DDA 4230: Reinforcement learning **CUHKSZ**

Assignment 1

TA: Sheng Xu

Due Date: Oct. 9th, 11:59 pm
Total points available: 100 pts.

Note: *Please note that external references are allowed only if you give appropriate reference. There is no required format of reference. Please elaborate on your answers as well. (not just give a number etc.)*

Problem 1: Markov Decision Process [20 pts.]

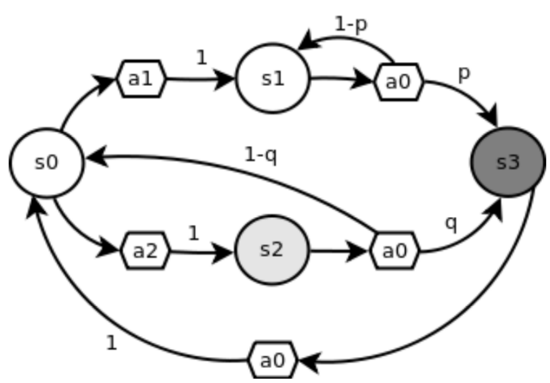

Figure 1: MDP for Problem 1. States are represented by circles and actions by hexagons. p, q denotes the transition probability and $p, q \in [0, 1]$. The reward is 10 for state s_3, 1 for state s_2 and 0 otherwise.

For this question, consider the infinite horizon (where time $t = \infty$) MDP represented by Figure 1 with discount factor $\gamma \in (0, 1)$.

1. List all the possible policies. [4 pts.]

2. Show the equations representing the optimal value functions for all states, including $V^*(s_0), V^*(s_1), V^*(s_2)$ and $V^*(s_3)$. [6 pts.]

For example, for $V^*(s_0)$, the representation is:

$$V^*(s_0) = \max_{a \in \{a_1, a_2\}} 0 + \gamma \sum_{s'} P(s' \mid s, a) V^*(s') = \gamma \max \{V^*(s_1), V^*(s_2)\} \tag{1}$$

3. Is there a value for p such that for all $\gamma \in (0,1)$ and $q \in [0,1], \pi^*(s_0) = a_2$? Explain. [5 pts.]

4. Is there a value for q such that for all $\gamma \in (0,1)$ and $p \in [0,1], \pi^*(s_0) = a_1$? Explain. [5 pts.]

Problem 2: Fixed Point [25 pts.]

Recall from lecture that the optimal value function can be written as:

$$V^*(s_t) = \max_a \ \mathbb{E}\left[r(s_t, a) + \gamma V^*(s_{t+1}) | a_t = a\right] .$$

Let $r(s,a)$ denotes the reward received of choosing action a on state s and $P(s'|s,a)$ be the probability of transiting to state s' when choosing action a on state s. The bellman operator can be defined as $B : \mathbb{R}^S \longrightarrow \mathbb{R}^S$, where

$$(BV)(s) = \max_a \ r(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s') .$$

In this problem, we are going to show a few properties about the bellman operator. We'll see that if we know the transition function $P(s'|s,a)$ and the reward function $r(s,a)$, then repeating this operator on our value function helps recover the optimal value function, in other words, we want to show that value iteration will converge to a unique fixed point V regardless of the starting point. An element V is a fixed point for an operator B (in this case the Bellman operator) if performance of B on V returns V, i.e., $BV = V$.

1. Contraction Property [5 pts.]

Prove that the Bellman operator B is a contraction operator for $\gamma \in (0,1)$ with respect to the infinity norm $\|\cdot\|_\infty$ (you may want to search up the definition for this). Specifically, we want to prove that

$$\|BV - BV'\|_\infty \leq \gamma \|V - V'\|_\infty \tag{2}$$

for any two value functions V and V', meaning if we apply it to two different value functions, the distance between value functions (in the ∞ norm) shrinks after application of the operator to each element.

2. Lead-up Proof [10 pts.]

According to the above contraction property, there are some helpful lead-up proofs for you to obtain final proof of the fixed point (i.e., optimal value function).

2.1. Let's define $\|V_{n+1} - V_n\|_\infty = \|BV_n - BV_{n-1}\|_\infty$, please prove by induction that $\|V_{n+1} - V_n\|_\infty \leq \gamma^n \|V_1 - V_0\|_\infty$.

2.2 Prove that for any $c > 0$, $\|V_{n+c} - V_n\|_\infty \leq \frac{\gamma^n}{1-\gamma}\|V_1 - V_0\|_\infty$

3. Fixed Point [5 pts.]

A *Cauchy sequence* is a sequence whose elements become arbitrarily close to each other as the sequence progresses. Formally a sequence $\{a_n\}$ in metric space X with distance metric d is a Cauchy sequence if given an $\epsilon > 0$ there exists k such that if m, n $> k$ then $d(a_m, a_n) < \epsilon$. Real Cauchy sequences are convergent. Using this information about Cauchy sequences, argue that the sequence $V_0, V_1, ...$ is a Cauchy sequence and is therefore convergent and must converge to some element V and this V is a fixed point.

4. Uniqueness [5 pts.]

Show that this fixed point is unique.

Problem 3: Practice of bandit problem [55 pts.]

Consider a two-armed bandit problem, where each arm's distribution is Bernoulli. Consider the following three problem variants, with respective Bernoulli distribution parameters specified for each arm:

Problem	Arm 1	Arm 2
P1	0.8	0.6
P2	0.8	0.7
P3	0.55	0.45
P4	0.5	0.5

Write a Python program to simulate each of the above bandit problems. Specifically, for each problem, you need to do:

1. Choose the horizon n as 10000.
2. For each algorithm, repeat the experiment 100 times.
3. Store the number of times an algorithm plays the optimal arm, for each round $t = 1, \cdots, n$.
4. Store the regret in each round $t = 1, \cdots, n$.
5. Plot the percentage of optimal arm played and regret against the rounds $t = 1, \cdots, n$,
6. For each plot, add standard error bars.

Do the above for the following bandit algorithms:

1. Explore-then-Commit (ETC) algorithm [20 pts.]

The explore-then-commit (ETC) algorithm with exploration parameter k chosen optimally so that the gap-dependent regret is minimum (this choice for k would require information about underlying gap).

2. ETC algorithm with a heuristic [20 pts.]

The ETC algorithm with a heuristic choice for exploration parameter k. Try different values for k and summarize your findings, say by tabulating regret for different k.

3. Interpretation and summarization [15 pts.]

Interpret the numerical results and submit your conclusions. In particular, discuss the following:

1. Explain the results obtained for ETC with optimal k and correlate the results to the theoretical findings.
2. Explain the results obtained for ETC with a heuristic choice for k. In particular, how does ETC with a k that is far from the optimal, perform?

Note that you need to submit: 1) Source code, preferably one that is readable with some comments; 2) Plots/tabulated results in a document (or you could submit printouts of plots); 3) Discussion of the results - either hand-written or typed-up.