Indian Institute of Technology Kharagpur
Quiz 02 2021-22

Date of Examination: **27 Aug. 2021** **Duration:** **30 minutes**
Subject No.: **CS60077** **Subject:** **Reinforcement Learning**
Department/Center/School: **Computer Science** **Credits:** **3** **Full marks:** **20**

Instructions

i. This question paper contains 3 pages and 3 questions. All questions are compulsory. Marks are indicated in parentheses. This question paper has been cross checked.
ii. Please write your name, roll number and date on top of the answer script.
iii. **Organize your work**, in a reasonably neat and coherent way. Work scattered all across the answer script without a clear ordering will receive very little marks.
iv. **Mysterious or unsupported answers will not receive full marks**. A correct answer, unsupported by calculations, explanation, will receive no marks; an incorrect answer supported by substantially correct calculations and explanations may receive partial marks.
v. In the online setting, you need to upload your answer scripts as **pdf file**. We will prefer a single pdf file. If you happen to have multiple files, please zip them and then uplaod as a single file. You can scan your worked out example or you can use latex to produce the pdf.

---

1. (a) (2 points) The Markov inequality for positive random variables is given by $P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$. Use this inequality to prove the following.

$$P\big[|X - \mathbb{E}[X]| \geq \varepsilon\big] \leq \frac{\sigma^2}{\varepsilon^2} \quad \forall \varepsilon \tag{1}$$

where σ^2 is the variance of X.

The inequality to prove is the Chebyshev inequality. It is a consequence of the Markov inequality. Let $D^2 = |X - \mathbb{E}[X]|^2$ be the squared deviation of X from the mean. Then Markov inequality applied to D with $a = \varepsilon^2$ gives,

$$P(D^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}[D^2]}{\varepsilon^2} = \frac{\mathbb{E}\big[|X - \mathbb{E}[X]|^2\big]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \quad \forall \varepsilon \tag{2}$$

Eqn. (1) follows when we note that $\{D^2 \geq \varepsilon^2\}$ and $|X - \mathbb{E}[X]| \geq \varepsilon$ are equivalent events.

(b) (2 points) Consider a finite, episodic and undiscounted MDP with states P and Q apart from the terminal state. Let the following two samples are observed when a Monte-Carlo evaluation is being carried out. For example a sample such as, $(P, +2) \to (P, +3) \to (Q, -2)$ means that the episode starts at P then goes to P again, then goes to Q and then terminates. On the way , the agent gets rewards of $+2, +3$ and -2 respectively.

- $(P, +2) \to (P, +3) \to (Q, -2) \to (P, +5) \to (Q, -3)$
- $(Q, -2) \to (P, +3) \to (Q, -3)$

Estimate the state value of both P and Q using first-visit Monte-Carlo evaluation.

The first visit return of state P in the first trajectory is $2 + 3 - 2 + 5 - 3 = 5$. The same for the second trajectory is $3 - 3 = 0$. So, the first visit MC estimate of the state-value of P is $\frac{5+0}{2} = \frac{5}{2}$.

The first visit return of state Q in the first trajectory is $-2 + 5 - 3 = 0$. The same for the second trajectory is $-2 + 3 - 3 = -2$. So, the first visit MC estimate of the state-value of Q is $\frac{0-2}{2} = -1$.

(c) (2 points) Estimate the state value of both P and Q using every-visit Monte-Carlo evaluation for the above problem.

There are 3 visits of the state P in the first trajectory with the corresponding returns of $2+3-2+5-3=5, 3-2+5-3=3$ and $5-3=2$. There is only one visit of the state P in the second trajectory. The return for the second trajectory thus, is $3-3=0$. So, the every visit MC estimate of the state-value of P is $\frac{5+3+2+0}{4}=\frac{5}{2}$.

There are 2 visits of the state Q in the first trajectory with the corresponding returns of $-2+5-3=0$ and $-3=-3$. There are 2 visits of the state Q in the second trajectory. The returns for the second trajectory thus, are $-2+3-3=-2$ and $-3=-3$. So, the every visit MC estimate of the state-value of Q is $\frac{0-3-2-3}{4}=-2$.

(d) (2 points) In the space of real numbers and considering infinity norm prove that the function $\mathcal{T}(\mathbf{x})=\frac{\mathbf{x}}{2}$ is a contraction mapping.

$||\mathcal{T}(\mathbf{x})-\mathcal{T}(\mathbf{y})||_\infty=||\frac{\mathbf{x}}{2}-\frac{\mathbf{y}}{2}||_\infty=\frac{1}{2}||\mathbf{x}-\mathbf{y}||_\infty$, which means $||\mathcal{T}(\mathbf{x})-\mathcal{T}(\mathbf{y})||_\infty \le \lambda||\mathbf{x}-\mathbf{y}||_\infty$ for any $\frac{1}{2}<\lambda<1$. Thus this is a contraction mapping.

2. (7 points) You want to compute the mean of $f(X)$ where X be a continuous random variable and $p(X)$ be its probability distribution function. Let $q(X)$ be another distribution such that $q(X)=0 \Rightarrow f(X)p(X)=0$. What is the importance sampling estimator $\hat{\mu}_q$ for $\mu=\mathbb{E}_p[f(X)]$? Show that the expected value of $\hat{\mu}_q$ is μ *i.e.*, $\mathbb{E}_q[\hat{\mu}_q]=\mu$. Also compute the variance of $\hat{\mu}_q$. You can assume the number of samples drawn is n.

Let D be the domain where $p(X)\neq 0$. Further, $q(X)=0$ implies $f(X)p(X)=0$. Define $Q=\{x: q(X)>0\}$. The mean of $f(X)$ is given by $\mathbb{E}_p[f(X)]=\mu=\int_D f(X)p(X)dX$.

The importance sampling estimator $\hat{\mu}_q$ is given by,

$$\hat{\mu}_q=\frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)p(X_i)}{q(X_i)}, \quad X_i\sim q \tag{3}$$

Computing the mean of $\hat{\mu}_q$,

$$\begin{aligned}
\mathbb{E}_q[\hat{\mu}_q] &= \mathbb{E}_q\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)p(X_i)}{q(X_i)}\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_q\Big[\frac{f(X_i)p(X_i)}{q(X_i)}\Big] \\
&= \mathbb{E}_q\Big[\frac{f(X_1)p(X_1)}{q(X_1)}\Big] \quad (\ X_i \text{ are i.i.d so } \mathbb{E}[\textstyle\sum g(X_i)]=\sum\mathbb{E}[g(X_i)]\ ) \\
&= \int_Q \frac{f(X)p(X)}{q(X)}q(X)dX \\
&= \int_Q f(X)p(X)dX \\
&= \int_D f(X)p(X)dX+\int_{Q\cap D^c} f(X)p(X)dX-\int_{D\cap Q^c} f(X)p(X)dX \\
&= \int_D f(X)p(X)dX=\mu \quad (\text{In } Q\cap D^c, p(X)=0 \text{ and in } D\cap Q^c, f(X)=0\ )
\end{aligned}$$

Hence $\hat{\mu}_q = \mu$ or $\hat{\mu}_q$ is an unbiased estimator.

Calculating $\text{Var}(\hat{\mu}_q)$,

$$\begin{aligned}
Var(\hat{\mu}_q) &= Var\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)p(X_i)}{q(X_i)}\Big) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var\Big(\frac{f(X_i)p(X_i)}{q(X_i)}\Big) \quad (\ X_i \text{ are i.i.d so } \mathbb{E}[\textstyle\sum g(X_i)] = \sum \mathbb{E}[g(X_i)]\ ) \\
&= \frac{1}{n} Var\Big(\frac{f(X_1)p(X_1)}{q(X_1)}\Big) \\
&= \frac{1}{n}\mathbb{E}_q\Big[\Big(\frac{f(X_1)p(X_1)}{q(X_1)}\Big)^2\Big] - \mathbb{E}_q\Big[\frac{f(X_1)p(X_1)}{q(X_1)}\Big]^2 \\
&= \frac{1}{n}\Big[\int_Q \frac{(f(X)p(X))^2}{q(X)}dX - \mu^2\Big]
\end{aligned}$$

Therefore, $Var(\hat{\mu}_q) = \frac{1}{n}\Big[\int_Q \frac{(f(X)p(X))^2}{q(X)}dX - \mu^2\Big]$.

3. (5 points) Let T^π be a modified form of the Bellman Operator for action value function defined as follows,

$$T^\pi Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}\sim p}[V(s_{t+1})] \tag{4}$$

where a modified form of the state value function is given by,

$$V(s_t) = \mathbb{E}_{a_t\sim\pi}[Q(s_t, a_t) - \log\pi(a_t|s_t)] \tag{5}$$

Let us consider evaluating a fixed policy π using iterative application of this modified Bellman operator *i.e.*, applying $Q^{k+1} = T^\pi Q^k$ starting with some arbitrary Q^k at $k = 0$. Prove that the sequence Q^k will converge as $k \to \infty$. Assume $|\mathcal{A}| < \infty$.

Hint: Trying to see if the modified Bellman operator can be reduced to a known contraction mapping.

From the given definitions, the operator T^π can be written as,

$$T^\pi Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}\sim p, a_{t+1}\sim\pi}[Q(s_{t+1}, a_{t+1}) - \log\pi(a_{t+1}|s_{t+1})] \tag{6}$$

Taking $r'(s_t, a_t) = r(s_t, a_t) - \mathbb{E}_{s_{t+1}\sim p, a_{t+1}\sim\pi}[\log\pi(a_{t+1}|s_{t+1})]$, we can rewrite the operator T^π as,

$$T^\pi Q(s_t, a_t) = r'(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}\sim p, a_{t+1}\sim\pi}[Q(s_{t+1}, a_{t+1})] \tag{7}$$

For policy iteration, π is constant and since $|A| < \infty$, $r'(s_t, a_t)$ is bounded. Although not important for the solution but interesting to note, $-\mathbb{E}_{a_{t+1}\sim\pi}[\log\pi(a_{t+1}|s_{t+1})]$ is the definition of entropy i.e. $H(\pi(\cdot|s_{t+1})$.

Equation 7 is now the standard Bellman operator for policy evaluation using Q function. Hence can be easily shown to be a contraction.