Homework 6

Graded

2 Hours, 19 Minutes Late

Student

Jinzhi Shen

Total Points

66.5 / 72 pts

Question 1

1 — **26** / 27 pts

1.1 **a** — **1** / 2 pts

– **0 pts** Correct

✔ – **1 pt** Incorrectly specified type

– **1 pt** Difference incorrect

– **2 pts** Incorrect

1.2 **b** — **3** / 3 pts

✔ – **0 pts** Correct

– **1 pt** Missing Exhaustive search

– **1 pt** Missing policy iteration

– **1 pt** Missing value iteration

1.3 **c** — **6** / 6 pts

✔ – **0 pts** Correct

– **1 pt** No policy graph

– **1.5 pts** One value function incorrect

– **3 pts** Multiple value functions incorrect

– **6 pts** Missing

1.4 **d** — **6** / 6 pts

✔ – **0 pts** Correct

– **1 pt** No policy graph

– **1.5 pts** One value function incorrect

– **3 pts** Multiple value functions incorrect

– **6 pts** Missing

1.5 **e** — **6** / 6 pts

✔ – **0 pts** Correct

– **1 pt** No/Wrong policy graph

– **1.5 pts** One value function incorrect/missing

– **3 pts** Multiple value functions incorrect

– **6 pts** Missing

1.6 f **4** / 4 pts

✔ **– 0 pts** Correct

– 1 pt No explanation

– 2 pts Incorrect policy

– 4 pts Incorrect/Missing

Question 2

Q-Learning [Written] **20.5** / 23 pts

2.1 **a** **5** / 5 pts

✔ **– 0 pts** Correct

– 2.5 pts Major mistake

– 1 pt Minor mistake

– 5 pts Missing

2.2 **b** **5** / 5 pts

✔ **– 0 pts** Correct

– 2.5 pts Major Mistake

– 1 pt Minor Mistake

– 5 pts Missing

2.3 **c** **4** / 4 pts

✔ **– 0 pts** Correct

– 2 pts Mistake in explanation

– 1 pt Minor mistake in explanation

– 4 pts Missing / wrong explanation

2.4 **d** **4** / 4 pts

✔ **– 0 pts** Correct

– 2 pts Mistake in explanation, or lacks details

– 1 pt Minor mistake in explanation

– 4 pts Missing explanation

2.5 **e** **2.5** / 5 pts

– 0 pts Correct

– 1 pt Minor calculation mistake

✔ **– 2.5 pts** Major calculation mistake

2 The tables for each step are expected to be written in your solution.

– 3 pts Incorrect

– 5 pts Missing

Question 3

Q-Learning [Coding] **5** / 5 pts

3.1 **d** **2.5** / 2.5 pts

✔ **– 0 pts** Correct

– 2.5 pts Missing

3.2 **e** **2.5** / 2.5 pts

✔ **– 0 pts** Correct

– 2.5 pts Missing

Question 4

REINFORCE [Written + Coding] **15** / 17 pts

4.1 **a** **3** / 3 pts

✔ **– 0 pts** Correct

– 1 pt Minor mistake

– 2 pts Major mistake

4.2 **b** **2** / 3 pts

– 0 pts Correct

✔ **– 1 pt** Minor Mistake

1

– 2 pts Major mistake

– 3 pts Missing answer

4.3 **c** **2** / 3 pts

– 0 pts Correct

✔ **– 1 pt** Minor mistake

– 2 pts Major Mistake

– 3 pts Missing Answer

4.4 **d** **3** / 3 pts

✔ **– 0 pts** Correct

– 1 pt Minor mistake

– 2 pts Major mistake

4.5 **e** **5** / 5 pts

✔ **– 0 pts** Correct

– 2 pts MLE Distribution not provided

– 1 pt MLE Distribution incorrect

– 2 pts REINFORCE distribution not provided

– 1 pt Reinforce distribution incorrect

– 1 pt Explanation not provided or insufficient

Questions assigned to the following page: 1.5, 1.3, 1.6, 1.1, 1.2, and 1.4

Homework 6

Question 1

(a)

In a deterministic MDP, the same action always results in the same outcome, while in a stochastic MDP, the same action can result in different outcomes.

(b)

1. exhaustive search
2. policy iteration
3. value iteration

(c)

The graph is shown in Figure 1.

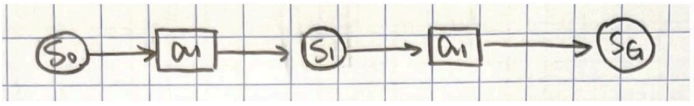

Figure 1: Question 1(c) graph.

$$V^{\pi}(s_1) = 3, \qquad V^{\pi}(s_0) = 5$$

(d)

The graph is shown in Figure 2

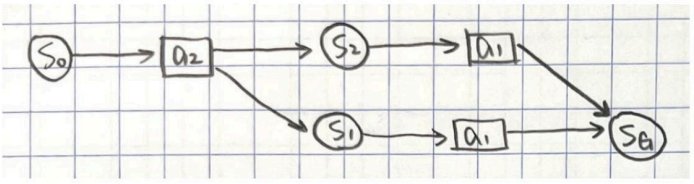

Figure 2: Question 1(d) graph.

$$V^{\pi}(s_2) = 1, \qquad V^{\pi}(s_1) = 3, \qquad V^{\pi}(s_0) = 0.25(4+1) + 0.75(2+3) = 5$$

(e)

The graph is shown in Figure 3

$$\begin{aligned} &V^{\pi}(s_2) = 0.75V^{\pi}(s_0) + 0.25 \times 2 = 0.75V^{\pi}(s_0) + 0.5 \\ &V^{\pi}(s_1) = 3 \\ &V^{\pi}(s_0) = 0.25(4 + V^{\pi}(s_2) + 0.75 \times (2+3)) = 0.25V^{\pi}(s_2) + 4.75 \end{aligned} \tag{1}$$

Questions assigned to the following page: 2.5, 1.6, 2.4, 2.1, 2.2, and 2.3

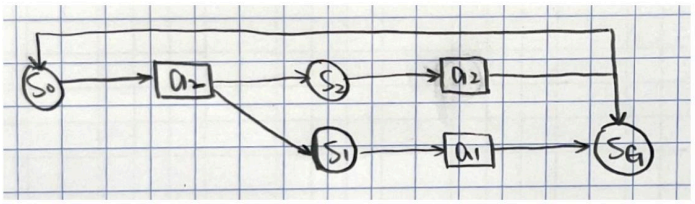

Figure 3: Question 1(e) graph.

By solving the above system of linear equations, we have

$$V^{\pi}(s_2) = 5, \qquad V^{\pi}(s_1) = 3, \qquad V^{\pi}(s_0) = 6$$

(f)

1(c)-1(e) consider all possible policies, which is an exhaustive search. From these three policies, the policy in 1(e):

$$\pi(s_0) = a_2, \qquad \pi(s_2) = a_2$$

has the largest $V^{\pi}(s_0)$, so it is the optimal policy for the MDP given in Fig. 1.

Question 2

(a)

$$Q^*(s,a) = \sum_{s' \in S} P(s'|s,a)[R(s,a,s') + \max_{a' \in A_{s'}} Q^*(s',a')]$$

(b)

Update of the Q-function at the observed state-action pair (s,a):

$$Q(s,a) \longleftarrow (1-\alpha)Q(s,a) + \alpha[r + \gamma \max_{a' \in A_{s'}} Q(s',a')]$$

(c)

It helps balance between exploration and exploitation via randomly selecting between exploration and exploitation. By random sampling of actions, estimated reward values will converge to their true values.

(d)

It makes it possible to remember and reuse the past old experiences. It breaks the correlation between consecutive data.

(e)

At the beginning after the initialization, the Q-table entries are:

$$Q(s_1,a_1) = Q(s_1,a_2) = Q(s_2,a_1) = Q(s_2,a_2) = 0$$

After the first transition:

$$Q(s_1,a_1) = 0.5 \times 0 + 0.5[-10 + 0.5 \times 0] = -5$$

After the second transition:

$$Q(s_1,a_2) = 0.5 \times 0 + 0.5[-10 + 0.5 \times 0] = -5$$

After the third transition:

$$Q(s_2,a_1) = 0.5 \times 0 + 0.5[18.5 + 0.5 \times (-5)] = 8$$

After the fourth transition:

$$Q(s_1,a_2) = 0.5 \times (-5) + 0.5[-10 + 0.5 \times 8] = 0.5$$

Questions assigned to the following page: 3.1, 4.1, 4.2, 3.2, and 2.5

The resulted Q-table entries are:

$$Q(s_1, a_1) = -5, \quad Q(s_1, a_2) = 0.5, \quad Q(s_2, a_1) = 8, \quad Q(s_2, a_2) = 0$$

The optimal policy after having observed the four transitions is:

$$\pi(s_1) = argmax_{a \in A_{s_1}} Q^*(s_1, a) = a_2, \quad \pi(s_2) = argmax_{a \in A_{s_2}} Q^*(s_2, a) = a_1$$

Question 3

(d)

The figure for this question is shown in Fig 4.

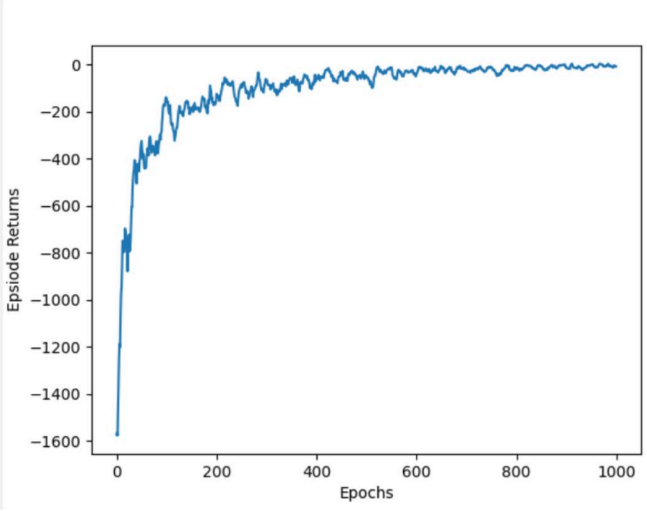

Figure 4: Question 3(d) graph.

(e)

The figures for this question are shown in Fig 5, Fig 6 and Fig 7.

Question 4

(a)

The cost function is

$$C(\theta) = \sum_{y \in \mathcal{D}} log\ p_\theta(y) = \sum_{y \in \mathcal{D}} \left(F_\theta(y) - log\ \left(\sum_{\hat{y} \in \mathcal{D}} exp(F_\theta(\hat{y})) \right) \right)$$

Its gradient is

$$\frac{\partial C(\theta)}{\partial \theta} = \sum_{y \in \mathcal{D}} \left(\frac{\partial F_\theta(y)}{\partial \theta} - \frac{\sum_{\hat{y} \in \mathcal{D}} exp(F_\theta(\hat{y})) \frac{\partial F_\theta(\hat{y})}{\partial \theta}}{\sum_{\hat{y} \in \mathcal{D}} exp(F_\theta(\hat{y}))} \right) \tag{2}$$

(b)

For each sample which follows the probability distribution $p_\theta(y)$, its expected utility is $E_{p_\theta}[R(y)]$, so the expected utility for N samples is

$$N \cdot E_{p_\theta}[R(y)] = N \cdot \sum_{y \in \mathcal{Y}} p_\theta(y) R(y)$$

which enables us to approximate the utility with N samples from the probability distribution $p_\theta(y)$

Questions assigned to the following page: 4.2 and 4.3

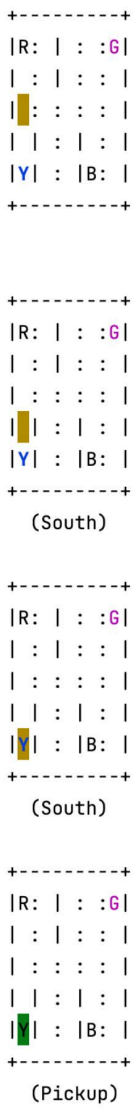

Figure 5: Question 3(e)1 graph.

(c)

The gradient of the utility $U(\theta)$ w.r.t. θ:

$$\begin{aligned}
\nabla_\theta U(\theta) &= \sum_{y \in \mathcal{Y}} \frac{\partial p_\theta(y)}{\partial \theta} \cdot R(y) \\
&= \sum_{y \in \mathcal{Y}} p_\theta(y) \frac{1}{p_\theta(y)} \frac{\partial p_\theta(y)}{\partial \theta} \cdot R(y) \\
&= \mathbb{E}_{p_\theta} \left[\frac{1}{p_\theta(y)} \frac{\partial p_\theta(y)}{\partial \theta} \cdot R(y) \right] \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \frac{1}{p_\theta(\tilde{y}_i)} \frac{\partial p_\theta(\tilde{y}_i)}{\partial \theta} \cdot R(\tilde{y}_i)
\end{aligned} \tag{3}$$

Questions assigned to the following page: 4.4 and 4.3

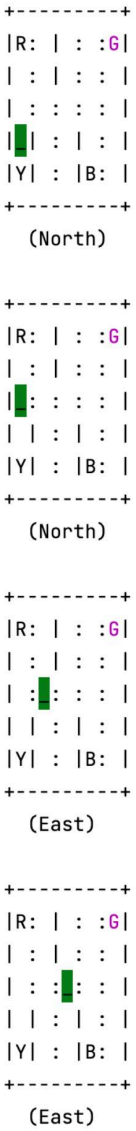

Figure 6: Question 3(e)2 graph.

(d)

$$
\begin{aligned}
\textit{Approximated gradient} &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{p_\theta(\tilde{y}_i)}\frac{\partial p_\theta(\tilde{y}_i)}{\partial\theta}\cdot R(\tilde{y}_i) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{expF_\theta(\tilde{y}_i)\frac{\partial F_\theta(\tilde{y}_i)}{\partial\theta}\sum_{\hat{y}\in\mathcal{Y}}exp(F_\theta(\hat{y})) - expF_\theta(\tilde{y}_i)\sum_{\hat{y}\in\mathcal{Y}}exp(F_\theta(\hat{y}))\frac{\partial F_\theta(\hat{y})}{\partial\theta}}{(\sum_{\hat{y}\in\mathcal{Y}}exp(F_\theta(\hat{y})))^2} \\
&\quad\cdot\frac{R(\tilde{y}_i)}{p_\theta(\tilde{y}_i)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{expF_\theta(\tilde{y}_i)\frac{\partial F_\theta(\tilde{y}_i)}{\partial\theta}\sum_{\hat{y}\in\mathcal{Y}}exp(F_\theta(\hat{y})) - expF_\theta(\tilde{y}_i)\sum_{\hat{y}\in\mathcal{Y}}exp(F_\theta(\hat{y}))\frac{\partial F_\theta(\hat{y})}{\partial\theta}}{(\sum_{\hat{y}\in\mathcal{Y}}exp(F_\theta(\hat{y})))^2} \\
&\quad\cdot\frac{R(\tilde{y}_i)\cdot\sum_{\hat{y}\in\mathcal{Y}}expF_\theta(\hat{y})}{expF_\theta(\tilde{y}_i)} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial F_\theta(\tilde{y}_i)}{\partial\theta} - \frac{\sum_{\hat{y}\in\mathcal{Y}}expF_\theta(\hat{y})\frac{\partial F_\theta(\hat{y})}{\partial\theta}}{\sum_{\hat{y}\in\mathcal{Y}}expF_\theta(\hat{y})}\right)\cdot R(\tilde{y}_i)
\end{aligned}
\tag{4}
$$

Questions assigned to the following page: 4.4 and 4.5

Compared to the result obtained in part (a), this gradient also considers the reward function $R(y)$. It weighs the term in part (a)'s result with reward function $R(y)$.

(e)

(i)

The distribution learned with the maximum likelihood approach is:
tensor([0.0900, 0.1590, 0.2450, 0.2520, 0.1750, 0.0790])

(ii)

The distribution learned with the REINFORCE approach is:
tensor([1.3905e-04, 2.4305e-04, 9.9809e-01, 1.2220e-03, 1.6481e-04, 1.4586e-04])

(iii)

For the maximum likelihood approach, the goal is to find the ground truth probability distribution, so the learned distribution is close to the ground truth distribution, which is expected

For the REINFORCE approach, the goal is to maximize the utility function, which is the expected value of the reward $R(y)$. For this goal, higher probability should be assigned to y that has high reward $R(y)$ value. The reward for $y = 3, 4$ is the highest. The learned probability distribution assigns almost all probability to these two y values (larger than 0.999 probability), which is expected.

Question assigned to the following page: 4.5

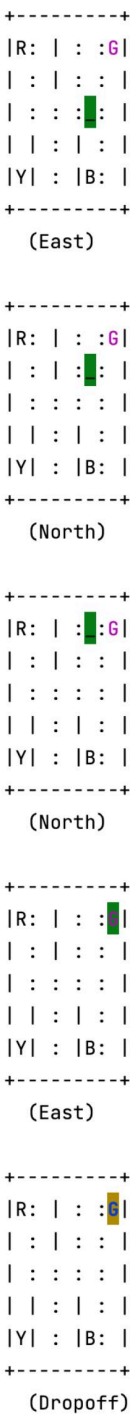

Figure 7: Question 3(e)3 graph.