Question 5:

(1) Let π be a policy, the state-value function satisfies the bellman equation:

$$Q^{\pi}(s,a) = E_{p^{\pi}}[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) \mid s_0 = s, a_0 = a]$$

$$= E_{p^{\pi}}[r(s,a) + \sum_{t=1}^{\infty} \gamma^t r(s_t,a_t) \mid s_0 = s, a_0 = a]$$

$$= r(s,a) + \gamma E_{p^{\pi}}[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1},a_{t+1}) \mid s_0 = s, a_0 = a]$$

$$= r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) E_{p^{\pi}}[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) | s_0 = s', a_0 = a']$$

$$= r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) Q^{\pi}(s',a') = E_{(s',a') \sim p^{\pi}(.|s,a)}[r(s,a) + \gamma Q^{\pi}(s',a')]$$

(2) Now let π^* be the optimal policy, using the result above, we obtain:

$$Q^*(s,a) = E_{(s',a') \sim p^{\pi^*}(.|s,a)}[r(s,a) + \gamma Q^{\pi^*}(s',a')]$$

$$= E_{s' \sim \pi^*(.|s,a)}[r(s,a) + \gamma Q^{\pi^*}(s',\pi^*(s'))]$$

$$= E_{s' \sim \pi^*(.|s,a)}[r(s,a) + \gamma \max_{a'} Q^{\pi^*}(s',a')]$$

$$= E_{s' \sim \pi^*(.|s,a)}[r(s,a) + \gamma \max_{a'} Q^*(s',a')]$$

(3) From the last part, the optimal policy π^* satisfies:

$$E_{s' \sim \pi^*(.|s,a)}[r(s,a) + \gamma \max_{a'} Q^*(s',a') - Q^*(s,a)] = 0$$

Hence, a natural idea is to take the norm of this quantitiy as the loss function:

$$\mathcal{L}(\theta) = E_{s' \sim \pi^*(.|s,a)} \| r(s,a) + \gamma \max_{a'} Q(s',a',\theta) - Q(s,a,\theta) \|^2.$$