AI21MTECH12003

ASSIGNMENT - 2

① ⓐ from the description in question we get that V_{k+1} is the ~~laste~~ last iteration of the algorithm, and also

$$\|V_{k+1} - V_k\|_\infty \leq \varepsilon$$

let B be the Bellman evaluation backup

$$V_{k+1} = B V_k$$

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

(triangular inequality over norms)

$$= \|V_k - V_{k+1}\|_\infty + \|B V_k - B V^\pi\|_\infty$$

(Bellman Evaluation backup)

$$\|V_k - V^\pi\|_\infty \leq \varepsilon + \gamma \|V_k - V^\pi\|_\infty$$

$$\|V_k - V\|_\infty \leq \frac{\varepsilon}{1-\gamma}$$

$$\|V_{k+1} - V^\pi\|_\infty = \|B V_k - B V^\pi\|_\infty \leq \gamma \|V_k - V^\pi\|_\infty$$

$$\|V_{k+1} - V^\pi\|_\infty \leq \frac{\gamma\varepsilon}{1-\gamma}$$

① (b) from the previous question we get that

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

similarly

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

$$\vdots$$

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

using this in eq ①

$$\|V_{k+1} - V^\pi\|_\infty \le \gamma^2 \| V_{k-1} - V^\pi\|_\infty$$

$\cancel{\|V_{k+1} - V^\pi\|_\infty \le \gamma^k \| V_k}$ similarly replacing each equation on RHS

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

① (c) given that action set A is finite, we can say that there exists $a_1, a_2 \in A$ such that for a fixed state $s \in S$

$$L(u(s)) = \left[R(s, a_1) + \gamma \sum_{s'} P(s'/s, a_1) u(s) \right] \quad — ①$$

and also

$$L(v(s)) = \left[R(s, a_2) + \gamma \sum_{s'} P(s'/s, a_2) v(s) \right]$$

from the definition of optimality operator a_2 ~~st~~ $\in A$ gives the maximum value for L, we can say that

classmate

Date

Page

$$L(V(S)) \geq \left[R(S,a_1) + \gamma \sum_{s'} P(s'/s,a_1)\, V(s) \right] \quad — ②$$

subtract ② from ①,

$$L(u(S)) - L(V(S)) \leq \left[\gamma \sum_{s'} P(s'/s,a_1) \, (u(s) - V(s)) \right]$$

since $u(s) \leq V(S)$;

$$L(u(S)) - L(V(S)) \leq \left[\gamma \sum_{s'} P(s'/s,a_1) \, (V(s) - V(s)) \right]$$

$$= 0$$

$$L(u(s)) - L(V(S)) \leq 0$$

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

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

hence proved

②ⓐ for all $u, v \in V$, we can write

$$\| P \circ Q (v) - P \circ Q (u) \| = \| P(Q(v)) - P(Q(u)) \|$$

$$\leq r_p \| Q(v) - Q(u) \|$$

$$= r_p r_q \| v - u \|$$

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

r_p and r_q both belong to $[0, 1)$
hence there product also belongs to $[0, 1)$
which makes $P \circ Q$ compostion to be
contraction on the same normed
vector space.

$$\| Q \circ P (v) - Q \circ P (u) \| = \| Q(P(v)) - Q(P(u)) \|$$

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

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

similarly, $Q \circ P$ compostion is also contraction
on the same normed vector space.

②ⓑ from the ②ⓐ we can see that
the contraction coefficient for $P \circ Q$ and
$Q \circ P$ is $r_p r_q$.

②ⓒ For a unique solution to exist, the
operator B as $F \circ L$ must be contraction
which means that both F and L
must be contraction under the max-norm.

classmate

Date

Page

3 (a) Trajectories starting from s looks like below

$$\{\underbrace{S, S, S \ldots \ldots S}_{K \text{ occurrences of } S}, A\}$$

~~(b) first time estimate of $\hat{V(s)}$ for s state is 1~~ ~~as the reward~~

3 (b) Depending on the reward for for state s which is 1, from the state s, if the typical trajectory has K number of occurrences of S then the first visit estimate of $V(S)$ is K.

3 (c) Given one trajectory, we can construct K - 1 sub-trajectories (no. of times state s is visited) where t th trajectory returns K-t. (k is the no. of occurrences of states)

$$\hat{V}(s) \text{ (estimated V)} = \frac{1}{K} \sum_{t=0}^{K-1} (K-t)$$

$$= \frac{1}{K} \sum_{t'=1}^{K} t'$$

$$= \frac{1}{K} \left(\frac{K(K+1)}{2} \right)$$

$$\hat{V}(s) = \frac{K+1}{2}$$

3 (d)

$$V(S) = 1 + (1-\beta) V(S) + 0 \cdot \beta$$

$$V(S) = 1 + V(S) - \beta(V(S))$$

$$V(S) = \frac{1}{\beta}$$

$$\mathbb{E}[\hat{V}] = \mathbb{E}[K] = \frac{1}{p} = V^{\pi}(s)$$

K occurrences of state s

So the true value of $V(s)$ is $\frac{1}{p}$ as it's average length of the trajectory

3 e The value function $\hat{v}(s)$ estimation for every visit is given as $\frac{k+1}{2}$

The corresponding expectation can be written as

$$\mathbb{E}\left[\frac{k+1}{2}\right] = \frac{1}{2}\mathbb{E}[K] + \frac{1}{2}$$

$$= \frac{1}{2}\left[\frac{1}{p} + 1\right]$$

$$= \frac{1+p}{2p} \neq V^{\pi}(s)$$

This means that the every visit MC estimate is biased estimator.

3 f. First visit MC convergences just rely on the law of large numbers whereas every visit MC convergence may need to more care because the samples for mean calculation are all independent.

4 a $$\delta_t = r_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t)$$

$$\mathbb{E}[\delta_t \mid S_t = s] = \mathbb{E}[r_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t) \mid S_t = s]$$

$$= \mathbb{E}[r_{t+1} + \gamma V^\pi(S_{t+1}) \mid S_t = s] - V^\pi(s)$$

$$= V^\pi(s) - V^\pi(s)$$

$$\mathbb{E}[\delta_t \mid S_t = s] = 0$$

4 b $$\mathbb{E}[\delta_t \mid S_t = s, A_t = a]$$

$$= \mathbb{E}[r_{t+1} + \gamma V^\pi(S_{t+1}) - V^\pi(S_t) \mid S_t = s, A_t = a]$$

$$= \mathbb{E}[r_{t+1} + \gamma V^\pi(S_{t+1}) \mid S_t = s, A_t = a] - V^\pi(s)$$

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

$$\mathbb{E}[\delta_t \mid S_t = s, A_t = a] = Q^\pi(s,a) - V^\pi(s)$$

4 c $\eta(\lambda)$ denote the time by which the weighting sequence would have fallen half of its initial value.
The half life occurs when weighting drops by half. Therefore

$$\lambda^{\eta(\lambda)} = \frac{1}{2}$$

$$\eta(\lambda) = \frac{\ln(\frac{1}{2})}{\ln(\lambda)}$$

for 3 step return, the time taken is 2

$$\eta(\lambda) = 2 \Rightarrow 2 = \frac{\ln(\frac{1}{2})}{\ln(\lambda)}$$

$$\Rightarrow \lambda^2 = (\tfrac{1}{2})$$

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

(5) The harmonic series' generalization is the p-series defined as $\sum_{i=1}^{\infty} \frac{1}{i^p}$

$$\sum_{i=1}^{\infty} \frac{1}{i^p}$$ for any positive real number p

The p-series converges for all $p > 1$ and diverges for all $p \leq 1$.

(i) $\alpha_t = \frac{1}{t} \Rightarrow \sum_{t=1}^{\infty} \alpha_t = \infty$

$$\sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

Algo doesnot converge with given α_t.

(ii) $\alpha_t = \frac{1}{t^2} \Rightarrow \sum_{t=1}^{\infty} \alpha_t = \infty$

$$\sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

classmate
Date
Page

Algo converges with given value of α_t

(iii) $\alpha_t = \frac{1}{t^{2/3}} \Rightarrow \sum_{t=1}^{\infty} \alpha_t = \infty$

$\sum_{t=1}^{\infty} \alpha_t^2 < \infty$

Algo converges with given value of α_t

(iv) $\alpha_t = \frac{1}{t^{1/2}} \Rightarrow \sum_{t=1}^{\infty} \alpha_t = \infty$

$\sum_{t=1}^{\infty} \alpha_t^2 = \infty$

Algo doesnot converges with given value of α_t.

generalizing above result we can write

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

$\sum_{t=1}^{\infty} \frac{1}{t^{2p}} < \infty \qquad 2p > 1$

$p > \frac{1}{2}$

The Robbins-Monroe condition is true when $p \in (\frac{1}{2}, 1)$.