1. **[CA Session] Last Visit Monte Carlo**

Prove that last visit Monte Carlo is not guaranteed to converge almost surely to $V^\pi$ for all finite MDPs with bounded rewards and $\gamma \in [0, 1]$. You may reference Khinchine's Strong Law of Large Numbers:

**[Khintchine Strong Law of Large Numbers]**

Let $\{X_i\}_{i=1}^{\infty}$ be independent and identically distributed random variables. Then $(\frac{1}{n} \sum_{i=1}^{n} X_i)_{n=1}^{\infty}$ is a sequence of random variables that converges almost surely to $\mathbb{E}[X_1]$.

**Solution** Define the MDP with $S = \{s_1, s_{end}\}$, $A = \{a_1\}$, $P(s_1, a_1, s_{end}) = 0.5$, $P(s_1, a_1, s_1) = 0.5$, $R_t = 1$ if $S_{t+1} = s_1$, and $R_t = 0$ if $S_{t+1} = s_{end}$. The starting state is $s_1$, $s_{end}$ is a terminal state, and $\gamma = 0.5$. Let $\pi$ be the only policy that always selects action $a_1$. Notice that $V^\pi(s_1) = 1$.

Last visit Monte Carlo computes returns from state $s_1$ and since it only uses the last visit, the returns from state $s_1$ will all be for the transition to $s_{end}$, where the return is zero. Thus, the last visit Monte Carlo estimate for $V^\pi(s_1)$ after $n$ episodes will be $\frac{1}{n} \sum_{i=1}^{n} 0$.

Hence, $\lim_{n \to \infty} \hat{V}^\pi(s_1) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} 0 = 0$.

Finally, $\Pr[\lim_{n \to \infty} \hat{V}^\pi(s_1) = V^\pi(s_1)] = \Pr[\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} 0 = 1] = 0$. 

---

1
Consider a finite MDP with bounded rewards, $M = (S, A, R, P, \gamma)$. Let $\gamma < 1$. Let $\pi^*$ be a deterministic optimal policy for this MDP. Let $M' = (S', A', R', P', \gamma')$ be a new MDP that is the same as $M$, except that a positive constant, $c$, is subtracted from $R_t$ if $A_t$ is not the action that $\pi^*$ would select. Is $\pi^*$ necessarily always an optimal policy for $M'$? Prove your answer. If it is not, prove that it is not, and if it is, prove that it is.

Solution

First, notice that for all $s \in S$ and $a \in A$, $R(s, a) \geq R'(s, a)$. Notice for all $s \in S$:

$$V_{M}^{\pi^*}(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = s, \pi^*, M \right]$$

$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = s, \pi^*, M' \right]$$

$$= V_{M'}^{\pi^*}(s), \text{ because when } A_t \sim \pi^*, R_t \text{ is unchanged.}$$

Next, we see for all $\pi$ and for all $s \in S$,

$$V_{M}^{\pi}(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = s, \pi, M \right]$$

$$= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R(S_{t+k}, A_{t+k}) | S_t = s, \pi, M \right]$$

$$\geq \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R'(S_{t+k}, A_{t+k}) | S_t = s, \pi, M' \right]$$

$$= V_{M'}^{\pi}(s).$$

Finally, notice that because $\pi^*$ is optimal in $M$, we have that for all $\pi$ and $s \in S$, $V_{M}^{\pi^*}(s) \geq V_{M}^{\pi}(s)$.

Combining equations, we have that for all $\pi$ and $s \in S$,

$$V_{M}^{\pi^*}(s) = V_{M}^{\pi^*}(s) \geq V_{M}^{\pi}(s) \geq V_{M'}^{\pi}(s).$$

Thus, $\pi^* \geq \pi$ for all policies $\pi$, and therefore $\pi^*$ is optimal in $M'$. 

Questions 1 and 2 are borrowed from Phil Thomas.  

---

Footnote: 1 https://people.cs.umass.edu/~pthomas/courses/CMPSCI_687_Fall2018/687_F18_main.pdf
3) [Breakout Rooms] Bellman Operator with Function Approximation

Consider an MDP $M = (S, A, R, P, \gamma)$ with finite discrete state space $S$ and action space $A$. Assume $M$ has dynamics model $P(s'|s, a)$ for all $s, s' \in S$ and $a \in A$ and reward model $R(s, a)$ for all $s \in S$ and $a \in A$.

Recall that the Bellman operator $B$ applied to a function $V : S \to \mathbb{R}$ is defined as

$$B(V)(s) = \max_a (R(s, a) + \gamma \sum_{s'} P(s'|s, a)V(s'))$$

(8)

(a) Now, consider a new operator which first applies a Bellman backup and then applies a function approximation, to map the value function back to a space representable by the function approximation. We will consider a linear value function approximator over a continuous state space. Consider the following graphs:

The graphs show linear regression on the sample $X_0 = \{0, 1, 2\}$ for hypothetical underlying functions. On the left, a target function $f$ (solid line), that evaluates to $f(0) = f(1) = f(2) = 0$ and its corresponding fitted function $\hat{f}(x) = 0$. On the right, another target function $g$ (solid line) that evaluates to $g(0) = 0$ and $g(1) = g(2) = 1$, and its fitted function $\hat{g}(x) = \frac{7}{12}x$.

What happens to the distance between points $\{f(0), f(1), f(2)\}$ and $\{g(0), g(1), g(2)\}$ after we do the linear approximation? In other words, compare $\max_{x \in X_0} |f(x) - g(x)|$ and $\max_{x \in X_0} |\hat{f}(x) - \hat{g}(x)|$.

**Solution** We compute $\max_{x \in X_0} |f(x) - g(x)| = 1$ and $\max_{x \in X_0} |\hat{f}(x) - \hat{g}(x)| = \frac{7}{6}$. Note $\max_{x \in X_0} |f(x) - g(x)| < \max_{x \in X_0} |\hat{f}(x) - \hat{g}(x)|$. The distance between the points increases after the linear approximation.
(b) Is the linear function approximator here a contraction operator? Explain your answer.

**Solution**  Let $L$ be the linear approximation operator such that $\hat{f} = Lf$ and $\hat{g} = Lg$. From part a), we see that $||Lf - Lg||_\infty > ||f - g||_\infty$ where $||\cdot||_\infty$ is the infinity norm. Then, the linear function approximator $L$ is not a contraction operator.

(c) Will the new operator be guaranteed to converge to a single value function? If yes, will this be the optimal value function for the problem? Justify your answers.

**Solution**  While the Bellman operator $B$ is a contraction operator, the composite operator $L \circ B$ that first applies a Bellman backup and then the linear approximation is not necessarily a contraction operator because the linear function approximator $L$ is not a contraction operator. Since we do not have the contraction property, the composite operator does not necessarily converge.