HW 1

Problem 1

(a)

(i)

$$\max_{\pi} V^{\pi}(s) = \max_{\pi} \sum_{a} \pi(a|s) \cdot Q^{\pi}(s,a) \le \max_{\pi} Q^{\pi}(s, \pi(s)) = Q^{*}(s,a)$$

$V^{*}(s)$

$\because 0 \le \pi(a|s) \le 1$

Suppose $V^{*}(s) < Q^{*}(s,a)$,

$$\rightarrow \max_{\pi} \sum_{a} \pi(a|s) \cdot Q^{\pi}(s,a) < \max_{\pi} Q^{\pi}(s,a)$$

take $a^{*} = \underset{a \in A}{\operatorname{argmax}}\, Q^{\pi}(s,a) \wedge \pi(a^{*}|s) = 1$

$$\rightarrow \max_{\pi} Q^{\pi}(s,a^{*}) < \max_{\pi} Q^{\pi}(s,a^{*}) \rightarrow 0 < 0 \quad (*)$$

$\therefore V^{*}(s) = Q^{*}(s,a)$ #

(ii)

$$\max_{\pi} Q(s,a) = \max_{\pi} \{ R_{s,a} + \gamma \cdot \sum_{s'} P^{a}_{ss'} \cdot \max_{\pi} V^{*}(s') \}$$

$$\rightarrow \max_{\pi} Q(s,a) = R_{s,a} + \gamma \cdot \sum_{s'} P^{a}_{ss'} \cdot \max_{\pi} V^{*}(s')$$

$$\rightarrow Q^{*}(s,a) = R_{s,a} + \gamma \cdot \sum_{s'} P^{a}_{ss'} \cdot V^{*}(s')$$

(b)

$$\| T^*(Q) - T^*(Q') \|_\infty$$

$$= \max_{(s,a)} \left| \left(R_{s,a} + \gamma \cdot \sum_{s'} P^a_{ss'} \cdot \max_{a'} Q(s', a') \right) - \left(R_{s,a} + \gamma \cdot \sum_{s'} P^a_{ss'} \cdot \max_{a'} Q'(s', a') \right) \right|$$

$$= \max_{(s,a)} \gamma \cdot \left| \sum_{s'} P^a_{ss'} \cdot \left(\max_{a'} Q(s', a') - \max_{a'} Q'(s', a') \right) \right|$$

$$\leq \max_{(s,a)} \gamma \cdot \left| \sum_{s'} P^a_{ss'} \cdot \max_{a'} \left(Q(s', a') - Q'(s', a') \right) \right| \quad \left(|\max A - \max B| \leq \max(A - B) \right)$$

$$\leq \gamma \cdot \max_{(s,a)} \max_{a'} \left| Q(s', a') - Q'(s', a') \right| \quad (\because 0 < \text{entry of } P^a_{ss'}(s,a) < 1)$$

$$\leq \gamma \cdot \| Q - Q' \|_\infty$$

$\therefore$ T^* is a γ-contraction operator #

Problem 2

(a)

$$\| T^\pi_\Omega V(s) - T^\pi_\Omega V'(s) \|_\infty$$

$$= \max_s \left| \left(R^\pi_s + \Omega(\pi(\cdot|s)) + \gamma \cdot P^\pi_{ss'} \cdot V \right) - \left(R^\pi_s + \Omega(\pi(\cdot|s)) + \gamma \cdot P^\pi_{ss'} \cdot V' \right) \right|$$

$$= \max_s \left| \gamma \cdot P^\pi_{ss'} \cdot (V - V') \right|$$

$$\leq \gamma \cdot \max_s |V - V'| = \gamma \cdot \| V - V' \|_\infty$$

$\therefore$ T^π_Ω is γ-contraction operator in ℓ_∞ norm #

(b)

define $[T_{\Omega}^{\pi} V](s) = R_s^{\pi} + \Omega(\pi(\cdot|s)) + \gamma \cdot P_{ss'}^{\pi} \cdot V$

<algorithm>

step 1. initialize $k = 0$ and $V_0(s) = 0 \ \forall s \in S$

step 2. $V_{k+1} \leftarrow T_{\Omega}^{\pi}(V_k)$

$\Leftrightarrow V_{k+1}(s) = \max_{a \in A} \{ R_s^a + \underline{\Omega(\pi(\cdot|s))} + \gamma \cdot P_{ss'}^{\pi} \cdot V_k \}$

also consider the randomness of π

step 3.

$Q_{k+1}(s,a) = R_{s,a} + \gamma \cdot \sum_{s' \in S} P_{ss'}^{a} \cdot V_{k+1}(s')$

<explain>

in step 2, $\because T_{\Omega}^{\pi}$ is γ-contraction operator, it will converge to final V

$\therefore$ we compute T_{Ω}^{π} iteratively

in step 3, we get Q_{k+1} by equation (7),

if $V_{k+1}(s) = V_{\Omega}^{*}(s) \Rightarrow Q_{k+1}(s,a) = Q_{\Omega}^{*}(s,a)$

#

Problem 3

$$RHS = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s\sim d_\mu^{\pi_\theta}} \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[f(s,a)]$$

$$= \frac{1}{1-\gamma} \sum_{s_0\in S} \sum_{s\in S} \sum_{a\in A} f(s,a) \cdot \mu(s_0) \cdot \pi_\theta(a|s) \cdot (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \cdot P(s_t = s \mid s_0, \pi)$$

$$= \sum_\tau \sum_{t=0}^{\infty} \gamma^t \cdot f(s,a) \cdot P_\mu^{\pi_\theta}(\tau) \quad \text{(by lemma 1)}$$

$$= \mathbb{E}_{\tau\sim P_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t \cdot f(s,a)\right] \quad \#$$

lemma 1. let trajectory $\tau = (s_0, a_0, s_1, a_1, \cdots, s_t, a_t)$

$$P_\mu^\pi(\tau) = \mu(s_0) \cdot \pi_\theta(a_0|s_0) \cdots \pi(a_t|s_t) \cdot P(s_t|s_{t-1}, a_{t-1}, \pi) \cdot \pi(a_t|s_t)$$

by Markov property →
$$= \mu(s_0) \cdot \pi(a_t|s_t) \cdot \prod_{i=1}^{t} \pi(a_{i-1} | \bigcap_{j=0}^{i-1} s_j \cdot \bigcap_{j=0}^{i-1} a_j) \cdot P(s_i | \bigcap_{j=0}^{i-1} s_j, \bigcap_{j=0}^{i-1} a_j, \pi)$$

by def of conditional probability →
$$= \mu(s_0) \pi(a_t|s_t) \frac{\pi(a_0 \mid s_0)}{\pi(s_0)} \cdot \frac{P(a_0, \bigcap_{i=0}^{1} s_i, \pi)}{P(a_0, s_0, \pi)} \cdot \cdots \cdot \frac{\pi(\bigcap_{i=0}^{t-1} s_i, \bigcap_{i=0}^{t-1} a_i)}{\pi(\bigcap_{i=0}^{t-1} s_i, \bigcap_{i=0}^{t-1} a_i)} \cdot \frac{P(\bigcap_{i=0}^{t} s_i, \bigcap_{i=0}^{t-1} a_i, \pi)}{P(\bigcap_{i=0}^{t-1} s_i, \bigcap_{i=0}^{t-1} a_i, \pi)}$$

↘
$$= \mu(s_0) \cdot \pi(a_t|s_t) \cdot P(\bigcap_{i=0}^{t} s_i, \bigcap_{i=0}^{t-1} a_t \mid s_0, \pi)$$

lemma 2.

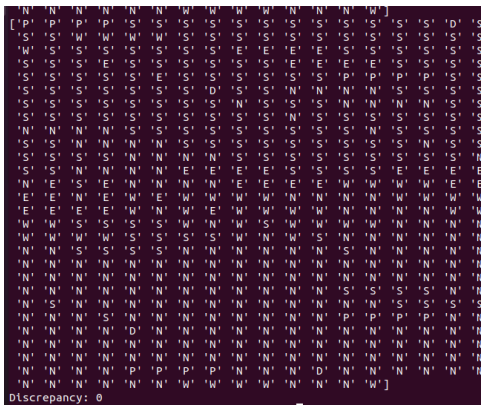
$$\sum_\tau P_\mu^\pi(\tau) = \sum_\tau \mu(s_0) \cdot \pi(a_t|s_t) \cdot P(\bigcap_{i=0}^{t} s_i, \bigcap_{i=0}^{t-1} a_i \mid s_0, \pi) \quad \text{(by lemma 1)}$$

extend τ trajectory →
$$= \sum_{s_0} \sum_{s_1} \cdots \sum_{s_t} \sum_{a_0} \cdots \sum_{a_t} \mu(s_0) \cdot \pi(a_t|s_t) \cdot P(\bigcap_{i=0}^{t} s_i, \bigcap_{i=0}^{t-1} a_i \mid s_0, \pi)$$

$$= \sum_{s_0} \sum_{s_t} \sum_{a_t} \mu(s_0) \cdot \pi(a_t|s_t) \cdot \left[\sum_{s_1} \sum_{s_2} \cdots \sum_{s_{t-1}} \sum_{a_0} \cdots \sum_{a_{t-1}} P(s_t, \bigcap_{i=0}^{t-1} s_i, \bigcap_{i=0}^{t-1} a_i \mid s_0, \pi)\right]$$

by total probability theorem →
$$= \sum_{s_0} \sum_{s_t} \sum_{a_t} \mu(s_0) \cdot \pi(a_t|s_t) \cdot P(s_t|s_0, \pi)$$

Problem 4



← this is the result of the PI algorithm

More detail please refer to the code

Problem 5

1. Original dataset : 'maze1d-umaze-v1'

   Format :

```
[[ 1.0856489   1.9745734   0.00981035  0.02174424]
 [ 1.0843927   1.97413    -0.12562364 -0.04433781]
 [ 1.0807577   1.9752754  -0.3634883   0.11453988]
 ...
 [ 1.1328583   2.8062387  -4.484303    0.09555068]
 [ 1.0883482   2.8068895  -4.4510083   0.06509537]
 [ 1.0463258   2.8074222  -4.202244    0.05324839]]
load datafile: 100%|
```

   Describe :

   it is a N x dim-observation array . where each observation is a float from -5 to 2

2. Mujoco dataset : 'halfcheetah-random-v2'

   Format :

```
[[-2.8627589e-02 -6.3696302e-02  8.6608730e-02 ... -8.6148446e-03
  -4.0926412e-03 -1.5979043e-01]
 [-4.3868057e-02 -2.2352228e-02  1.1845701e-02 ...  1.7004584e+00
  -5.4827838e+00  2.5702784e+00]
 [-8.1463926e-02 -3.5814878e-02 -8.6687684e-02 ... -7.7921929e+00
   1.4037144e+01 -8.9725056e+00]
 ...
 [-5.3456318e-01  3.2453218e+00 -9.1072828e-02 ...  1.3767828e+01
  -9.2018442e+00  1.5579005e+00]
 [-4.9477938e-01  3.1914997e+00 -3.9144486e-01 ... -2.0965147e+00
   2.5790567e+00  1.0007437e+01]
 [-5.1447099e-01  3.2184551e+00 -2.8413078e-01 ... -4.2252550e+00
   9.4621916e+00 -3.1550488e+00]]
load datafile: 100%|
```

   Describe :

   it is a N x dim-observation array . where each observation is a float ranging from -1 to 10