AI3000 – ASSIGNMENT 2

VAISHNAVI
AI20BTECH11025

1a. $\|V_{k+1} - V_k\|_\infty \le \epsilon$ for $\epsilon > 0$

Consider $\|V_k - V^\pi\|_\infty$

$$\|V_k - V^\pi\|_\infty \le \underbrace{\|V_k - V_{k+1}\|_\infty}_{\epsilon} + \|V_{k+1} - V^\pi\|_\infty$$

$$\le \epsilon + \|(R + \gamma P V_k) - (R + \gamma P V^\pi)\|_\infty$$

$$\le \epsilon + \gamma \|P(V_k - V^\pi)\|$$

$$\le \underbrace{\|P\|}_{\le 1} \cdot \|V_k - V^\pi\|$$

[$\because \|AB\| \le \|A\| \, \|B\|$]

$$\le \epsilon + \gamma \|V_k - V^\pi\|$$

$$\Rightarrow \|V_k - V^\pi\| \le \frac{\epsilon}{1-\gamma}$$

Consider $\|V_{k+1} - V^\pi\|_\infty = \|(R + \gamma P V_k) - (R + \gamma P V^\pi)\|_\infty$

$$\le \gamma \|P(V_k - V^\pi)\|$$

$$\le \underbrace{\|P\|}_{\le 1} \, \|V_k - V^\pi\|$$

$$\le \gamma \|V_k - V^\pi\|_\infty$$

$$= \frac{\gamma\epsilon}{1-\gamma}$$

1b. $\|V_{k+1} - V^\pi\|_\infty = \|(R + \gamma P V_k) - (R + \gamma P V^\pi)\|_\infty$

$$\le \gamma \|V_k - V^\pi\|_\infty$$

|||ly $\|V_k - V^\pi\| \le \gamma \|V_{k-1} - V^\pi\|_\infty$

$\vdots$

$\|V_2 - V^\pi\| \le \gamma \|V_1 - V^\pi\|_\infty$

$$\Rightarrow \|V_{k+1} - V^\pi\| \le \gamma^k \|V_1 - V^\pi\|$$

1c. Bellman optimality operator

$$L(V) = \max_{a \in A} \left[R^a + \gamma P^a V \right]$$

For value function 'u' let a_1 - optimal action at s - state

$$L(u) = \left[R(s, a_1) + \gamma \sum_{s'} P(s' \mid s, a_1) \cdot u(s) \right]$$

Value function V $\quad a_2$ - Optimal action at state s

$$\begin{aligned} L(V) &= \left[R(s, a_2) + \gamma \sum_{s'} P(s' \mid s, a_2) \cdot V(s) \right] \\ &\geq \left[R(s, a_1) + \gamma \sum_{s'} P(s' \mid s, a_1) V(s) \right] \end{aligned}$$

$$\Rightarrow \quad L(u) - L(V) \leq \underbrace{\gamma}_{\geq 0} \sum_{s'} \underbrace{P(s' \mid s, a_1)}_{\geq 0} \underbrace{\left[u(s) - v(s) \right]}_{\leq 0}$$

$$\left[\begin{aligned} &\because\ u(s) \leq v(s) \ \text{(given)} \\ &\Rightarrow u(s) - v(s) \leq 0 \end{aligned} \right]$$

$$\Rightarrow \quad L(u) - L(V) \leq 0$$

$$L(u) \leq L(V)$$

Hence, bellman optimality operator (L) satisfies the monotonicity property.

2. On contractions

a) $\exists$ P, Q are contractions on normed vector space $\langle V, \|\cdot\| \rangle$

~~$\Rightarrow \| P(u) - P(v) \| \leq$~~

$\Rightarrow$ $\exists$ $\gamma_p, \gamma_q \in [0,1)$ s.t

$$\| P(u) - P(v) \| \leq \gamma_p \| u - v \|$$

$$\| Q(u) - Q(v) \| \leq \gamma_q \| u - v \| \quad \forall\ u, v \text{ in } V$$

- Composition $P \circ Q$

$$\| P \circ Q(u) - P \circ Q(v) \| \leq \gamma_p \| Q(u) - Q(v) \|$$

$$\leq \gamma_p \cdot \gamma_q \| u - v \|$$

$\gamma_p \in [0,1)$, $\gamma_q \in [0,1)$

$\Rightarrow \gamma_p \cdot \gamma_q \in [0,1)$

$\therefore$ $P \circ Q$ is a contraction in the same normed vector space

- Composition $Q \circ P$

$$\| Q \circ P(u) - Q \circ P(v) \| \leq \gamma_q \| P(u) - P(v) \|$$

$$\leq \gamma_q \cdot \gamma_p \| u - v \|$$

$\gamma_q \in [0,1)$, $\gamma_p \in [0,1)$

$\Rightarrow \gamma_q \cdot \gamma_p \in [0,1)$

$\therefore$ $Q \circ P$ is a contraction in the same normed vector space.

b) From above, we have

Composition $P \circ Q$: $\| P \circ Q(u) - P \circ Q(v) \| \leq \gamma_p \cdot \gamma_q \| u - v \|$

$\Rightarrow$ $\gamma_p \cdot \gamma_q$ is the suitable Lipschitz coeff.

Composition $Q \circ P$: $\| Q \circ P(u) - Q \circ P(v) \| \leq \gamma_p \cdot \gamma_q \| u - v \|$

$\Rightarrow$ $\gamma_p \gamma_q$ is the suitable Lipschitz coeff.

2c For the operator to converge to a unique solution, the operator $F \circ L$ must be a contraction.

From (a), we have that a composition is a contraction if both the functions F & L are contractions.

$\Rightarrow$ $F \circ L$ converges to unique sol when F, L are contractions under max-norm.

~~2a.~~

Q3.

$$\begin{tikzcd} S \arrow[loop above, "1-p"] \arrow[r, "p"'] & A \end{tikzcd}$$

3a. Typical trajectory starting at S : SS.....A

b. For a trajectory of length $l+1$ ie l - 'S's

$$V(S)_{\text{first visit MC estimate}} = l$$

c.
$$V(S)_{\text{every visit MC estimate}} = \frac{l + l-1 + l-2 + \cdots 1}{l}$$
$$= \frac{l(l+1)}{2l} = \frac{l+1}{2}$$

d.
$$V(A) = 0$$
$$V(S) = 1 + p \cdot \cancel{V(A)} + (1-p)\, V(S)$$
$$\Rightarrow V(S)[1 - 1 + p] = 1$$
$$V(S) = \frac{1}{p}$$

e. Every visit MC is biased as not all returns are not iid

Proof :

Expected length of episode $= p + (1-p)p(2) + (1-p)^2 p(3) + \cdots]$

$= p[1 + (1-p)2 + (1-p)^2 3 + \cdots]$

$= p\left[\frac{1}{p} + \frac{1(1-p)}{p^2}\right] = \frac{1}{p}$

$V(S)$ for every visit $= \dfrac{\left(\frac{1}{p} + 1\right)}{2}$ which is $\frac{1}{2}$ times that of true value.

$\Rightarrow$ Every visit MC is biased

3f. First visit MC: All the returns used in the calculation of value of a state are from diff episodes sampled randomly i.e i.i.ds. By the ~~law~~ of large numbers it converges to the true value as the num of episodes increases.

Every visit MC: Again assuming large num of episodes and exploring diferent starts garantees the converges of the algorithm. Though the returns are not all i.i.ds, the bias decreases consistently with increasing num of episodes (it asymptotically goes to zero)

Q4. Temporal Difference Methods

MDP - M

Policy - π

One step TD error: $\delta_t = r_{t+1} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)$

a) $E_{\pi}(\delta_t \mid s_t = s) = E_{\pi}(r_{t+1} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t) \mid s_t = s)$

From linearity of expectation $= E_{\pi}(r_{t+1} + \gamma V^{\pi}(s_{t+1}) \mid s_t = s) - \underbrace{E[V^{\pi}(s_t) \mid s_t = s]}_{V^{\pi}(s)}$

[$\because$ we are using true state value function V^{π}]

$= \underbrace{\sum_a \pi(a|s) \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V^{\pi}(s')]}_{= V^{\pi}(s) \text{ (From def of } V^{\pi}(s))} - V^{\pi}(s)$

$= 0$

b) $E_{\pi}(\delta_t \mid s_t = s, a_t = a) = E_{\pi}(r_{t+1} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t) \mid s_t = s, a_t = a)$

$= E_{\pi}(r_{t+1} + \gamma V^{\pi}(s_{t+1}) \mid s_t = s, a_t = s) - \underbrace{E[V^{\pi}(s_t) \mid s_t = s, a_t = a]}_{V^{\pi}(s)}$

$= \underbrace{\sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V^{\pi}(s')]}_{Q(s,a) \text{ [From the def of } Q(s,a)]} - V^{\pi}(s)$

$= Q^{\pi}(s,a) - V^{\pi}(s)$

c) TD(λ) algorithm, λ return target

$$G_t^{\lambda} = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

where $G_t^{(n)} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots \gamma^{n-1} r_{t+n} + \gamma^n V(s_{t+n})$

let $\eta(\lambda)$ denote time by which weighing seq reduces by half.

$$\Rightarrow \frac{1}{2} \le \frac{(1-\lambda)\lambda^{\eta(\lambda)-1}}{(1-\lambda)\lambda^{1-1}}$$

$$-\ln 2 \le [\eta(\lambda) - 1] \ln \lambda$$

$$\boxed{1 - \frac{\ln 2}{\ln \lambda} \le \eta(\lambda)}$$

Value of λ for which wts drop to half after 3 steps

ie $\eta(\lambda) = 3$

$$1 - \frac{\ln 2}{\ln \lambda} = 3$$

$$-2 = \frac{\ln 2}{\ln \lambda}$$

$$\ln \lambda = \ln 2^{-1/2}$$

$$\lambda = \frac{1}{\sqrt{2}}$$

Q5. Consider the p-series $\sum_{n=1}^{\infty} \frac{1}{n^p}$

We know that if $p \leq 1 \rightarrow$ divergent

$p > 1 \rightarrow$ convergent

Proof: Let us look at the convergence of the cores. integral

$$(p>0) \quad \int_1^{\infty} \frac{1}{x^p}\,dx = \lim_{m\to\infty} \int_1^m \frac{1}{x^p}$$

$p = 1$

$$\lim_{M\to\infty} [\ln x]_1^M = \ln\infty - 0 = \infty$$

Diverges

$p \neq 1$

$$\lim_{m\to\infty} \left[\frac{x^{-p+1}}{-p+1}\right]_1^m$$

$$\frac{1}{1-p} \lim_{m\to\infty} m^{1-p} - \frac{1}{1-p}$$

$1-p > 0$, $1 > p$: Diverges

$1-p < 0$, $1 < p$: Converges

(1) $\alpha_t = \frac{1}{t}$

$$\sum_{t=0}^{\infty} \alpha_t = \sum_{t=0}^{\infty} \frac{1}{t} \quad p = 1 \rightarrow \text{diverges} = \infty$$

$$\sum \alpha_t^2 = \sum_{t=0}^{\infty} \frac{1}{t^2} \quad p = 2 \rightarrow \text{converges} < \infty$$

$\Rightarrow$ It obeys Robbins-Monroe condition thus converges

(2) $\alpha_t = \frac{1}{t^2}$

$\sum \alpha_t = \sum \frac{1}{t^2}$ $p > 1$ Converges $< \infty$

$\sum \alpha_t^2 = \sum \frac{1}{t^4}$ $p = 4 > 1$ Converges $< \infty$

$\Rightarrow$ Doesn't obey Robbins - Monroe condition thus doesn't converge.

(3) $\alpha_t = \frac{1}{t^{2/3}}$

$\sum \alpha_t = \sum \frac{1}{t^{2/3}}$ $p = 2/3 < 1$ Diverges $= \infty$

$\sum \alpha_t^2 = \sum \frac{1}{t^{4/3}}$ $p = 4/3 > 1$ Converges $< \infty$

$\Rightarrow$ Obeys Robbins - Monroe condition thus converges.

(4) $\alpha_t = \frac{1}{t^{1/2}}$

$\sum \alpha_t = \sum \frac{1}{t^{1/2}}$ $p = 1/2 < 1$ Diverges $= \infty$

$\sum \alpha_t^2 = \sum \frac{1}{t}$ $p = 1$ Diverges $= \infty$

$\Rightarrow$ Doesn't obey Robbins - Monroe condition thus doesn't converge.

For learning rate $\alpha_t = \frac{1}{t^p}$

$\sum \alpha_t = \sum \frac{1}{t^p} = \infty \Rightarrow p \leq 1$

$\sum \alpha_t^2 = \sum \frac{1}{t^{2p}} < \infty \Rightarrow 2p > 1$

$p > 1/2$

$\Rightarrow p \in (1/2, 1]$ for α_t to converge to $V(S)$.