1. (a) We first prove $V^*(s) \le \max_a Q^*(s,a) \Rightarrow$

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

$\max_\pi \left(\sum_{a\in A}\pi(a|s)\cdot\left(\max_{a\in A} Q^\pi(s,a')\right)\right) \le \max_\pi \sum_{a\in A}\pi(a|s)\cdot$

$\max_{a\in A} Q^*(s,a) = \max_a Q^*(s,a)$, so $V^*(s) \le \max_a Q^*(s,a)$

Then we show $V^*(s) < \max_a Q^*(s,a)$ cannot happen,

Suppose $V^*(s) < \max_a Q^*(s,a)$, so there exist π' s.t.

$V^{\pi'} = \sum_{a\in A}\pi'(a|s)\cdot Q^*(s,a) > V^*(s)$, but $V^*(s)$ is the

best $V(s) \Rightarrow$ contradiction $\Rightarrow V^*(s) = \max_a Q^*(s,a)$

We know that $Q^*(s,a) = \max_\pi Q^\pi(s,a)$, and $Q^\pi(s,a)$

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

$= R_{s,a} + \gamma\sum_{s'} P^a_{ss'} \times \left(\max_\pi V^\pi(s')\right) = R_{s,a} + \gamma\sum_{s'} P^a_{ss'} V^*(s')$

(b) For any two action-value functions Q, Q', we have

$\|T^*(Q) - T^*(Q')\|_\infty = \max_{(s,a)} \left|[T^*(Q)](s,a) - [T^*(Q')](s,a)\right|$

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

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

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

$\leq \gamma \cdot \|(Q - Q')\|_\infty$ (due to all entrys in $P^a_{ss'} \leq 1$)

so T^* is a γ-contraction operator in terms of ∞-norm ($\gamma < 1$)

---

2. (a) For any two value functions V, V', we have

$$\|T^\pi_\Omega(V) - T^\pi_\Omega(V')\|_\infty = \max_{s \in S} \left| [T^\pi_\Omega V](s) - [T^\pi_\Omega V'](s) \right|$$

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

$$= \max_{s \in S} \left| \gamma \cdot P^\pi_{ss'} \cdot (V(s') - V'(s')) \right| \leq \max_{s \in S} \left| \gamma \cdot (V(s) - V'(s)) \right|$$

(due to all entrys in $P^\pi_{ss'} \leq 1$)

$= \gamma \cdot \|(V - V')\|_\infty$, so T^π_Ω is a contraction operator in L_∞ norm

---

(b) $k \leftarrow 0$, for all states $s \in S$, $V^0_\Omega(s) \leftarrow 0$ (initialize)

while V^k_Ω doesn't converge do (直至收敛才停止)

$V^{k+1}_\Omega(s) \leftarrow \max_{\pi \in \Pi} (R^\pi_s + \gamma P^\pi_s V^k_\Omega)$ for all states $s \in S$ (update value function)

$k \leftarrow k + 1$ (k increment)

end while (迴圈停止)

$Q^*_\Omega(s, a) \leftarrow R_{s,a} + \gamma E_{s' \sim P(\cdot|s,a)}[V^k_\Omega(s')]$ (此時 $V^k_\Omega \equiv V^*_\Omega$, get optimality Q-value functio

for all actions $a \in A$, for all states $s \in S$

Return $V^*_\Omega(s)$, $Q^*_\Omega(s, a)$

3. $\frac{1}{1-\gamma} E_{s\sim d_\mu^{\pi_\theta}} E_{a\sim\pi_\theta(\cdot|s)}[f(s,a)] = \frac{1}{1-\gamma} \sum_{s\in S} \sum_{a\in A} d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot f(s,a)$

$= \frac{1}{1-\gamma} \sum_{s\in S} \sum_{a\in A} \sum_{s_0\in M} M(s_0) \cdot d_{s_0}^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot f(s,a)$

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

$= \sum_{s_0\in M} M(s_0) \cdot \sum_{t=0}^{\infty} \sum_{s\in S} \sum_{a\in A} \gamma^t \cdot f(s,a) \cdot \underbrace{P(s_t=s | s_0, \pi_\theta) \cdot \pi_\theta(a|s)}_{\tau}$

$= E_{\tau\sim P_M^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t \cdot f(s_t, a_t) \right]$

Problem 5

```
Downloading dataset: http://rail.eecs.berkeley.edu/datasets/offline_rl/maze2d/maze2d-umaze-sparse-v1.hdf5
load datafile: 100%|████████████| 8/8 [00:00<00:00, 25.66it/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, 25.72it/s]
```

- 這是由 maze2d-umaze-v1 所得到的 dataset, 我觀察到其是由一個大 list 包住許多的小 list, 且每個小 list 都有四個 elements. 第一個 element 都差不多在 1.08 左右, 第二個 element 從 1.9 慢慢到 2.8, 第三個 element 從 0 到 -4, 第四個 element 沒有什麼規律

```
Downloading dataset: http://rail.eecs.berkeley.edu/datasets/offline_rl/gym_mujoco_v2/walker2d_random-v2.hdf5
load datafile: 100%|████████████| 9/9 [00:01<00:00,  4.71it/s]
[[ 1.2491173e+00  3.2758405e-03  3.5445758e-03 ... -3.2404759e-03
   1.3873719e-03 -3.3544931e-03]
 [ 1.2489719e+00  6.0818223e-03  4.9435999e-03 ...  8.4943724e-01
  -5.0627989e-01  5.1634040e+00]
 [ 1.2480459e+00  3.8860261e-03  3.9840965e-03 ...  3.2524174e-01
  -2.5208545e+00  3.3501066e-02]
 ...
 [ 1.2507056e+00  2.1851482e-03  1.2356882e-03 ...  4.6311049e-03
  -1.2648222e-03 -1.7198029e-03]
 [ 1.2494115e+00  3.5594066e-04 -3.6252677e-05 ...  1.9545360e+00
  -7.1042180e+00  1.0000000e+01]
 [ 1.2474208e+00 -1.3153229e-02 -1.0459671e-02 ...  7.5775456e-01
  -6.7979965e+00  9.4499722e+00]]
load datafile: 100%|████████████| 9/9 [00:02<00:00,  4.23it/s]
```

- 這是由 Walker2d-v2 所得到的 dataset, 我觀察到其和上一個 dataset 皆是由一個大 list 包住許多的小 list, 不過每個小 list 的element 數量較多, 且可以發現某些 element 的 value 很小, 只有 10^{-3} 左右