CIS 510 Assignment 3

Steven Walton

June 6, 2019

Problem P1

Problem Q1

Part 3)

Consider a first price sealed-bid auction with n risk-neutral agents whose valuations, $v_1, \cdots, v_n$, are independently drawn from a uniform distribution on the interval $[0, b]$. Prove that $\left(\frac{n-1}{n}v_1, \cdots, \frac{n-1}{n}v_n\right)$ is a Bayes-Nash equilibrium.
We will follow a similar formula to the two player game that we did in class. We will let $\mathcal{V}$ be the space of players.

$$
\begin{aligned}
&\int_{\mathcal{V},0}^{b} u_1(s_1)dv \\
=&\int_{\mathcal{V},j}^{s_1} u_1(s_1)dv + \int_{\mathcal{V},s_1}^{b} u_1(s_1)dv \\
=&\int_{\mathcal{V},0}^{s_1} u_1(s_1)dv + 0 \\
=&\int_{\mathcal{V},0}^{s_1} (v_1 - s_1)dv \\
=&(v_1 - s_1)\int_{\mathcal{V},0}^{s_1} dv \\
=&(v_1 - s_1)(s_1^{n-1} - a) \\
=&s_1^{n-1}v_1 - s_1^n - av_1 + as_1
\end{aligned}
$$

$$
\begin{aligned}
&\left(\frac{d}{ds_1}\{s_1^{n-1} - s_1^n\}\right) \\
&= (n-1)v_1 - ns_1 \\
ns_1 &= (n-1)v_1 + a \\
s_1 &= \frac{n-1}{n}v_1
\end{aligned}
$$

This method can similarly be used for each player following the same pattern. We should see that v_1, s_1 can be replaced with v_i, s_i and we will get a similar result.
Original problem had bounds $[a, b]$ which creates an offset by a. This results in the profit smaller than a being obtained.

Problem Q2

Image an unknown game which has three states $\{A, B, C\}$ and in each state the agent has two actions to choose from $\{Up, Down\}$. Suppose a game agent chooses actions according to some policy π and generates the following sequence of actions and rewards in the unknown game:

t	s_t	a_t	s_{t+1}	r_t
0	A	Down	B	2
1	B	Down	C	3
2	C	Up	B	-2
3	B	Down	B	0
4	B	UP	A	1
5	A	Down	C	-3
6	C	Down	A	2
7	A	Up	C	1
8	C	Down	B	2
9	B	Down	A	2
10	A	Up	B	3

Table 1: $\gamma = 0.5$ and $\alpha = 0.5$

Part a)

Assume that all Q-values are initialized to 0. What are the Q-values learned by running Q-learning with the above experience sequence?
We have the algorithm for updating values

$$Q(s,a) = T(s,a,s')[R(s,a,s') + \gamma V(s')$$

and

$$V^\pi(s) = (1-\alpha)V(s) + \alpha[R(s,\pi(s),s') + \gamma V^\pi(s')$$

Using these we will iterate over the values

$$Q_{init}(s,a) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$V = (1-0.5)0 + 0.5(2+0.5*0) = 1$$

$$Q_0(s,a) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$V = (1-0.5)0 + 0.5(3+0.5*0) = 1.5$$

$$Q_1(s,a) = \begin{bmatrix} 0 & 1 \\ 0 & \frac{3}{2} \\ 0 & 0 \end{bmatrix}$$

$$V = (1-0.5)0 + 0.5(-2+0.5*\frac{3}{2}) = -\frac{5}{8}$$

$$Q_2(s,a) = \begin{bmatrix} 0 & 1 \\ 0 & \frac{3}{2} \\ -\frac{5}{8} & 0 \end{bmatrix}$$

$$V = (1-0.5)\frac{3}{2} + 0.5(0+0.5*\frac{3}{2}) = \frac{9}{8}$$

$$Q_3(s,a) = \begin{bmatrix} 0 & 1 \\ 0 & \frac{9}{8} \\ -\frac{5}{8} & 0 \end{bmatrix}$$

$$V = (1-0.5)0 + 0.5(1+0.5*1) = \frac{3}{4}$$

$$Q_4(s,a) = \begin{bmatrix} 0 & 1 \\ \frac{3}{4} & \frac{9}{8} \\ -\frac{5}{8} & 0 \end{bmatrix}$$

$$V = (1-0.5)1 + 0.5(-3+0.5*0) = -1$$

$$Q_5(s,a) = \begin{bmatrix} 0 & -1 \\ \frac{3}{4} & \frac{9}{8} \\ -\frac{5}{8} & 0 \end{bmatrix}$$

$$V = (1-0.5)0 + 0.5(2+0.5*0) = 1$$

$$Q_6(s,a) = \begin{bmatrix} 0 & -1 \\ \frac{3}{4} & \frac{9}{8} \\ -\frac{5}{8} & 1 \end{bmatrix}$$

$$V = (1-0.5)0 + 0.5(1+0.5*1) = \frac{3}{4}$$

$$Q_7(s,a) = \begin{bmatrix} \frac{3}{4} & -1 \\ \frac{3}{4} & \frac{9}{8} \\ -\frac{5}{8} & 1 \end{bmatrix}$$

$$V = (1-0.5)1 + 0.5(2+0.5*\frac{9}{8}) = \frac{57}{32}$$

$$Q_8(s,a) = \begin{bmatrix} \frac{3}{4} & -1 \\ \frac{3}{4} & \frac{9}{8} \\ -\frac{5}{8} & \frac{57}{32} \end{bmatrix}$$

$$V = (1-0.5)\frac{9}{8} + 0.5(2+0.5*\frac{3}{4}) = \frac{28}{16}$$

$$Q_9(s,a) = \begin{bmatrix} \frac{3}{4} & -1 \\ \frac{3}{4} & \frac{28}{16} \\ -\frac{5}{8} & \frac{57}{32} \end{bmatrix}$$

$$V = (1-0.5)\frac{3}{4} + 0.5(3+0.5*\frac{28}{16}) = \frac{37}{16}$$

$$Q_9(s,a) = \begin{bmatrix} \frac{37}{16} & -1 \\ \frac{3}{4} & \frac{28}{16} \\ -\frac{5}{8} & \frac{57}{32} \end{bmatrix}$$

Part b)

In a model-based reinforcement learning, we first estimate the transition function $T(s, a, s')$ and the reward function $R(s, a, s')$. Write down the estimates of T and R, estimated from the experience above. Write "n/a" if not applicable or undefined.

To figure this out we're going to reorder the above table for more clarity Here we can start calculating

s_t	a_t	s_{t+1}	r_t
A	Down	B	2
A	Down	C	-3
A	Up	B	3
A	Up	C	1
B	Down	A	2
B	Down	B	0
B	Down	C	3
B	UP	A	1
C	Down	A	2
C	Down	B	2
C	Up	B	-2

Table 2: Sorted by states and actions

the transition states by taking a given (s, a) pair and determining the probability of going to another state, s_{t+1}. We can determine the reward by normalizing.

$$T(A, Down, B) = \frac{1}{2}$$
$$T(A, Down, C) = \frac{1}{2}$$
$$R(A, Down, B) = 1$$
$$R(A, Down, C) = -\frac{3}{2}$$
$$T(A, Up, B) = \frac{1}{2}$$
$$T(A, Up, C) = \frac{1}{2}$$
$$R(A, Up, B) = \frac{3}{4}$$
$$R(A, Up, C) = \frac{1}{4}$$
$$T(B, Down, A) = \frac{1}{3}$$

$$T(B, Down, B) = \frac{1}{3}$$
$$T(B, Down, C) = \frac{1}{3}$$
$$R(B, Down, A) = \frac{2}{5}$$
$$R(B, Down, B) = 0$$
$$R(B, Down, C) = \frac{3}{5}$$
$$T(C, Down, A) = \frac{1}{2}$$
$$T(C, Down, B) = \frac{1}{2}$$
$$R(C, Down, A) = \frac{1}{2}$$
$$R(C, Down, B) = \frac{1}{2}$$
$$T(C, Up, B) = 1$$
$$R(C, Up, B) = -2$$

Part c)

Assume we had a different experience and ended up with the following estimates of the transition and reward functions

s	a	s'	$\hat{T}(s,a,s')$	$\hat{R}(s,a,s')$
A	Up	A	1	12
A	Down	B	0.5	2
A	Down	C	0.5	-3
B	Up	B	1	8
B	Down	C	1	-6
C	Down	C	1	12
C	Up	C	0.5	2
C	Up	B	0.5	-2

(i) Give the optimal policy $\hat{\pi}^*(s)$ and $\hat{V}^*(s)$ for the MDP with transition function $\hat{T}$ and reward function $\hat{R}$. Explain your answers.
Our two easiest policies are for being in states A and C where we already have the maximal reward in the MDP.
So given state A, $\hat{\pi}^*(A) =$ Up we always pick A and similarly in state C we have the policy $\hat{\pi}^*(C) =$ Down to stay in C. Where in A we will always pick Up and in state C we will always pick Down. Because they have the same reward we know that finding one will result in the other.
We have the infinite equation

$$
\begin{aligned}
V^* =& \hat{R}(s,a,s')(1+\gamma+\gamma^2+\cdots) \\
=& \hat{R}(s,a,s')\left(\frac{1}{1-\frac{1}{2}}\right) \\
=& \hat{R}(s,a,s')2 \\
=& 24
\end{aligned}
$$

This gives us the value for both A and C, where $\hat{R}(A, Up, s') = \hat{R}(C, Down, s')$.
B is a little more difficult to find, but we can see that once we get to A or C we will use the above values.
We can simply look at $\pi(B) =$ Up and see that we will always get a reward of 8, giving us $V(B, Up) = 16$, similarly to above. We need to also look at $\pi(B) =$ Down. We see that we get $-6 + \gamma V^*(C) = 6$. Here we know that $16 > 8 \therefore \hat{\pi}^*(B) =$ Up with $V^*(B) = 16$.

(ii) If we repeatedly feed this new experience sequence through our Q-learning algorithm, what values will it converget to? Assume that convergence is guarenteed.