CMPSCI 687 Homework 2

Due October 17, 2017, 11pm Eastern Time

Instructions: This homework assignment consists of only a written portion. You may discuss concepts related to the written portion with other students, but should not discuss how to solve the specific questions with other students. Submissions must be typed (hand written and scanned submissions will not be accepted). We recommend that you use LaTeX. The assignment should be submitted as a single .pdf on Moodle. The automated system will not accept assignments after 11:55pm on the due date specified above.

Written Portion (65 Points Total)

1. (10 Points) Apply value iteration to the gridworld used in class (with stochastic state transitions and zero reward for hitting obstacles, as it was originally presented). Remember that R_t is -10 if S_{t+1} is the state with water, and R_t is $+10$ if S_{t+1} is the bottom-right state. Use $\gamma = 1.0$. Begin with the value of every state being zero. Draw the value function as a 5×5 grid with two cells missing (the obstacles), with numbers in each cell of the grid correspond to the current estimate of the value of that state. After computing v_{k+1}, round all values to three decimal places before continuing (your answer should include three decimal places, and future computations should use the rounded values). Show the first ten iterations of value iteration. Below is the initial value function:

v_0:

0	0	0	0	0
0	0	0	0	0
0	0	N/A	0	0
0	0	N/A	0	0
0	0	0	0	0

v_1 :

0.000	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000
0.000	0.000	N/A	0.000	0.000
0.000	0.000	N/A	0.000	8.000
0.000	0.000	-2.000	8.000	0.000

v_2 :

0.000	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	0.000
0.000	0.000	N/A	0.000	6.400
0.000	0.000	N/A	6.800	9.200
0.000	0.000	4.000	9.200	0.000

v_3 :

0.000	0.000	0.000	0.000	0.000
0.000	0.000	0.000	0.000	5.120
0.000	0.000	N/A	5.760	8.320
0.000	0.000	N/A	8.840	9.720
0.000	0.000	6.160	9.720	0.000

v_4 :

0.000	0.000	0.000	0.000	4.096
0.000	0.000	0.000	4.864	7.424
0.000	0.000	N/A	8.352	9.312
0.000	0.000	N/A	9.588	9.900
0.000	0.000	7.008	9.900	0.000

v_5 :

0.000	0.000	0.000	4.096	6.554
0.000	0.000	3.891	7.539	8.806
0.000	0.000	N/A	9.389	9.734
0.000	0.000	N/A	9.853	9.964
0.000	0.000	7.322	9.964	0.000

v_6 :

0.000	0.000	3.471	6.769	8.233
0.000	3.113	6.615	8.900	9.485
0.000	0.000	N/A	9.777	9.901
0.000	0.000	N/A	9.947	9.987
0.000	0.000	7.436	9.987	0.000

	0.000	2.932	6.267	8.382	9.161
	2.490	5.603	8.286	9.517	9.789
v_7 :	0.000	2.490	N/A	9.919	9.964
	0.000	0.000	N/A	9.981	9.995
	0.000	0.000	7.477	9.995	0.000

	2.470	5.734	8.060	9.223	9.624
	4.731	7.460	9.170	9.791	9.915
v_8 :	2.117	4.856	N/A	9.971	9.987
	0.000	1.992	N/A	9.993	9.998
	0.000	0.000	7.491	9.998	0.000

	5.194	7.681	9.046	9.639	9.837
	6.670	8.611	9.611	9.910	9.966
v_9 :	4.345	6.802	N/A	9.989	9.995
	1.793	4.184	N/A	9.997	9.999
	0.000	1.468	7.497	9.999	0.000

	7.257	8.819	9.549	9.836	9.930
	8.033	9.274	9.822	9.961	9.986
v_{10} :	6.328	8.126	N/A	9.996	9.998
	3.954	6.159	N/A	9.999	10.000
	1.508	3.369	7.499	10.000	0.000

2. (10 Points) Prove that multiplying all rewards (of a finite MDP with bounded rewards) by a positive scalar does not change which policies are optimal, using either of the definitions of optimal policies that we covered in class (that is, show it for at least one of the definitions that we covered in class).

Consider a finite MDP, M, with bounded rewards. If $\pi^\star$ is an optimal policy for M, then we will show that it is an optimal policy for all MDPs M_α that are identical to M, except that their rewards, R'_t are a positive constant times the rewards of M, i.e., $R'_t = \alpha R_t$ for some $\alpha > 0$.

Let $J_\alpha(\pi) \coloneqq \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t \alpha R_t\right]$. That is, J_α is the objective function for the MDP M_α, and $J_1 = J$ (recall that J is the objective function for M, as defined in class). If $\pi^\star$ is an optimal policy for M, then for all $\pi \in \Pi$:

$$\begin{aligned} J(\pi^\star) &\geq J(\pi) \\ \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| \pi^\star\right] &\geq \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| \pi\right]. \end{aligned}$$

Multiplying both sides by α we have that for all $\pi \in \Pi$:

$$\begin{aligned} \alpha\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| \pi^\star\right] &\geq \alpha\mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| \pi\right] \\ \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t \alpha R_t \middle| \pi^\star\right] &\geq \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t \alpha R_t \middle| \pi\right] \\ J_\alpha(\pi^\star) &\geq J_\alpha(\pi). \end{aligned}$$

Thus, $\pi^\star$ is a maximizer of J_α and therefore an optimal policy for M_α.

3. (5 Points) Prove that adding a positive constant to all rewards (of a finite MDP with bounded rewards) can change which policies are optimal, using

either of the definitions of optimal policies that we covered in class.

Consider the MDP shown in Figure 1. S0 is the start state. π_1 denotes the policy that takes action 1, and π_2 denotes the policy that takes action 2. Initially, let every transition give a deterministic reward of -1. Then, $J(\pi_1) = -1$ and $J(\pi_2) = -2$, so π_1 is the optimal policy. Next, add 2 to the reward given by every transition, so every transition now gives a reward of 1. In this case, $J(\pi_1) = 1$ and $J(\pi_2) = 2$. Now, π_2 is the optimal policy; this shows that adding a constant to all rewards can change which policies are optimal.

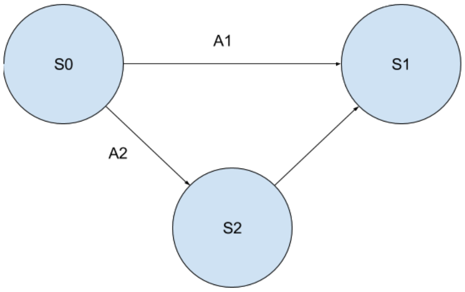

Figure 1: MDP for Question 3

4. (5 Points) Your boss asked you to estimate the state-value function associated with a known policy, π, for a specific MDP. You misheard and instead estimated the action-value function. This estimation was very expensive, and so you do not want to do it again. Explain how you could easily retrieve the value of any state given what you have already computed.

   Use the property: $v^\pi(s) = \sum_{a\in\mathcal{A}} \pi(s,a) q^\pi(s,a)$. That is, when asked for $v^\pi(s)$, compute the right side of this equation, which uses q.

5. (10 Points) Consider a finite MDP with bounded rewards, where all rewards are negative. That is, $R_t < 0$ always. Let $\gamma = 1$. The MDP is finite horizon, with horizon L, and also has a deterministic transition function and initial state distribution (rewards may be stochastic). Let $H = (S_0, A_0, R_0, S_1, A_2, R_1, \ldots, S_{L-1}, A_{L-1}, R_{L-1})$ be any history that can be generated by a deterministic policy, π. Prove that the sequence $v^\pi(S_0), v^\pi(S_1), \ldots, v^\pi(S_{L-1})$ is strictly increasing.

$$\begin{aligned}
v^\pi(S_t) &\overset{\text{(a)}}{=} v^\pi(S_t) \\
&\overset{\text{(b)}}{=} v^\pi(s_t) \\
&\overset{\text{(c)}}{=} \mathbf{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s_t, \pi\right] \\
&\overset{\text{(d)}}{=} \sum_{k=0}^{\infty} \mathbf{E}\left[R_{t+k} | S_t = s_t, \pi\right] \\
&\overset{\text{(e)}}{=} \sum_{k=0}^{\infty} \mathbf{E}\left[R_{t+k} | \pi\right] \\
&= \mathbf{E}\left[R_t | \pi^\star\right] + \sum_{k=0}^{\infty} \mathbf{E}\left[R_{t+k+1} | \pi\right] \\
&\overset{\text{(f)}}{=} \mathbf{E}\left[R_t | \pi\right] + \sum_{k=0}^{\infty} \mathbf{E}\left[R_{t+k+1} | S_{t+1} = s_{t+1}, \pi\right] \\
&\overset{\text{(g)}}{=} \mathbf{E}\left[R_t | \pi\right] + v^\pi(s_{t+1}) \\
&\overset{\text{(h)}}{=} \mathbf{E}\left[R_t | \pi\right] + v^\pi(S_{t+1}) \\
&\overset{\text{(i)}}{\leq} v^\pi(S_{t+1}),
\end{aligned}$$

where **(a)** holds because $v^\star$ is the state-value function associated with all optimal policies, $\pi^\star$, as discussed in class, **(b)** holds because $S_t = s_t$ (for some $s_t \in \mathcal{S}$) deterministically due to the transition function and initial state distributions being deterministic, **(c)** comes from the definition of the state-value function, **(d)** holds because $\gamma = 1$, **(e)** and **(f)** hold because the sequence of states is deterministic and so conditioning on $S_t = s_t$ or $S_{t+1} = s_{t+1}$ is conditioning on an event that always happens, **(g)** holds by the definition of $v^{\pi^\star}$, **(h)** holds because $S_{t+1} = s_{t+1}$ always, and **(i)** holds because $R_t < 0$ (this was specified in the problem statement).

6. (15 Points) The Bellman operator for q-functions is:

$$\mathcal{T} : \mathcal{Q} \to \mathcal{Q},$$

where $\mathcal{Q}$ is the set of all functions, $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and

$$Tq(s,a) \coloneqq \sum_{s'} P(s,a,s') \left(R(s,a,s') + \gamma \max_{a'} q(s',a') \right).$$

Prove that the Bellman operator for q-functions is a contraction mapping.

To prove that this is a contraction mapping, we must prove that

$$||Tq - Tq'|| \leq \gamma ||q - q'||$$

where we have defined

$$||q|| = \max_{s,a} |q(s,a)|.$$

This gives

$$\begin{aligned}
||Tq - q'|| &\overset{\text{a}}{=} \max_{s,a} |Tq(s,a) - Tq'(s,a)| \\
&\overset{\text{b}}{=} \max_{s,a} |\sum_{s'} P(s,a,s')(R(s,a,s') + \gamma \max_{a'} q(s',a')) \\
&\quad - \sum_{s'} P(s,a,s')(R(s,a,s') + \gamma \max_{a'} q'(s',a'))| \\
&\overset{\text{c}}{=} \max_{s,a} |\sum_{s'} P(s,a,s')\gamma(\max_{a'} q(s',a') - \max_{a'} q'(s',a'))| \\
&\overset{\text{d}}{\leq} \max_{s,a} \max_{s'} \gamma| \max_{a'} q(s',a') - \max_{a'} q'(s',a')| \\
&\overset{\text{e}}{\leq} \max_{s,a} \max_{s'} \max_{a'} \gamma|q(s',a') - q'(s',a')| \\
&\overset{\text{f}}{=} \gamma \max_{s',a'} |q(s',a') - q'(s',a')| \\
&\overset{\text{g}}{=} \gamma||q - q'||].
\end{aligned}$$

(a) holds by the definition of the norm; **(b)** holds by the definition of T; **(c)** holds due to algebraic simplification; **(d)** holds because $P(s,a,s') \leq 1$; **(e)** holds because $|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$, as shown in class; **(f)** holds because the expression does not depend on s or a; **(g)** holds by the definition of the norm. This proves that the Bellman operator for q-functions is a contraction mapping, as desired.

7. (10 Points) A researcher proposes an estimator, $\hat{J}$, of J. The estimator uses data to estimate the performance of a policy. That is, $\hat{J}(\pi, H)$ corresponds to the estimator's estimate of $J(\pi)$, where H is a history produced by running π for one episode. Specifically:

$$\hat{J}(\pi, H) = \sum_{t=0}^{\infty} \gamma^t \left(R_t - R(S_t, A_t, S_{t+1})\right) + \sum_{t=0}^{\infty} \gamma^t \sum_{s'} P(S_t, A_t, s') R(S_t, A_t, s').$$

Now consider the case where we have a data set, D_n, that includes $n \in \mathbb{N}_{>0}$ i.i.d. histories, i.e., $D_n = (H_1, \dots, H_n)$, each produced by running the policy π. We construct a new estimator, $\hat{J}_n(\pi, D_n) = \frac{1}{n}\sum_{i=1}^{n} \hat{J}(\pi, H_i)$. Prove that $\hat{J}_n(\pi, D_n)$ converges in probability to $J(\pi)$. That is, for all ϵ,

$$\lim_{n \to \infty} \Pr\left(|\hat{J}_n(\pi, D_n) - J(\pi)| > \epsilon\right) = 0.$$

We begin by showing that $\hat{J}(\pi, H)$ is an unbiased estimator of $J(\pi)$. Below all expected values are conditioned on using the policy π (we omit this to avoid clutter):

$$\begin{aligned}
\mathbf{E}[\hat{J}(\pi, H)] =& \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t (R_t - R(S_t, A_t, S_{t+1}) + \sum_{t=0}^{\infty} \gamma^t \sum_{s'} P(S_t, A_t, s') R(S_t, A_t, s')] \\
=& \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t R_t] - \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t, S_{t+1})] + \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t \sum_{s'} P(S_t, A_t, s') R(S_t, A_t, s')] \\
\overset{a}{=}& \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t R_t] - \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t, S_{t+1})] + \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t, S_{t+1})] \\
=& \mathbf{E}[\sum_{t=0}^{\infty} \gamma^t R_t] \\
\overset{b}{=}& J(\pi)
\end{aligned}$$

Note that $\sum_{s'} P(S_t, A_t, s') R(S_t, A_t, s')$ is the expected reward for taking action A_t in state S_t. Next states come from i.i.d. histories, so this is equal to $R(S_t, A_t, S_t)$, which explains the substitution in step **a**. Then, step **b** holds due to the definition of $J(\pi)$. Since we have shown that $\hat{J}(\pi, H)$ is an unbiased estimator of $J(\pi)$, it follows immediately from the weak law of large numbers that $\hat{J}(\pi, H)$ converges in probability to $J(\pi)$.