CS234 Problem Session

Week 4: Feb 2

1) [CA Session] Last Visit Monte Carlo

Prove that last visit Monte Carlo is not guaranteed to converge almost surely to V^π for all finite MDPs with bounded rewards and $\gamma \in [0, 1]$. You may reference Khintchine'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 $\mathbb{E}[X_1]$.

2) [CA Session] Optimal Policy in Modified MDP

Consider a finite MDP with bounded rewards, $M = (\mathcal{S}, \mathcal{A}, R, P, \gamma)$. Let $\gamma < 1$. Let π^* be a deterministic optimal policy for this MDP. Let $M' = (\mathcal{S}', \mathcal{A}', R', P', \gamma')$ 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 π^* would select. Is π^* 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.

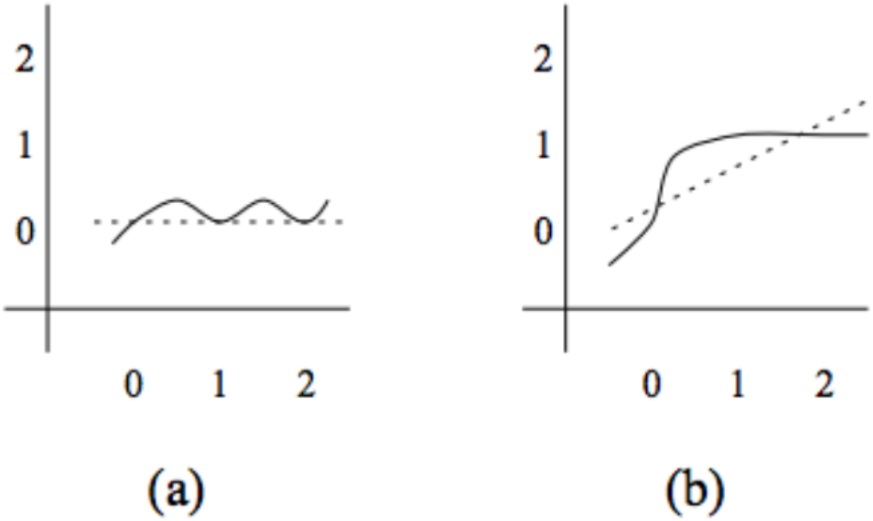
3) [Breakout Rooms] Bellman Operator with Function Approximation

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

$$B(V)(s) = max_a(R(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')) \tag{1}$$

(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:



(a)

(b)

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 $max_{x \in X_0}|\hat{f}(x) - \hat{g}(x)|$.

(b) Is the linear function approximator here a contraction operator? Explain your answer.

(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.