HW1

Problem 1.

a. For the first optimal equation, we can prove it as follow:

   By Total Probability Theorem, we know that

$$V^\pi(s) = \sum_{a \in A} \pi(a|s) Q^\pi(s,a)$$

   Let state $s \in \mathcal{S}$, we define

$$a^* = \arg\max_{a \in \mathcal{A}} Q^*(s,a)$$

   Then for any policy π, we have:

$$\begin{aligned} V^\pi(s) &= \sum_{a \in A} \pi(a|s) Q^\pi(s,a) \leq \sum_{a \in A} \pi(a|s) Q^*(s,a) \leq \sum_{a \in A} \pi(a|s) Q^*(s,a^*) \\ &= Q^*(s,a^*) \sum_{a \in A} \pi(a|s) = Q^*(s,a^*) \cdot 1 = Q^*(s,a^*) \\ &= \max_a Q^*(s,a) \end{aligned}$$

   The above equation is hold for any policy, so it must be true that

$$V^*(s) = \max_\pi V^\pi(s) = V^{\pi^*}(s) \leq \max_a Q^*(s,a)$$

   Now consider $V^*(s) < \max_a Q^*(s,a)$, by the assumption we can get there exist a policy ϕ such that $V^\phi(s) = \max_a Q^*(s,a) > V^*(s)$, but it's a contradiction with the definition of $V^*(s) = \max_\pi V^\pi(s)$. So $V^*(s) = \max_a Q^*(s,a)$ for all state $s \in \mathcal{S}$.

   For the second optimal equation, which can be proved as follow:

$$\begin{aligned} Q^*(s,a) &= \max_\pi Q^\pi(s,a) \\ &= \max_\pi \left(R_{s,a} + \gamma \sum_{s' \in S} P^a_{ss'} V^\pi(s')\right) \\ &= R_{s,a} + \gamma \sum_{s' \in S} P^a_{ss'} \max_\pi V^\pi(s') \\ &= R_{s,a} + \gamma \sum_{s' \in S} P^a_{ss'} V^*(s') \end{aligned}$$

b. ***Proof.***

$$\begin{aligned} ||T^*(Q) - T^*(Q')||_\infty &= \max_{s,a} |[T^*(Q)](s,a) - [T^*(Q')](s,a)| \\ &= \max_{s,a} |(R_{s,a} + \gamma \sum_{s'} P^a_{ss'} \max_{a'} Q(s',a')) - (R_{s,a} + \gamma \sum_{s'} P^a_{ss'} \max_{a'} Q'(s',a')| \\ &= \max_{s,a} |\gamma \sum_{s'} P^a_{ss'} (\max_{a'} Q(s',a') - \max_{a'} Q'(s',a'))| \\ &\leq \max_{s,a} \max_a |\gamma \sum_{s'} P^a_{ss'} (Q(s',a') - Q'(s',a'))| \\ &= \max_{s,a} \max_a |\gamma (Q(s',a') - Q'(s',a'))| \\ &\leq \gamma ||Q - Q'||_\infty \end{aligned}$$

   Therefore, T^* is γ-contraction operator ($\gamma < 1$).

Problem 2.

a. ***Proof.*** For any two value function V and V'

$$\begin{aligned}||T_\Omega^\pi(V) - T_\Omega^\pi(V')||_\infty &= ||(R^\pi + \Omega + \gamma P^\pi V) - (R^\pi + \Omega + \gamma P^\pi V')||_\infty \\ &= \gamma||P^\pi(V - V')||_\infty \\ &\leq \gamma||V - V'||_\infty\end{aligned}$$

Therefore, T_Ω^π is γ-contraction operator ($\gamma < 1$).

b. First we define Bellman optimality operator for regularized MDPs:

$$[T_\Omega^* V](s) := \max_\pi [T_\Omega^\pi V](s) = \max_\pi \left(R_s^\pi + \Omega(\pi(\cdot|s)) + \gamma P_{ss'}^\pi V\right)$$

The pseudo code to solve $V_\Omega^*(s)$ is shown below:

i. Initialize $k = 0$ and set $V_0(s) = 0$ for all states.

ii. Repeat the following until convergence: $V_{k+1} \leftarrow T_\Omega^*(V_k)$.

Equivalently: for each state s:

$$\begin{aligned}V_{k+1}(s) &= \max_\pi \Big[R_s^\pi + \Omega(\pi(\cdot|s)) + \gamma \sum_{s' \in S} P_{ss'}^\pi V_k(s')\Big] \\ &= \max_\pi \Big[R_s^\pi - \sum_{a \in \mathcal{A}} \pi(a|s) \ln \pi(a|s) + \gamma \sum_{s' \in S} P_{ss'}^\pi V_k(s')\Big]\end{aligned}$$

Because T_Ω^* is a contraction operator and V_Ω^* is a fixed point of T_Ω^*, so when k is big enough, $V_k \to V_\Omega^*$ due to uniqueness. Then we can derive Q_Ω^* from the equation:

$$\text{For each action } a \text{ and state } s: \quad Q_\Omega^*(s,a) = R_s^a + \gamma E_{s' \sim P(\cdot|s,a)}[V_\Omega^*(s')]$$

Problem 3.

$$\begin{aligned}\frac{1}{1-\gamma}\mathbb{E}_{s \sim d_\mu^{\pi_0}}\,\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[f(s,a)] &= \frac{1}{1-\gamma}\sum_{s \in \mathcal{S}} d_\mu^{\pi_0}(s)\left(\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[f(s,a)]\right) \\ &= \frac{1}{1-\gamma}\sum_{s \in \mathcal{S}} d_\mu^{\pi_0}(s) \sum_{a \in \mathcal{A}} \pi_\theta(a|s) \cdot f(s,a) \\ &= \mathbb{E}_{s_0 \sim \mu}\Big[\sum_{s \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t P(s_t = s|s_0, \pi_0) \sum_{a \in \mathcal{A}} \pi_0(a|s) \cdot f(s,a)\Big] \\ &= \sum_\tau \sum_{t=0}^{\infty} \gamma^t P_\mu^{\pi_0}(\tau) \cdot f(s,a) \\ &= \mathbb{E}_{\tau \sim P_\mu^{\pi_0}}\Big[\sum_{t=0}^{\infty} \gamma^t f(s,a)\Big]\end{aligned}$$

Problem 5.

After setting up all the dependencies that D4RL package needs, I successfully ran `d4rl_sanity_check.py` and gain the following results:

```
load datafile: 100%|██████████| 8/8 [00:00<00:00, 15.80it/s]
[[ 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%|██████████| 8/8 [00:00<00:00, 18.19it/s]
```

From the documentation of Gym (link), we know that observation dataset is an `ndarray` of shape `(4,N)`. The elements of the array correspond to the following:

Num	Observation	Min	Max	Joint Name (in corresponding XML file)	Joint Type	Unit
0	x coordinate of the green ball in the MuJoCo simulation	-Inf	Inf	ball_x	slide	position (m)
1	y coordinate of the green ball in the MuJoCo simulation	-Inf	Inf	ball_y	slide	position (m)
2	Green ball linear velocity in the x direction	-Inf	Inf	ball_x	slide	velocity (m/s)
3	Green ball linear velocity in the y direction	-Inf	Inf	ball_y	slide	velocity (m/s)

Then I modified the `env = gym.make('maze2d-umaze-v1')` to `env = gym.make('hopper-medium-v0')`, a task of MuJoCo.

```
load datafile: 100%|██████████| 5/5 [00:00<00:00,  9.04it/s]
[[ 1.2444514e+00  2.7131591e-02 -8.8438280e-02 ... -2.9314532e+00
  -1.3299903e+00  6.1160928e-01]
 [ 1.2403151e+00  1.5472738e-02 -1.1442658e-01 ... -3.5621471e+00
  -1.9423494e+00  2.8149670e-01]
 [ 1.2367572e+00  1.7649246e-03 -1.4012979e-01 ... -2.8636725e+00
  -2.6759245e+00  8.3468568e-01]
 ...
 [ 7.8751677e-01  3.0673077e-02 -1.1401708e+00 ... -4.2804508e+00
   6.4485419e-01 -7.2679301e-03]
 [ 7.5421393e-01  1.7388938e-02 -1.1694210e+00 ... -3.0343819e+00
  -5.4834080e-01 -3.8771250e-03]
 [ 7.2064501e-01  1.3139286e-02 -1.1885022e+00 ... -1.7375422e+00
   4.9596998e-01 -5.9253159e-03]]
load datafile: 100%|██████████| 5/5 [00:00<00:00,  8.56it/s]
```

From the documentation of Gym (link), we know that observation dataset is an `ndarray` of shape `(11,N)`. The elements of the array correspond to the following:

Num	Observation	Min	Max	Name (in corresponding XML file)	Joint	Unit
0	z-coordinate of the top (height of hopper)	-Inf	Inf	rootz	slide	position (m)
1	angle of the top	-Inf	Inf	rooty	hinge	angle (rad)
2	angle of the thigh joint	-Inf	Inf	thigh_joint	hinge	angle (rad)
3	angle of the leg joint	-Inf	Inf	leg_joint	hinge	angle (rad)
4	angle of the foot joint	-Inf	Inf	foot_joint	hinge	angle (rad)
5	velocity of the x-coordinate of the top	-Inf	Inf	rootx	slide	velocity (m/s)
6	velocity of the z-coordinate (height) of the top	-Inf	Inf	rootz	slide	velocity (m/s)
7	angular velocity of the angle of the top	-Inf	Inf	rooty	hinge	angular velocity (rad/s)
8	angular velocity of the thigh hinge	-Inf	Inf	thigh_joint	hinge	angular velocity (rad/s)
9	angular velocity of the leg hinge	-Inf	Inf	leg_joint	hinge	angular velocity (rad/s)
10	angular velocity of the foot hinge	-Inf	Inf	foot_joint	hinge	angular velocity (rad/s)