of total expectation over and over until we obtain an infinite expression in terms of π, p, and R. We then replace the p and R terms with p' and R', and undo the same steps to obtain $v^\pi_{M'}(s)$.

$$
\begin{aligned}
v^\pi_M(s) &= \mathbf{E}_M[G_t | S_t = s; \pi] \\
&= \sum_{k=0}^{\infty} \gamma^k \mathbf{E}_M[R_{t+k} | S_t = s; \pi] \\
&= \sum_{a \in \mathcal{A}} \pi(s,a) \left(R(s,a) + \sum_{k=1}^{\infty} \gamma^k \mathbf{E}_M[R_{t+k} | S_t = s; \pi] \right) \\
&= \sum_{a \in \mathcal{A}} \pi(s,a) \left(R(s,a) + \sum_{s' \in \mathcal{S}} p(s,a,s') \sum_{a' \in \mathcal{A}} \pi(s',a')(\gamma^1 R(s',a') + \sum_{k=2}^{\infty} \gamma^k \mathbf{E}_M[R_{t+k} | S_t = s; \pi]) \right) \\
&= \gamma^0 \sum_{a \in \mathcal{A}} \pi(s,a) R(s,a) + \gamma^1 \sum_{a \in \mathcal{A}} \pi(s,a) \sum_{s' \in \mathcal{S}} p(s,a,s') \sum_{a' \in \mathcal{A}} \pi(s',a') R(s',a') \\
&\quad + \gamma^2 \sum_{a \in \mathcal{A}} \pi(s,a) \sum_{s' \in \mathcal{S}} p(s,a,s') \sum_{a' \in \mathcal{A}} \pi(s',a') \sum_{s'' \in \mathcal{S}} p(s',a',s'') \sum_{a'' \in \mathcal{A}} \pi(s'',a'') R(s'',a'') + \dots \\
&= \gamma^0 \sum_{a \in \mathcal{A}} \pi(s,a) R'(s,a) + \gamma^1 \sum_{a \in \mathcal{A}} \pi(s,a) \sum_{s' \in \mathcal{S}} p'(s,a,s') \sum_{a' \in \mathcal{A}} \pi(s',a') R'(s',a') \\
&\quad + \gamma^2 \sum_{a \in \mathcal{A}} \pi(s,a) \sum_{s' \in \mathcal{S}} p'(s,a,s') \sum_{a' \in \mathcal{A}} \pi(s',a') \sum_{s'' \in \mathcal{S}} p'(s',a',s'') \sum_{a'' \in \mathcal{A}} \pi(s'',a'') R'(s'',a'') + \dots \\
&\overset{\textbf{(a)}}{=} \mathbf{E}_{M'}[G_t | S_t = s; \pi] \\
&= v^\pi_{M'}(s),
\end{aligned}
$$

where **(a)** follows from reversing the first 4 steps and $\mathbf{E}_M$ indicates that the random variables inside the expectation are those associated with MDP M.

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.

Notice that with deterministic initial state distribution, transition function, and policy, the state and action sequences are deterministic. Only the rewards remain random. Hence, conditioning on events like $S_t = s_t$ doesn't change any probabilities or expected values, since this event always occurs. This means that we can freely add and remove such conditions on state values in statements of probability or expectation without changing their value, e.g., $\mathbf{E}[R_7 | S_7 = s_7; \pi] = \mathbf{E}[R_7; \pi]$. We show that for all $t < L$, $v^\pi(s_t) < v^\pi(s_{t+1})$ using these properties:

$$v^\pi(s_t) = \mathbf{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s_t; \pi\right] \tag{40}$$

$$\stackrel{(a)}{=} \mathbf{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k}; \pi\right] \tag{41}$$

$$= \mathbf{E}\left[R_t; \pi\right] + \mathbf{E}\left[\sum_{k=1}^{\infty} \gamma^k R_{t+k}; \pi\right] \tag{42}$$

$$\stackrel{(b)}{=} \mathbf{E}\left[R_t; \pi\right] + \gamma \mathbf{E}\left[\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k} \middle| S_{t+1} = s_{t+1}; \pi\right] \tag{43}$$

$$= \mathbf{E}\left[R_t; \pi\right] + \gamma v^\pi(s_{t+1}) \tag{44}$$

$$\stackrel{(c)}{<} v^\pi(s_{t+1}), \tag{45}$$

where **(a)** holds because conditioning on $S_t = s_t$ changes nothing since this event always occurs, **(b)** holds because conditioning on $S_{t+1} = s_{t+1}$ also changes nothing since this event always occurs, and **(c)** holds because $R_t < 0$ for $t < L$ as part of the problem statement (if $t \geq L$, then $R_t = 0$ and the inequality would not be strict).