Problem 1

(a) Proof of Eq (1): $V_*(s) = \max_a Q_*(s,a)$

① $V_*(s) = \max_\pi V^\pi(s)$

$$= \max_\pi \left(\sum_a \pi(a|s) Q^\pi(s,a) \right)$$

└ 對所有Q值取平均 ≤ 取 max over a

$\Rightarrow \sum_a \pi(a|s) Q^\pi(s,a) \le \max_a Q^\pi(s,a)$

$$\le \max_\pi \max_a Q^\pi(s,a) = \max_a \max_\pi Q^\pi(s,a) = \max_a Q_*(s,a)$$

② 對於比較policy的好壞，定義：

$\pi \ge \pi'$ if $V^\pi(s) \ge V^{\pi'}(s)$, $\forall s$

對於 optimal policy π_*，定義：

$\pi_* \ge \pi$, for all π

從①，得到 $V_*(s) \le \max_a Q_*(s,a)$

若 $V_*(s) < \max_a Q_*(s,a)$ 成立，表示我們可以找到一個比 optimal policy 更好的 policy，為在 state s 時執行 action a 使 $V_*(s) < \max_a Q_*(s,a)$ 成立。但這個 policy 不應該存在，因此 $V_*(s) < \max_a Q_*(s,a)$ 不成立。

所以最後可得 $V_*(s) = \max_a Q_*(s,a)$。

Proof of Eq (2): $Q_*(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^a V_*(s')$

$$Q_*(s,a) = \max_\pi Q^\pi(s,a)$$
$$= \max_\pi \left(R_s^a + \gamma \sum_{s'} P_{ss'}^a V^\pi(s') \right)$$
$$= R_s^a + \gamma \sum_{s'} P_{ss'}^a \cdot \max_\pi V^\pi(s')$$
$$= R_s^a + \gamma \sum_{s'} P_{ss'}^a V_*(s') \quad \left(V_*(s) = \max_\pi V^\pi(s) \right)$$

NAN PAO

Problem 1

(b)

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

$$= \max_{s,a} \left| R_s^a + \gamma \sum_{s'} P_{ss'}^a \max_{a'} Q(s',a') - R_s^a - \gamma \sum_{s'} P_{ss'}^a \max_{a'} Q'(s',a') \right|$$

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

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

$$\le \gamma \|Q-Q'\|_\infty$$

$\therefore$ T^* is a γ-contraction operator in terms of ∞-norm.

Problem 2

$$L(\pi)=\sum_{a\in A}\left(\pi(a|s)\,Q_{\Omega}^{\pi_k}(s,a)-\pi(a|s)\log\pi(a|s)\right)-\mu\left(\sum_{a\in A}\pi(a|s)-1\right)$$

and the optimal satisfies $\frac{\partial L(\pi)}{\partial \pi(a|s)}=0$ for every $a\in A$

假設 action 是離散的，則有 $a_1,a_2,a_3\cdots a_n\in A$，

且 optimal 在 $\frac{\partial L(\pi)}{\partial \pi(a_1|s)}=\frac{\partial L(\pi)}{\partial \pi(a_2|s)}=\cdots=\frac{\partial L(\pi)}{\partial \pi(a_n|s)}=0$

將 $L(\pi)$ 的 sigma 寫開。$L(\pi)$ 為以下式子的總和：

$$\begin{cases}\pi(a_1|s)\,Q_{\Omega}^{\pi_k}(s,a_1)-\pi(a_1|s)\log\pi(a_1|s)-\mu\pi(a_1|s)\\ \pi(a_2|s)\,Q_{\Omega}^{\pi_k}(s,a_2)-\pi(a_2|s)\log\pi(a_2|s)-\mu\pi(a_2|s)\\ \vdots\\ \pi(a_n|s)\,Q_{\Omega}^{\pi_k}(s,a_n)-\pi(a_n|s)\log\pi(a_n|s)-\mu\pi(a_n|s)-\mu\end{cases}$$

分別求得對 $\pi(a_1|s),\pi(a_2|s),\cdots,\pi(a_n|s)$ 的偏微分並令其 $=0$

$$\frac{\partial L(\pi)}{\partial \pi(a_1|s)}=Q_{\Omega}^{\pi_k}(s,a_1)-(\log\pi(a_1|s)+1)-\mu=0$$

$$\frac{\partial L(\pi)}{\partial \pi(a_2|s)}=Q_{\Omega}^{\pi_k}(s,a_2)-(\log\pi(a_2|s)+1)-\mu=0$$

$$\vdots$$

$$\frac{\partial L(\pi)}{\partial \pi(a_n|s)}=Q_{\Omega}^{\pi_k}(s,a_n)-(\log\pi(a_n|s)+1)-\mu=0$$

綜合以上關係可得聯立方程式：

$$\begin{cases}\pi(a_1|s)=e^{Q_{\Omega}^{\pi_k}(s,a_1)-\mu-1}\\ \pi(a_2|s)=e^{Q_{\Omega}^{\pi_k}(s,a_2)-\mu-1}\\ \vdots\\ \pi(a_n|s)=e^{Q_{\Omega}^{\pi_k}(s,a_n)-\mu-1}\\ \pi(a_1|s)+\pi(a_2|s)+\cdots+\pi(a_n|s)=1\end{cases}$$

解此方程：

$$e^{Q_{\Omega}^{\pi_k}(s,a_1)-\mu-1}+e^{Q_{\Omega}^{\pi_k}(s,a_2)-\mu-1}+\cdots+e^{Q_{\Omega}^{\pi_k}(s,a_n)-\mu-1}=1$$

$$\xrightarrow[e^{\mu+1}]{\text{同乘}} e^{Q_{\Omega}^{\pi_k}(s,a_1)}+e^{Q_{\Omega}^{\pi_k}(s,a_2)}+\cdots+e^{Q_{\Omega}^{\pi_k}(s,a_n)}=e^{\mu+1}$$

NAN PAO

改寫原聯立方程式:

$$\begin{cases} \pi(a_1|s) = e^{Q_\Omega^{\pi_k}(s,a_1)-\mu-1} = \dfrac{e^{Q_\Omega^{\pi_k}(s,a_1)}}{e^{\mu+1}} \\ \pi(a_2|s) = e^{Q_\Omega^{\pi_k}(s,a_2)-\mu-1} = \dfrac{e^{Q_\Omega^{\pi_k}(s,a_2)}}{e^{\mu+1}} \\ \vdots \\ \pi(a_n|s) = e^{Q_\Omega^{\pi_k}(s,a_n)-\mu-1} = \dfrac{e^{Q_\Omega^{\pi_k}(s,a_n)}}{e^{\mu+1}} \end{cases}$$

將 $e^{\mu+1} = e^{Q_\Omega^{\pi_k}(s,a_1)} + e^{Q_\Omega^{\pi_k}(s,a_2)} + \cdots + e^{Q_\Omega^{\pi_k}(s,a_n)} = e^{\mu+1}$ 代入,

可得:

$$\begin{cases} \pi(a_1|s) = \dfrac{e^{Q_\Omega^{\pi_k}(s,a_1)}}{e^{\mu+1}} = \dfrac{e^{Q_\Omega^{\pi_k}(s,a_1)}}{e^{Q_\Omega^{\pi_k}(s,a_1)} + e^{Q_\Omega^{\pi_k}(s,a_2)} + \cdots + e^{Q_\Omega^{\pi_k}(s,a_n)}} \\ \vdots \\ \pi(a_n|s) = \dfrac{e^{Q_\Omega^{\pi_k}(s,a_n)}}{e^{\mu+1}} = \dfrac{e^{Q_\Omega^{\pi_k}(s,a_n)}}{e^{Q_\Omega^{\pi_k}(s,a_1)} + e^{Q_\Omega^{\pi_k}(s,a_2)} + \cdots + e^{Q_\Omega^{\pi_k}(s,a_n)}} \\ \qquad = \dfrac{e^{Q_\Omega^{\pi_k}(s,a_n)}}{\sum_{a\in A} e^{Q_\Omega^{\pi_k}(s,a)}} \end{cases}$$

改寫上式可得知 $L(\pi)$ 在

$$\pi(\cdot|s) = \frac{e^{Q_\Omega^{\pi_k}(s,\cdot)}}{\sum_{a\in A} e^{Q_\Omega^{\pi_k}(s,a)}}$$ 有最大值。