AI3000: Assignment 2 (theory)

Raj Patil - CS18BTECH11039

2021-10-22 Fri

Contents

1 Model Free Prediction and Control

1.1 (A)

$$
\begin{aligned}
V(A) &= \frac{14+15+17+16+15}{5} = \frac{77}{5} = 15.4 \\
V(B) &= \frac{13+14+16+15+14}{5} = \frac{72}{5} = 14.4 \\
V(C) &= \frac{12+13+15+14+13}{5} = \frac{67}{5} = 13.4 \\
V(D) &= \frac{12+12+12+11}{4} = \frac{47}{4} = 11.75 \\
V(E) &= \frac{11+11+11+10+9}{5} = \frac{52}{5} = 10.4 \\
V(F) &= \frac{10+10+10+10+9}{5} = \frac{49}{5} = 9.8 \\
V(G) &= \frac{0+0+0+0+0}{5} = 0
\end{aligned}
$$

1.2 (B)

The states that are visited more than once in at least one episode exhibit a possibility of being assigned a different $V(s)$. In this case, those states would be:

$$B, C, E, F$$

The reason being attributed to the counter being incremented more number of times in the every-visit case compared to that of the first visit case. Note that the value for these states may still be the same (depends on the rewards received) but a possibility exists that they will be different (improbable for the states that occur less than once in the all episodes).

1.3 (C)

Given the policy π_f that always moves forward, the true value (computing backward in the chain) of the states would be :

$$
\begin{aligned}
&V^{\pi_f}(G) = 0 \\
&V^{\pi_f}(F) = 10 + V^{\pi_f}(G) = 10 \\
&V^{\pi_f}(E) = 1 + V^{\pi_f}(F) = 11 \\
&V^{\pi_f}(D) = 1 + V^{\pi_f}(E) = 12 \\
&V^{\pi_f}(C) = 0.5 * (1 + V^{\pi_f}(D)) + 0.5 * (4 + V^{\pi_f}(E)) = \frac{13+15}{2} = 14 \\
&V^{\pi_f}(B) = 1 + V^{\pi_f}(C) = 15 \\
&V^{\pi_f}(A) = 1 + V^{\pi_f}(B) = 16
\end{aligned}
$$

1.4 (D)

Given the episodes 2,3 note that we only need to obtain the most likely transition probabilities when one chooses to jump right from C. This is because the rest of the MDP (Transition Probabilities and Rewards) stays the same (deterministic). Out of the three episodes:

1. $C \rightarrow E$ occurs once (in 3)

2. $C \to D$ occurs once (in 2)

Hence,

$$MLE(\mathcal{P}^{jump}_{CD}) = \frac{1}{2}$$

$$MLE(\mathcal{P}^{jump}_{CE}) = \frac{1}{2}$$

Consequently, the answer will be the same:

$$\begin{aligned}
&V^{\pi_f}(G) = 0\\
&V^{\pi_f}(F) = 10 + V^{\pi_f}(G) = 10\\
&V^{\pi_f}(E) = 1 + V^{\pi_f}(F) = 11\\
&V^{\pi_f}(D) = 1 + V^{\pi_f}(E) = 12\\
&V^{\pi_f}(C) = 0.5 * (1 + V^{\pi_f}(D)) + 0.5 * (4 + V^{\pi_f}(E)) = \frac{13 + 15}{2} = 14\\
&V^{\pi_f}(B) = 1 + V^{\pi_f}(C) = 15\\
&V^{\pi_f}(A) = 1 + V^{\pi_f}(B) = 16
\end{aligned}$$

1.5 (E)

By Law of Large Numbers, the MC method will definitely converge to the true V^{π_f} as it is an unbiased estimator. TD(0) method however exhibits some bias due to the initial bootstrap. However, that bias should vanish asymptotically and one should reach at the same conclusion as with the MC method. This is due to the fact that the actual V^{π_f} will be a fixed point for the TD update (even for still significant learning rate) and over time one should converge to that. Hence, yes, they should converge to the same true value function

1.6 (F)

given the following transitions:

s	a	r	s	a	r	s	a	r	s	a	r	s
C	jump	4	E	right	1	F	left	-2	E	right	1	F

The updates are as follows (given $\alpha = 0.5, \gamma = 1$) i.e. each update will be the average of the old and the newly received recommendation

	Q(C,left)	Q(C,jump)	Q(E,left)	Q(E,right)	Q(F,left)	Q(F,right)
Initial	0	0	0	0	0	0
Transition 1	0	(updated)2	0	0	0	0
Transition 2	0	2	0	(updated)0.5	0	0
Transition 3	0	2	0	0.5	(updated)-0.75	0
Transition 4	0	2	0	(updated)0.5	-0.75	0

1.7 (G)

Constructing a greedy $\pi : \{C, E, F\} \to \{left, right, jump\}$ using the following:

$$\pi(S) = \underset{a \in \mathcal{A}}{argmax} Q(S, a)$$

the greedy policy emerges as :

state	prescribed action
C	jump
E	right
F	right

2 On Learning Rates

Drawing upon the p-series test for convergence, with the intuition for the same: For a series

$$\sum_{t=1}^{\infty} \frac{1}{t^p}$$

One can lower bound these by using the intuitive definition of an area under the curve. Note that the above sum will be greater than the corresponding degenerated integral(setting dt = 1 and considering areas of rectangles of unit width)

$$\sum_{t=1}^{\infty} \frac{1}{t^p} = \frac{1}{1^p} \cdot 1 + \frac{1}{2^p} \cdot 1 \cdots > \int_{t=1}^{\infty} \frac{1}{t^p} dt$$

where the term on the left represents a coarse upper bound for width of the approximation rectangles being set to one with their height as the value of the function. Now given $p \in \mathbb{R}^+$, three cases arise:

2.1 $p \in (0, 1)$

$$\int_{t=1}^{\infty} \frac{1}{t^p} dt = [\frac{t^{1-p}}{(1-p)}]_1^{\infty} = \infty$$

hence for $p < 1$ the series diverges

2.2 $p = 1$

$$\int_{t=1}^{\infty} \frac{1}{t} dt = [ln(t)]_1^{\infty} = \infty$$

hence for $p = 1$ the series diverges

2.3 $p \in (1, \infty)$

$$\int_{t=1}^{\infty} \frac{1}{t^p} dt = [\frac{t^{1-p}}{(1-p)}]_1^{\infty} = \frac{1}{p-1} < \infty$$

hence for p $>$ 1 the series converges

2.4 Application

2.4.1 (1)

$p = 1, 2p = 2 :$ WILL converge

2.4.2 (2)

$p = 2, 2p = 4 :$ WON'T converge

2.4.3 (3)

$p = \frac{2}{3}, 2p = \frac{4}{3} :$ WILL converge

2.4.4 (4)

$p = \frac{1}{2}, 2p = 1$: WON'T converge

3 Q-Learning

3.1 (A)

Given these are for a finite-horizon MDP with horizon 1, γ is of no importance

$$V^*(s) = \mathbb{E}[G_t|s_t = s] = \mathbb{E}[r_{t+1}|s_t = s] = c$$

(as expectation is given the same for both actions)

$$Q(s, a_1) = \mathbb{E}[G_t|s_t = s, a_t = a_1] = c$$

$$Q(s, a_2) = \mathbb{E}[G_t|s_t = s, a_t = a_2] = c$$

3.2 (B)

For the n independent episodes observed, each action was taken for equal number of times (given). The MLE estimate for the Q-function will be as follows: Note that the sum is over only the rewards corresponding to that particular action.

$$\hat{Q}(s, a_1) = \frac{2}{n} \sum_{i:a^i=a_1} r_i$$

$$\hat{Q}(s, a_2) = \frac{2}{n} \sum_{i:a^i=a_2} r_i$$

Now $\hat{\pi}$ is defined as

$$\hat{\pi}(s) = \underset{a}{argmax}\ \hat{Q}(s, a)$$

then $\hat{V}^{\hat{\pi}}$ is given by

$$\hat{V}^{\hat{\pi}}(s) = \underset{a}{max}\ \hat{Q}(s, \hat{\pi}(s)) = \underset{a}{max}\ \hat{Q}(s, \underset{a}{argmax}\ \hat{Q}(s, a)) = \underset{a}{max}\hat{Q}(s, a)$$

$$\therefore \hat{V}^{\hat{\pi}} = max\{\hat{Q}(s, a_1), \hat{Q}(s, a_2)\}$$

recalling that for an estimator $\hat{\theta}$ of θ, the bias $b(\hat{\theta}, \theta)$ is defined as:

$$b(\hat{\theta}, \theta) = \mathbb{E}_{obs|\theta}[\hat{\theta} - \theta] = \mathbb{E}_{obs|\theta}[\hat{\theta}] - \theta$$

i.e. the expectation of the difference between the true and estimated value. here the estimator is $\hat{V}^{\hat{\pi}}(s) = max(\hat{Q}(s, a_1), \hat{Q}(s, a_2))$ and $V^*(s) = c$. Also note that $\mathbb{E}[\hat{Q}(s, a_1)] = \mathbb{E}[\hat{Q}(s, a_2)] = \frac{2}{n} \cdot \frac{n}{2} \cdot c = c$

Now,

$$max(\hat{Q}(s, a_1), \hat{Q}(s, a_2)) \geq \hat{Q}(s, a_1)$$

$$max(\hat{Q}(s, a_1), \hat{Q}(s, a_2)) \geq \hat{Q}(s, a_2)$$

adding and dividing by two:

$$max(\hat{Q}(s, a_1), \hat{Q}(s, a_2)) \geq \frac{\hat{Q}(s, a_1) + \hat{Q}(s, a_2)}{2}$$

$$\therefore \mathbb{E}[max(\hat{Q}(s, a_1), \hat{Q}(s, a_2))] \geq \mathbb{E}[\frac{\hat{Q}(s, a_1) + \hat{Q}(s, a_2)}{2}] = \frac{c + c}{2} = c$$

$$\therefore \mathbb{E}[max(\hat{Q}(s,a_2),\hat{Q}(s,a)2))] - c \geq 0$$

$$\therefore \mathbb{E}[\hat{V}^{\hat{\pi}}(s)] - V^*(s) \geq 0$$

$$\therefore b(\hat{V}^{\hat{\pi}}, V^*) \geq 0$$

That is $\hat{V}^{\hat{\pi}}$ is a biased estimator of V^*, hence shown: ■

3.3 (C)

a_1 is the better action in expectation (using Linearity of Expectation)as :

$$\mathbb{E}[r|a_1] = c > c - 0.2 = \mathbb{E}[r|a_2]$$

Given finite examples, Temporal difference based methods may favor a_2 as well as one can receive, with non-zero probability , rewards greater than c. If this happens more than half the time action 2 was chosen (which is highly probable given finite examples), a_2 will be suggested as the optimal action by the algorithms. So $\boxed{NO}$, TD methods might not favor the methods that are best in expectation always for finite examples but should do so asymptotically.

4 Importance Sampling

Note that as this is a single state MDP, $\pi(a)$ has been considered equivalent to $\pi(a|\cdot)$

4.1 (A)

For a single sample (a,r):

$$\hat{V}^{\pi} = \mathbb{E}_{\pi_b}[\mathcal{R}^a \cdot \frac{\pi(a)}{\pi_b(a)}] = \frac{r \cdot \pi(a)}{\pi_b(a)}$$

Note that this will be an unbiased estimate of V^{π} as can be verified by considering its expectation which is

$$\mathbb{E}_{a\sim\pi_b}[r \cdot \frac{\pi(a)}{\pi_b(a)}] = \mathbb{E}_{a\sim\pi}[r]$$

which is by definition equal to V^{π}

$$\therefore \mathbb{E}[\hat{V}^{\pi}] - V^{\pi} = 0$$

i.e. it is an unbiased estimate

4.2 (B)

$$\mathbb{E}_{\pi_b}[\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)}] = \sum_{a\in\mathcal{A}}(\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)} \cdot \pi_b(a|\cdot)) = \sum_{a\in\mathcal{A}}\pi(a|\cdot) = 1$$

4.3 (C)

For π_b being uniform stochastic and π being deterministic, The I.S. ratio is given by:

$$I.S.(a) = \frac{\pi(a)}{\pi_b(a)} = \begin{cases} K, & \pi(a) = 1 \\ 0, & otherwise \end{cases}$$

4.4 (D)

Given that the reward is a deterministic constant for all the actions, for the sampling policy being uniform; introducing a shorthand for the importance sampling ratio for an action, for convenience:

$$\rho(a) \triangleq \frac{\pi(a)}{\pi_b(a)}$$

$$\begin{aligned} var_{a\sim\pi_b}(\rho(a)\cdot r) &= r^2 \cdot var_{a\sim\pi_b}(\rho(a)) \\ &= r^2(\mathbb{E}_{a\sim\pi_b}[(\rho(a))^2] - (\mathbb{E}_{a\sim\pi_b}[\rho(a)])^2) \end{aligned}$$

Note that the second term is the same asked in part (b) of this question and was found to be 1. For the first term, using what was found in part (c) of this question: $\rho(a)$ takes value K for exactly one action i.e. with probability $\frac{1}{K}$.

$$\therefore \mathbb{E}[\rho(a)^2] = \frac{1}{K}\cdot K^2 + 0\cdot\frac{K-1}{K} = K$$

substituting these in the above expression

$$\begin{aligned} var_{a\sim\pi_b}(\rho(a)\cdot r) &= r^2(K-(1)^2) \\ &= r^2(K-1) \end{aligned}$$

4.5 (E)

For the general case, using the following:

$$var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$$

$$\therefore var(\rho r) < \mathbb{E}_{a\sim\pi_b}(\rho^2 r^2) \leq \mathbb{E}_{a\sim\pi_b}(\rho^2) = K^2\cdot\frac{1}{K} = K$$

$$\boxed{\therefore var(\frac{\pi(a)}{\pi_b(a)}\cdot r) \leq K}$$

4.6 (F)

For τ representing a trajectory, the joint probability of certain actions evolving given a policy π would be,

$$\pi(a_0, a_1, a_2 \ldots a_\infty | s_0, s_1, s_2 \ldots s_\infty)$$

Using Markov Property, this can be written as:

$$\pi(a_0|s_0)\pi(a_1|s_1)\ldots\pi(a_\infty|s_\infty) = \prod_{i=0}^{\infty}\pi(a_i|s_i)$$

$$\boxed{\therefore \frac{P(\tau)}{Q(\tau)} = \prod_{i=0}^{\infty}\frac{\pi_b(a_i|s_i)}{\pi(a_i|s_i)}}$$