Total Marks: 25, Total Time: 6 hrs

Instructions

1. Work on your own. You can discuss with your classmates on the problems, use books or web. However, the solutions that are submitted must be your own and you must acknowledge all the sources (names of people you worked with, books, webpages etc., including class notes.) Failure to do so will be considered cheating. Identical or similar write-ups will be considered cheating as well.
2. In your submission, add the following declaration at the outset:
   "I pledge that I have not copied or given any unauthorized assistance on this exam."
3. The exam is divided into two sections. For the first section, either provide a proof or disprove using a counterexample. For the second section, provide a detailed answer, showing all the necessary steps.

I Prove or disprove

2 1. Recall from the class notes the projected fixed point equation $\Phi r^* = \Pi T_\pi(\Phi r^*)$ for a given policy π. Here $T_\pi(J) = g + \alpha PJ$ is the contraction mapping underlying policy π in a discounted MDP. The remaining notation is as in the class notes. Let J_π denote the expected cost (or value function) corrresponding to policy π, and let $S = \{\Phi r \mid r \in \mathbb{R}^d\}$ denote the linear space, where we find an approximation to J_π.

   Suppose that $J_\pi \in S$. Then, $J_\pi = \Phi r^*$.

II Long answer problems

5 2. **Discounted MDP**
   Consider a machine repair problem in an infinite horizon discounted MDP setting with discount factor $\alpha < 1$. In each stage, the machine can either be running or broken. If it is running, it provides a reward of 1 unit, i.e., a cost of -1 unit. A running machine may break down w.p. 0.1. If it is broken, then it can be repaired with a cost a^2, where $a \in [0, 1]$. Following a repair, the machine will start working w.p. a, and remain broken w.p. $1 - a$.

   Formulate this problem in the discounted MDP framework, and determine the optimal policy for repair.

3. **Approximate policy evaluation**
   Recall from the class notes that the projected fixed point equation $\Phi r^* = \Pi T_\pi(\Phi r^*)$ can be written in matrix form as follows:

$$Cr^* = d, \text{ where } C = \Phi^\top D(I - \alpha P)\Phi, \text{ and } d = \Phi^\top Dg,$$

   where the quantities Φ, D, P, g are as defined in the class notes.

   Under the assumption that Φ has linearly independent columns, show that

3 (a) C is positive definite in the following sense: $x^\top Cx > 0$ for all $x \neq 0$;

2 (b) The eigenvalues of C have positive real parts.

4. A gambler engages in a game of successive i.i.d. flipping of a coin that falls heads w.p.$p \in (0,1)$. When the coin turns up heads, the gambler wins a rupee. When the coin turns up tails, the gambles loses all his/her earnings so far.

   Answer the following:

2 (a) Model this problem as a Markov chain with a state variable that denotes the current earnings of the gambler. Show that this chain is positive recurrent by exhibiting a stationary distribution.

2 (b) Suppose there are two actions for the gambler in each round. The first action is to stop and take home the current earnings, and the second action is to continue flipping. Formulate this problem as an infinite horizon discounted MDP with discount factor $\alpha < 1$. Write down the Q-values, and the associated Bellman equation.

2 (c) For the MDP in the part above, specify the update rule of the Q-learning algorithm, using the following template:

$$Q_{t+1}(i) = (1-\beta_t)Q_t(i) + \beta_t(HQ_t(i) + w_t(i)).$$

2 (d) Let ξ denote the stationary distribution obtain in the first part above, and let $\|\cdot\|_\xi$ denote the weighted ℓ_2 norm using the stationary distribution ξ. Show that the mapping HQ underlying the Q-learning algorithm is a contraction w.r.t. $\|\cdot\|_\xi$, i.e.,

$$\left\|HQ - HQ'\right\|_\xi \leq \alpha \left\|Q - Q'\right\|_\xi.$$

5. Consider the problem of Monte Carlo policy evaluation in the context of a stochastic shortest path (SSP) problem. Fix a proper policy π, and obtain l MDP trajectories using this policy. Use the every visit variant for policy evaluation, i.e., for a given state i, the estimate $\tilde{J}_l(i)$ of the expected cost $J_\pi(i)$ is formed by averaging the cost samples $\{\hat{J}(i,m,k), m = 1,\ldots,n_k, k = 1,\ldots,l\}$. Here n_k is the number of visits to state i in trajectory k, for $k = 1,\ldots,l$ and $\hat{J}(i,m,k)$ is sum total of the costs starting from the mth visit point until the end of the trajectory.

   Answer the following:

1.5 (a) Is $\tilde{J}_l(i)$ an unbiased estimate of $J_\pi(i)$? Justify your answer.

3.5 (b) Prove a result in the spirit of the strong law, i.e.,

$$\tilde{J}_l(i) \to J_\pi(i) \text{ a.s. as } l \to \infty.$$