یادگیری تقویتی   امیرضا [illegible] 400102222   تمرین 2

سوال 1

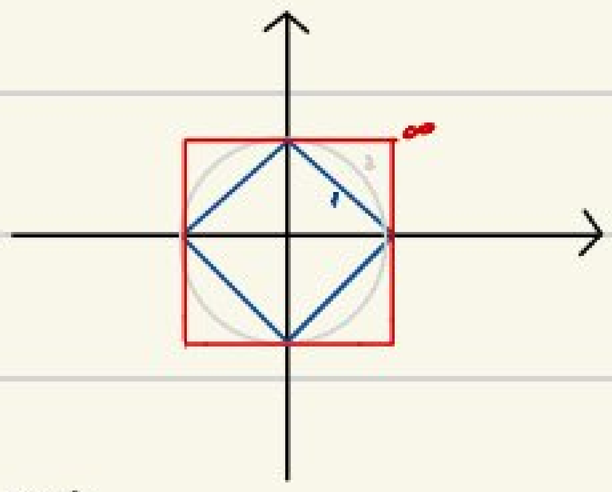
الف)

To show that $U(V)$ is a Contraction mapping; we have to prove that $\|U(V_1)-U(V_2)\| \leq \|V_1-V_2\|$; i.e., $U(.)$ makes two Value function closer to each other in any norm.

It's orbitary to show that maximum differnce between two vectors is their infinity norm distance; i.e.:

$$\|U(V_1)-U(V_2)\|_\infty \leq \|V_1-V_2\|_\infty \Longrightarrow \|U(V_1)-U(V_2)\| \leq \|V_1-V_2\|$$

A short proof can be appreared by visualation:



Thus The objection can be reduced to ∞-norm case.

$$\|U(V_1)-U(V_2)\|_\infty = \|R+\gamma P V_1 - R - \gamma P V_2\| = \gamma\|P(V_1-V_2)\|_\infty \leq \|P(V_1-V_2)\|_\infty \overset{\#}{\leq} \|P\|_\infty \|V_1-V_2\|_\infty \overset{*}{=} \|V_1-V_2\|$$

where in # I used the Cauchy-Schwartz theorem and in * I used the fact that P is a transition matrix; thus it's sum of rows is 1 and hence it's ∞-norm is equal to one.

ب)

$$\lim_{n\to\infty} U^n(V) = V^\pi \Leftrightarrow \lim_{n\to\infty} U^n(V) - V^\pi = 0 \Leftrightarrow \lim_{n\to\infty} U(U^{n-1}(V)) - V^\pi = 0 \Leftrightarrow \lim_{n\to\infty} U(U^{n-1}(V) - V^\pi) = 0 \Leftrightarrow \lim_{n\to\infty} \|U(U^{n-1}(V) - V^\pi)\|_\infty = 0$$

$$\overset{\#}{\Leftrightarrow} \lim_{n\to\infty} \|U(U^{n-1}(V) - V^\pi)\|_\infty \leq \gamma \lim_{n\to\infty} \|U^{n-1}(V) - V^\pi\|_\infty \leq \dots \leq \lim_{n\to\infty} \gamma^n \|V - V^\pi\|_\infty \Leftrightarrow \lim_{n\to\infty} \gamma^n = 0$$

we know that $\lim_{n\to\infty} \gamma^n$ converges to zero for any bounded γ which is strictly less than zero. Also in # we used the fact that $U(V^\pi) = V^\pi$ which is obvious; i.e. $U(V^\pi) = V^\pi$ which means mapping has converged.

پ)

$$\|U^k(V) - U^{k-1}(V)\|_\infty < \varepsilon \overset{?}{\Longrightarrow} \|V^\pi - U^k(V)\|_\infty < \frac{\varepsilon}{1-\gamma}$$

$$\|V^\pi - U^{k+1}(V) + U^{k+1}(V) - U^k(V)\|_\infty \leq \|V^\pi - U^{k+1}(V)\|_\infty + \|U^{k+1}(V) - U^k(V)\|_\infty \leq \|V^\pi - U^{k+1}(V)\|_\infty + \gamma\|U^{k+1}(V) - U^k(V)\|_\infty$$

$$\leq \|V^\pi - U^{k+1}(V)\|_\infty + \gamma\varepsilon \leq \|V^\pi - U^{k+2}(V)\|_\infty + \|U^{k+2}(V) - U^{k+1}(V)\| + \gamma\varepsilon$$

$$\leq \|V^\pi - U^{k+2}(V)\|_\infty + \|U(U(U^k(V))) - U(U(U^{k-1}(V)))\| + \gamma\varepsilon \leq$$

$$\leq \|V^\pi - U^{k+2}(V)\|_\infty + \gamma^2\varepsilon + \gamma\varepsilon$$

$$\vdots$$

$$\leq \lim_{n\to\infty} \underbrace{\|V^\pi - U^n(V)\|}_{=0} + \sum_{i=1}^{n} \gamma\varepsilon = \frac{\varepsilon}{1-\gamma}$$

سوال ②

$G_{14} = +10 \quad G_{13} = 10\gamma \quad G_{12} = 10\gamma^2 \quad G_{11} = 10\gamma^3 \quad G_{10} = 10\gamma^4 - 10 \quad G_9 = 10\gamma^5 - 10\gamma \quad G_8 = 10\gamma^6 - 10\gamma^2 - 10$

$G_7 = 10\gamma^7 - 10\gamma^3 - 10\gamma \quad G_6 = 10\gamma^8 - 10\gamma^4 - 10\gamma^2 \quad G_5 = 10\gamma^9 - 10\gamma^5 - 10\gamma^3 \quad G_4 = 10\gamma^{10} - 10\gamma^6 - 10\gamma^4 \quad G_3 = 10\gamma^{11} - 10\gamma^7 - 10\gamma^5$

$G_2 = 10\gamma^{12} - 10\gamma^8 - 10\gamma^6 \quad G_1 = 10\gamma^{13} - 10\gamma^9 - 10\gamma^7$

الف)

Every Visit Monte Carlo

$V(1) = G_1$; $V(2) = G_2$; $V(3) = G_3$; $V(7) = G_5$; $V(8) = G_4$; $V(12) = G_6$; $V(16) = G_7$; $V(17) = G_{13}$

$V(20) = G8$; $V(18) = G_{14}$; $V(21) = \frac{G_9 + G_{11}}{2}$; $V(22) = \frac{G_{10} + G_{12}}{2}$

ب)

First Visit Monte Carlo

$V(1) = G_1$; $V(2) = G_2$; $V(3) = G_3$; $V(7) = G_5$; $V(8) = G_4$; $V(12) = G_6$; $V(16) = G_7$; $V(17) = G_{13}$

$V(20) = G8$; $V(18) = G_{14}$; $V(21) = G_{11}$; $V(22) = G_{10}$

سوال ③

الف)

$V^{\pi}(s) = E^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid S_0 = s\right] = E^{\pi}\left[r_0 + \sum_{t=1}^{\infty} \gamma^t r_t \mid S_0 = s\right] = E^{\pi}\left[r_0 + \gamma \sum_{t'=0}^{\infty} \gamma^{t'} r_{t'} \mid S_0 = s\right] = E^{\pi}[r_0 \mid S_0] + \gamma E^{\pi}\left[\sum_{t'=0}^{\infty} \gamma^{t'} r_{t'} \mid S_0 = s\right]$ *

* $E^{\pi}\left[\sum_{t'=0}^{\infty} \gamma^{t'} r_{t'} \mid S_0 = s\right] = E_{s' \sim p(s'|s,a)}\left[E^{\pi}\left[\sum_{t'=0}^{\infty} \gamma^{t'} r_{t'} \mid S_0' = s' \mid S_0 = s\right]\right] \xrightarrow{MDP^*} = E_{s' \sim p(s'|s,a)}\left[E^{\pi}\left[\sum_{t'=0}^{\infty} \gamma^{t'} r_{t'} \mid S_0' = s'\right]\right] = E_{s' \sim p(s'|s,a)}[V^{\pi}(s')]$

$\Rightarrow V^{\pi}(s) = E^{\pi}\left[E_{s' \sim p(s'|s,a)}[r_0 + \gamma V^{\pi}(s')]\right]$ ⟶ independent of S_0 * $p(S_2|S_1, S_0) \xrightarrow{MDP} = p(S_2|S_1)$

ب)

define $\pi_1^*(s) = \arg\max_a E_{s' \sim p(s'|s,a)}[r_0 + \gamma V^{\pi}(s')] \xrightarrow{\times \gamma} \pi_2^*(s) = \arg\max_a E_{s' \sim p(s'|s,a)}[\gamma r_0 + \gamma \gamma V^{\pi}(s')] =$

$= \gamma E_{s' \sim p(s'|s,a)}[r_0 + \gamma V^{\pi}(s')]$

π_1^* and π_2^* are the same because their objective function has been multiplied by a constant factor and hence can be factorized and ignored in an optimization seeking the optimal argument.

ج)

$V^{\pi}(s) = E^{\pi}\left[\sum_{t=0}^{T} \gamma^t r_t \mid S_0 = s\right] \Rightarrow V_2^{\pi}(s) = E^{\pi}\left[\sum_{t=0}^{T} \gamma^t (r_t + C) \mid S_0 = s\right] = E^{\pi}[\gamma^t r_t \mid S_0] + C E^{\pi}\left[\sum_{t=0}^{T} \gamma^t \mid S_0 = s\right] = V_1^{\pi}(s) + C E^{\pi}\left[\frac{1-\gamma^{T+1}}{1-\gamma} \mid S_0 = s\right]$

It's obvious that $E^{\pi}\left[\frac{1-\gamma^{T+1}}{1-\gamma} \mid S_0 = s\right]$ is not a constant term, it obviously depends not only on π; but also on where we start which is not considered case in Bellman equation. Thus $V_2^{\pi}(s) \neq V_1^{\pi}(s)$ i.e. we are dealing with a new value function in second definition and there no constrains that optimal policy for $V_2^{\pi}(s)$ will converge $\pi_1^*(s)$ as it's dealing with different objective function.

د)

$V_2^{\pi}(s) - V_1^{\pi}(s) = C E^{\pi}\left[\lim_{T \to \infty} \frac{1-\gamma^{T+1}}{1-\gamma} \mid S_0 = s\right] = C E^{\pi}\left[\frac{1}{1-\gamma} \mid S_0 = s\right] = \frac{C}{1-\gamma}$ ⟹ like part b; $\pi_1^*(s)$ and $\pi_2^*(s)$ will be the same policies. Proof

independent of π and s

also can be obtained by defining $\pi_d^*(s) = \pi_2^*(s) - \pi_1^*(s) = \arg\max_a E_{s'\sim p(s'|s,a)}\left[r - r + \gamma V_1^{\pi}(s) + \frac{c}{1-\gamma} - \gamma V_1^{\pi}(s)\right] = \dots \Rightarrow \pi_d^*(s) = 0$

ه)

For first MDP; we have $\pi_1^*(s) = \arg\max_a E_{s'}[r + \gamma V_1^{\pi}(s)]$. This implies that for any state s: $\pi_1^*(s)$ chooses the optimal action a^*. To proove that $\pi_2^*(s) = \pi_1^*(s)$ we can show that for any states; $E_{s'}[r + \gamma V_2^{\pi}(s) \mid a = a^*] \geq E_{s'}[r + \gamma V_2^{\pi}(s) \mid a \neq a^*]$. Thus we have:

$E_{s'}[r + \gamma V_2^{\pi}(s) \mid a \neq a^*] = E_{s'}[r - c + \gamma V_2^{\pi}(s)] \overset{*}{=} E_{s'}[r + \gamma V_1^{\pi}(s)] - c \longrightarrow \pi_2^*(s) = \pi_1(s)$

Also at * we assumed that value functions are the same if we choose the optimal action; which is shown above.

سوال ④

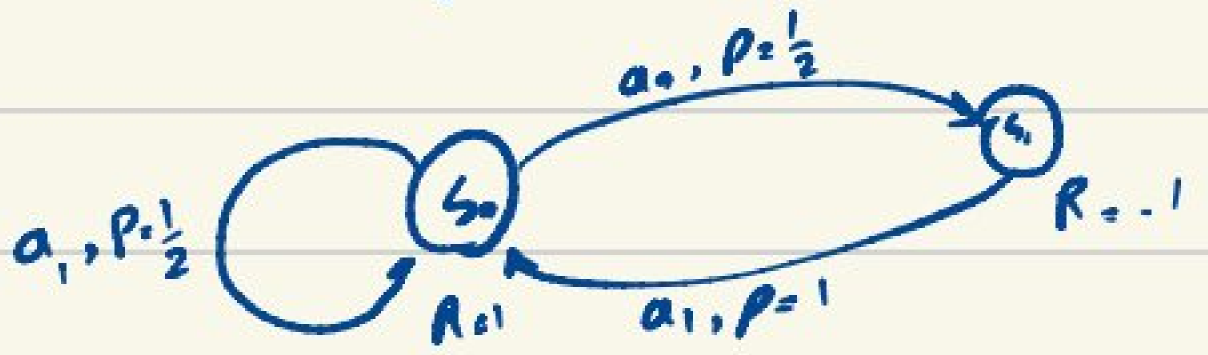
الف)

$$V^{\pi}(s') \overset{?}{=} \sum_s \sum_a P(s,a,s')\,\pi(s,a)\left[\frac{V^{\pi}(s) - R(s,a)}{\gamma}\right]$$

Bellman equation: $V^{\pi}(s) = \sum_{s'} \sum_a P(s'|s,a)\,\pi(a|s)\,(R(s,a) + \gamma V^{\pi}(s'))$

$P(s',a,s)$ describes path probability and $\pi(a,s)$ explains joint probability of states and actions. Consider the case where s' and s both have action a_0. then $P(s,s',a_0)$ doesn't have specific meaning because it can not obtain 2 diffrent values.

Counter example:



$\gamma = \frac{1}{2}$

Bellman equation: $\begin{cases} V^{\pi}(0) = \frac{1}{2}(1 + \frac{1}{2}V^{\pi}(0)) + \frac{1}{2}(-1 + \frac{1}{2}V^{\pi}(1)) = \frac{3}{4}V^{\pi}(0) - \frac{1}{4}V^{\pi}(1) \rightsquigarrow V^{\pi}(0) = \frac{1}{3}V^{\pi}(1) \\ V^{\pi}(1) = 1 + \frac{1}{2}V^{\pi}(0) = 1 + \frac{1}{6}V^{\pi}(1) \Rightarrow V^{\pi}(1) = \frac{6}{5},\ V^{\pi}(0) = \frac{2}{5} \end{cases}$

suggested equation: $\begin{cases} V^{\pi}(0) = \frac{2}{2}(V^{\pi}(1) + 1) + \frac{2}{2}(V^{\pi}(0) - 1) \Rightarrow V^{\pi}(1) = 0 \\ V^{\pi}(1) = 2(V^{\pi}(0) - 1) \Rightarrow V^{\pi}(0) = 1 \end{cases}$ Contradiction

ب)

$V^{(\pi)}(s) = E[G_t \mid S_t = s, \pi] \overset{\text{As shown in part 3.a}}{=} E^{\pi}\left[r_t + E_{s'\sim p(s'|S_t,\pi(s))}[V^{\pi}(s')] \mid S_t = s\right]$

$= E^{\pi}[r_t \mid S_t = s] + E^{\pi}\left[E_{s'\sim p(s'|S_t,\pi(s))}[V^{\pi}(s')] \mid S_t = s\right]$

from markov property $= E^{\pi}[R(s,\pi(s))] + E^{\pi}\left[E_{s'\sim p(s'|s,\pi(s))}[V^{\pi}(s')]\right]$

$= E^{\pi}[r_0 \mid S_0 = s] + E^{\pi}\left[E_{s'\sim p(s'|S_0,\pi(S_0))}[V^{\pi}(s')] \mid S_0 = s\right]$

$= E^{\pi}\left[r_0 + E_{s'\sim p(s'|S_0,\pi(S_0))}[V^{\pi}(s')] \mid S_0 = s\right] = E[G_0 \mid S_0 = s, \pi] = V^{\pi}(s)$

ج)

$$V^{\pi}(s_t) = E^{\pi}[r_t + E_{s_{t+1} \sim p(s_{t+1}|s,\pi(s))}[V^{\pi}(s_{t+1})]]$$

$= E^{\pi}[r_t + V^{\pi}(s_{t+1})]$ transition is deterministic

$= E^{\pi}[r_t] + V^{\pi}(s_{t+1})$ actions don't effect transition

$< V^{\pi}(s_{t+1})$ rewards are defintly negative.

سوال ۵

الف)

n-step TD learning: $V^{k+1}(s_t) \leftarrow V_t(s_t) + \alpha(\sum_{k=1}^{n} \gamma^{k-1} R_k + \gamma^n V^k(s_{t+n}) - V_t(s_t))$

$n=1 \mapsto V(E)$ gets updated

$n=2 \mapsto V(D), V(E)$ gets updated

$n=3 \mapsto$ Value function for all steps gets updated

ب)

As we increase α; at first TD learns more and converges better for limited iteration. Afterwards; by increasing α; temporal error makes bigger impact in algharithm and hence for big values of α; Value functions get unstable and don't converge to optima.

For large n; we have more presice temporal difference and better converges. If we increase n however makes algharithm needy for more iterations and thus converges slowly.

So proper n, α for TD algharithm is nothing huge and not neglectable amount.

ج)

A) Increasing states causes more leafs for state-action tree; i.e. we have bigger environment to explore in contrast to before. Thus more probably the error rate for same episode and n will increase.

B) From law of large numbers we know that more episodes mean more samples and thus less error probably. In fact; for good amount of n; we converge to $V^*(s)$ if we take ∞ samples.

C) For a proper amount of α (recall part ب); more repetition means Convergence! i.e. if α is small; running algorithm for $V(s)$ makes maximum usage of the explored data. However; if α is large; more repeat has no good cause it may sway.

د)

$$E_t(s) = \gamma\lambda E_{t-1}(s) + 1\{S_t = S_{t-1}\}$$

$E_t(s)$ reaches it's maximum increase "$E_{t-1}(s)$" when the state s is selected for accumulating traces.

thus: $E_t(s) = \frac{1}{4} \times \frac{4}{5} E_t(s) + 1 = \frac{1}{5} E_t(s) + 1 \mapsto E_t(s) = \frac{5}{4}$