Homework 4 of CS 165A (Spring 2023)

University of California, Santa Barbara

To be discussed on May 31, 2023 (Wednesday)

Notes:

- The homework is optional. You do not need to submit your solutions anywhere and you will not be evaluated by these.
- To maximize your learning, you should try understanding the problems and try solving them as much as you can before the discussion class.
- Feel free to discuss with your peers / form small groups to solve these problems.
- Feel free to discuss any questions with the instructor and the TA in office hours or on Piazza.

Why should I do this homework?

This optional homework provide practices for you to deepen your understanding on Reinforcement Learning (Problem 1 and Problem 2) and Logic (Problem 3,4,5).

Problem 1 (a) makes sure you understand the UCB rules and how it helps with bounding regret in multi-arm bandits problems. Problem 1(b) extrapolates the idea of optimism a bit for RL (not covered in the lectures!). Problem 1(c)(d)(e) involve coding up a simple (3 states, 2 action, deterministic) MDP environment and run Q-learning on it. Problem 2 gives an example that helps you to draw connections between multi-armed bandit to the problem of clinical trials, i.e., average treatment effect estimation. It also asks you to construct estimators based on the plug-in principle — "empirical average" as an estimate of the "mean", i.e., monte carlo estimates, of the value functions.

Problem 3 provide examples for you to understand the new jargons in logic and logic agents. Problem 4 talks about inferences in propositional logic. Problem 5 asks you to translates natural sentences into their first order logic description.

If you are time constrained, it is fine to skip Problem 1 and 2 (but do attend the discussion on Wednesday to gain insight) since you have already done quite a bit to those in Project 3. I strongly encourage everyone to work out Problem 3-5, because there will be questions similar to these in the final.

Problem 1 Reinforcement Learning Algorithms

*The easiest way to solve the problem is by writing Python code for Part (c) (d) (e) (f).

(a) After finishing tth round of a multi-armed bandit algorithm, the agent obtained an an estimator $\hat{r}_t$ of the vector of the expected reward (In our lecture we named $\hat{r}_t$ as Q_t). Suppose $\hat{r}_t$ obeys that

$$|\hat{r}_t(a) - r(a)| \leq \epsilon_t \text{ for all action } a \in \mathcal{A}$$

for $\mathcal{A} = \{1, 2, 3, 4, 5, ..., K\}$. Which action will the agent take in the $(t+1)$st round if the agent follows the Upper Confidence Bound(UCB) strategy?

Show that the (immediate, rather than cumulative) expected regret that the agent incurs at time $t+1$ is at most $2\epsilon_t$.

(b) (Challenge question) Let $\hat{Q}(s, a)$ be an approximation of $Q^*(s, a)$ in the sense that

$$\forall s \in \mathcal{S}, \forall a \in \mathcal{A},\ |\hat{Q}(s, a) - Q^*(s, a)| \leq \epsilon$$

Let $\hat{\pi}$ be a deterministic policy such that $\hat{\pi}(s) = \text{argmax}_a \hat{Q}(s, a)$ for all $s \in \mathcal{S}$. Show that for all $s \in \mathcal{S}$

$$V^*(s) - V^{\hat{\pi}}(s) \leq \frac{2\epsilon}{1-\gamma}.$$

(*This question is a bit more difficult and it is out of the scope of the course.But it should be intuitively clear that why should we care about this for model-free RL.

Hint 1: Follow the Bellman equation and apply the given condition of approximation.

Hint 2: Use the same arguments from Part (a).

Hint 3: Use geometric series.)

(c) (Q-learning with Greedy Actions) Consider the simple MDP environment in Figure 1. Let us initiate $Q_0(s, a) = 0$ for all s, a and run Q-learning for 20 iterations with learning rate parameter $\alpha = 0.5$. At the tth iteration, the action A_t will be taken as the greedy actions using Q_t (when tied, use $A_t = a_1$).

Write a python program that runs Q-learning as instructed above and print out $Q_t(s, a)$ for $t = 1, 2, 3, ..., 20$ (each one of them is a matrix of size 3×2).

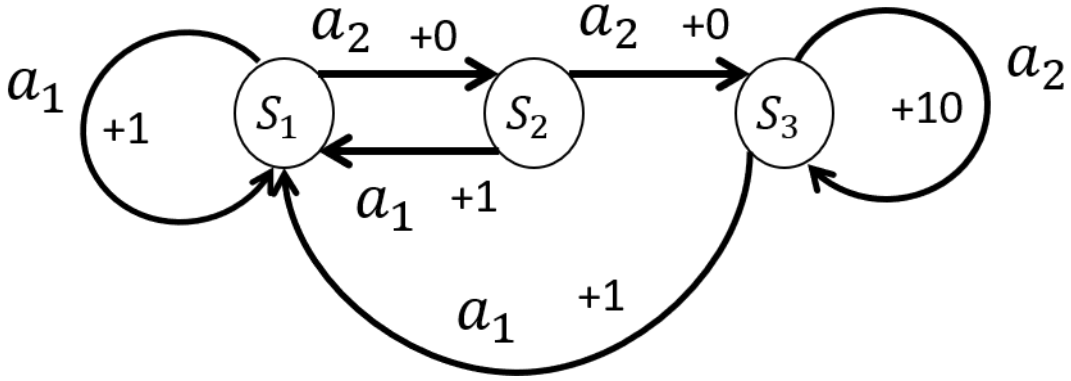


Figure 1: The state-space diagram of a simple infinite-horizon tabular MDP with 3 states and 2 actions. The transitions are deterministic, indicated by the edges in the state-space diagram. The corresponding immediate reward is given by the number on the edge. The initial state is s_1. Assume $\gamma = 0.9$.

(d) (Q-learning with Exploratory Policy) For the same environment, run Q-learning, but instead of using the greedy policy, use uniformly random actions.

Let the sequence of random bits you use for implementing the random policy be

```
1111101010000110001100011101000111110011...
```

(you will not use up these random bits)

When the bit is 0, choose action a_1 when the bit is 1, choose action a_2.

Again, write a Python of program to compute $Q_t(s, a)$ for $t = 1, 2, 3, ..., 20$.

(e) (Challenge question) Calculate the expectation of the cumulative discounted rewards of the algorithm in Part (c), Part (d) when we run them indefinitely.

(Hint: 1. Are the policy being executed changing over time in Part (c) and Part (d)?

2. for the calculations for the randomized policy, it is often helpful to calculate the transition matrix P^π under this policy π and then use the matrix version of the Bellman equation:

$$(I - \gamma[P^\pi]^T)V^\pi = r^\pi.$$

You can solve a linear system of equations of the form $Ax = b$ using numpy with:

```
x = np.linalg.solve(A, b)
```

)

(f) (Challenge question) Calculate the values of the greedy policies of Q_{20} from Part (c) and Part (d) respectively (assume that we start running them from scratch). What is the value of the optimal policy for this environment?

(Hint: 1. You may use Part (e) or use the first principle definition to calculate the value function of a policy (then substitute the initial state to get the value of a policy — a scalar).

2. To check if a policy is optimal, it suffices to check the Bellman Optimality Equation.)

Problem 2 Bandits for clinical trials

In standard clinical trials, the doctor tosses a (biased) coin when a patient arrives. And the patient receives the *new drug* (Treatment group) if the coin turns up "Head", or the patient will receive a *placebo* (control group) if the coin turns up "Tail".

The goal of clinical trial is to study whether the new drug is useful by estimating the so-called Average Treatment Effect (ATE), defined as follows:

$$\text{ATE} = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$$

In the above, T is a binary random variable denoting the treatment assignment. $T = 1$ indicates that the actual drug is given and $T = 0$ indicates that the patient is assigned to a control group where placebos are given instead. Y is a binary random variable denoting whether the patient has recovered.

Y depends both on the treatment T and unknown health conditions U of the patient, as indicated in Figure 2a.

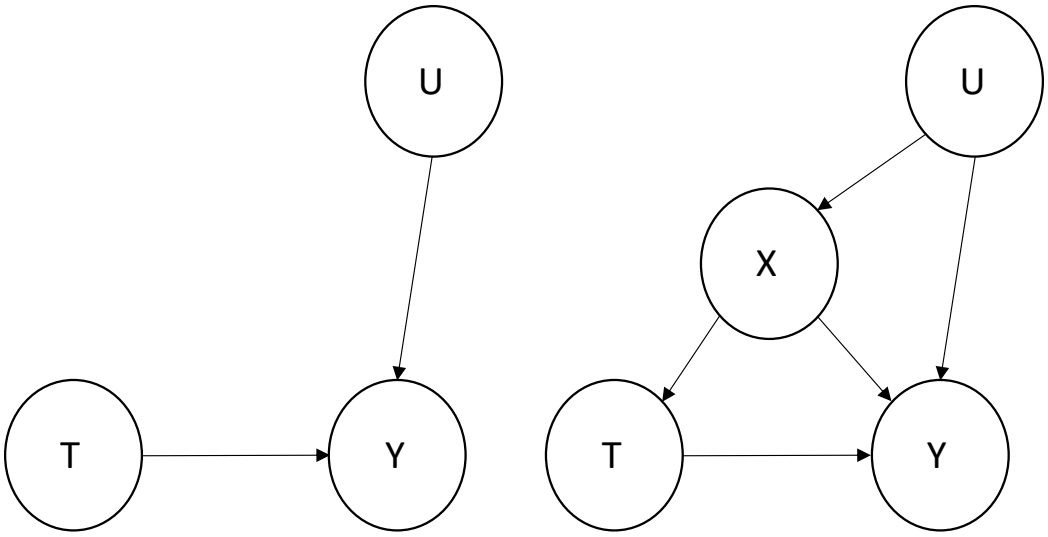
(a) If we consider the treatment T the action, Y the reward, then the problem described above is closely related to a two-armed bandit problem. Describe the "policy" π that the doctor runs, and rewrite the ATE as the value difference of two policies π_1 and π_2. Clearly state what these policies are.

(b) Let the patients be drawn independently from an identical distribution, i.e., $U_1, ..., U_n \sim \mathcal{D}$ i.i.d., and the doctors collect a dataset $(T_1, Y_1), ..., (T_n, Y_n)$. One can construct a plug-in (or Monte Carlo) estimate of the value of π_1 and π_2 from Part (a) hence construct an estimator by taking their differences

$$\widehat{\text{ATE}}_{\text{plug-in}} = \widehat{V}^{\pi_1} - \widehat{V}^{\pi_2}.$$

Assume both the treatment and the control are chosen at least once in the dataset (that is to say you can condition on the event that $\sum_{i=1}^n \mathbf{1}(T_i = 0) > 0$ and $\sum_{i=1}^n \mathbf{1}(T_i = 1) > 0$), show that this estimator is unbiased, i.e.,

$$\mathbb{E}\left[\widehat{\text{ATE}}_{\text{plug-in}} \middle| \sum_{i=1}^n \mathbf{1}(T_i = 0) > 0, \sum_{i=1}^n \mathbf{1}(T_i = 1) > 0\right] = \text{ATE}.$$



(a) BayesNet of basic clinical trial (b) BayesNet of advanced clinical trial

Figure 2: The Probabilistic Graphical models of the variables in clinical trials.

Note that we are not making any assumptions about the distribution of U and how it interacts with Y. Also, notice that we are not using parameter p.

(c) (Not covered in the class, but straightforward.) Next, consider another way of estimating ATE

$$\widehat{\text{ATE}}_{IS} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}(T_i = 1)}{p}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}(T_i = 0)}{1-p}Y_i$$

where IS stands for importance sampling.

Prove, using the definition of the expectation, that the expectation of this estimator is also equal to ATE.

(Hint: You may wish to use the provided graphical model to factorize the joint probability distribution.)

(d) (Not covered in the class, but straightforward.) Can you rewrite your expression from Part (b) into a form similar to that of Part (c)?

(Hint: Change p into something else!)

(e) Suppose you are to use a 2-armed bandit algorithm, e.g., UCB, rather than tossing a random coin in each round. Rewrite the *regret* of this 2-armed bandit problem in the notation of the ATE problem (in terms of random variables Y, T.)

(Hint: Think about what the optimal action is in this problem)

Explain in layman's term (so a doctor can understand) what a *sublinear* regret bound mean in this setting.

(f) (Challenge question) More advanced clinical trials determines the probability of Treatment and Control group assignment by taking patient's feature vector X into account,

i.e., $\mathbb{P}[T = 1|X] = \pi(X)$ and $\mathbb{P}[T = 0|X] = 1 - \pi(X)$, where $\pi : \mathcal{X} \to (0, 1)$. The new BayesNet describing the dependence of the random variables are given in Figure 2b.

Design the ATE estimators analogous to that in Part (b) and Part (c).

Problem 3 Logic Agent, Models in Propositional Logic

(a) Suppose the agent has progressed to the point shown in Figure 7.4(a), (page 239 in AIMA 3rd Ed and Page 213 in AIMA 4th Ed), having perceived nothing in [1,1], a breeze in [2,1], and a stench in [1,2], and is now concerned with the contents of [1,3], [2,2], and [3,1]. Each of these can contain a pit, and at most one can contain a wumpus. Following the example of Figure 7.5 (in AIMA Textbook), construct the set of possible worlds. (You should find 32 of them.) Mark the worlds in which the KB is true and those in which each of the following sentences is true:

α_2 = "There is no pit in [2,2]."

α_3 = "There is a wumpus in [1,3]."

Hence show that KB $\models \alpha_2$ and KB $\models \alpha_3$.

(b) Given the following, can you prove that the unicorn is mythical? How about magical? Horned?

If the unicorn is mythical, then it is immortal, but if it is not mythical, then it is a mortal mammal. If the unicorn is either immortal or a mammal, then it is horned. The unicorn is magical if it is horned.

Problem 4 Logical Inference with CNF and Resolution.

Consider the following sentence:

$$[(\text{Food} \Rightarrow \text{Party}) \vee (\text{Drinks} \Rightarrow \text{Party})] \Rightarrow [(\text{Food} \wedge \text{Drinks}) \Rightarrow \text{Party}].$$

(a) Determine, using enumeration, whether this sentence is valid, satisfiable (but not valid), or unsatisfiable.

(b) Convert the left-hand and right-hand sides of the main implication into CNF, showing each step, and explain how the results confirm your answer to (a).

(c) Prove your answer to (a) using resolution.

Problem 5 First Order Logic.

Consider a vocabulary with the following symbols:

Occupation(p, o): Predicate. Person p has occupation o.

Customer (p1, p2): Predicate. Person p1 is a customer of person p2.

Boss(p1, p2): Predicate. Person p1 is a boss of person p2.

Doctor, Surgeon, Lawyer, Actor: Constants denoting occupations.

Emily, Joe: Constants denoting people.

Use these symbols to write the following assertions in first-order logic:

(a) Emily is either a surgeon or a lawyer.

(b) Joe is an actor, but he also holds another job.

(c) All surgeons are doctors.

(d) Joe does not have a lawyer (i.e., is not a customer of any lawyer).

(e) Emily has a boss who is a lawyer.

(f) There exists a lawyer all of whose customers are doctors.

(g) Every surgeon has a lawyer.