E1 277 REINFORCEMENT LEARNING

Date: 25th of April, 2022 Time: 1400 to 1700 hrs

Semester: Vasanth/Spring 2022 Maximum Points: 30

Instructions:

(1) Please neatly write down the answers on blank sheets of paper. **It is recommended that you write the answer to each question on a new page.**

(2) Once you are ready to submit, you will need to scan the answer sheets in to a pdf. This pdf will then need to be uploaded to **gradescope**.

(3) There are **04 problems** in this question paper that you need to answer. Simply writing down the final answer will not get you the full points. Instead, you must give explicit reasons for each statement that you make in your submission. Your explanation should be good enough to make the TA understand how you arrived at your answer. There are marks for partial solutions.

(4) Mention all the assumptions you are making. Existing result from any literature can be used directly in your proof; however, please explicitly cite the same.

(5) Use a calculator when necessary.

(6) Do not cheat. Any academic misconduct will be dealt with strictly using all permissible laws of the institute.

(1) Suppose a person has n jobs to perform one after the other and needs to find the optimal order to schedule these jobs. The ith job requires a random time X_i for its execution and assume that $X_1, \ldots, X_n$ are all independent. If the ith job gets completed at time t, the person earns $\alpha^t R_i$, $i = 1, \ldots, n$, where $\alpha \in (0, 1)$ is a constant. No reward is earned from the job while it is being processed. Consider now two specific orderings of jobs: $O_1 = (i_1, \ldots, i_k, i, j, i_{k+3}, \ldots, i_n)$ and $O_2 = (i_1, \ldots, i_k, j, i, i_{k+3}, \ldots, i_n)$, respectively.

(a) Find $E[(\text{return from } O_1) - (\text{return from } O_2) \mid \sum_{l=1}^{k} X_{i_l} = t]$. [03 pts]

(b) Obtain a verifiable sufficient condition for

$$E[\text{ return from } O_1] \leq E[\text{return from } O_2]$$

using the above. [02 pts]

(c) Guess the form of the optimal policy and argue why it is optimal? [02 pts]

(2) Recall the infinite horizon discounted reward problem, where under any given policy π, the value function is defined by

$$J_\pi(i) = E_\pi \left[\sum_{n=0}^{\infty} \gamma^n R(X_n, a_n) \mid X_0 = i \right].$$

Do not assume the rewards $R(i, a)$ to be uniformly bounded. Instead, suppose that for each i, there exist numbers $B_i > 0$ and a constant $k > 0$ such that starting in i, we have

$$E_\pi[|R(X_{n-1}, a_{n-1})| \mid X_0 = i] \leq B_i n^k, \text{ for all policies } \pi.$$

(a) Under the above condition, show J_π is well defined for every policy π? [02 pts]

(b) Let f be the policy chosen by the optimality equation and J^* denote the optimal value function. Show that

$$J^*(i) = E_f[n\text{-stage return} \mid X_0 = i] + \gamma^n E_f[J^*(X_n) \mid X_0 = i],$$

for every $n \geq 1$? [02 pts]

(c) Show that $\lim_{n\to\infty} \gamma^n |E_f[J^*(X_n) \mid X_0 = i]| = 0$? [02 pts]

(d) Now argue that as $n \to \infty$, $J^*(i) = J_f(i)$? [02 pts]

(3) For a random variable X with a strictly positive density function, its value-at-risk at $\alpha \in [0, 1]$ is the number $v \in \mathbb{R}$ such that

$$\mathbb{P}\{X \leq v\} = \alpha.$$

Consider the MDP $(\mathcal{S}, \mathcal{A}, \mathbb{P}, \gamma, r)$ and let μ be a stationary policy. Let $v \in \mathbb{R}^{|\mathcal{S}|}$ be the vector whose s-th coordinate $v(s)$ denotes the value-at-risk associated with the random variable

$$X(s) := \sum_{t=0}^{\infty} \gamma^t r(s_t),$$

where $s_0 = s$ and, for $t \geq 0$, $a_t \sim \mu(\cdot|s_t)$ and $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$. Note that the reward at time t is a deterministic function of the state at time t.

(a) Design an algorithm for estimating v using linear function approximation. You may assume that the given feature matrix is Φ and that $X(s)$ has a strictly positive density function for all $s \in \mathcal{S}$. (**Note**: You don't need to discuss the convergence of this algorithm, but each step leading to your design should be properly justified.) [05 pts]

(b) Discuss if your algorithm is implementable in the spirit of the TD(0) algorithm for policy evaluation with linear function approximation. [02 pts]

(4) Consider the MDP $(\mathcal{S}, \mathcal{A}, \mathbb{P}, \gamma, r)$ and let μ be a stationary policy with value function J_μ. Suppose the initial state s_0 is sampled from an arbitrary initial distribution and, thereafter, $a_t \sim \mu(\cdot|s_t)$ and $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$. Suppose we use the hypothetical algorithm

$$w_{t+1} = w_t + \alpha_t[U_{t+1} - \phi(s_t)^\top w_t]\phi(s_t)$$

for policy evaluation with linear function approximation, where $\phi(s)$ is the feature vector associated with state s, α_t is some stepsize, and U_{t+1} is random variable that, conditional on the past, has an expected value of $J_\mu(s_t)$ and finite second moment. Discuss the convergence of this algorithm. You may assume that the iterates of this algorithm are stable. Explicitly state all other assumptions. [08 pts]