Total Marks: 25, Total Time: 3 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. In the first section, the first question requires no justification, while the second question requires a proof or a counterexample. In the second section, provide a detailed answer, showing all the necessary steps for each problem.

I Short answer questions

[1] 1. Policy iteration algorithm for solving MDPs was proposed by

(a) Blackwell (b) Bellman

(c) Howard (d) Bertsekas

[3] 2. Prove or disprove: In a discounted-cost MDP, let π_1 and π_2 be two policies with corresponding expected cost (or value function) J_{π_1} and J_{π_2}, respectively. Consider a policy π_3 formed using π_1, π_2 as follows:

$$\pi_3(i) = \begin{cases} \pi_1(i) & \text{if } J_{\pi_1}(i) \leq J_{\pi_2}(i), \\ \pi_2(i) & \text{otherwise.} \end{cases}$$

Then, $J_{\pi_3} \leq \min(J_{\pi_1}, J_{\pi_2})$.

II Long answer problems

3. **Discounted MDP**
Consider an MDP with a finite state space $\mathcal{X} = \{1, \cdots, n\}$. Consider a deterministic and bounded reward function $r : \mathcal{X} \to \mathbb{R}$. For each $x \in \mathcal{X}$, a fixed policy π determines a stochastic transition to a subsequent state $x' \in \{\mathcal{X}\}$ with probability $\mathbb{P}(x' \mid x)$. We denote by x_t the state at time t, where $t = 0, 1, 2, \ldots$.

The cumulative discounted reward $R \in \mathbb{R}$ is defined by

$$R = \sum_{t=0}^{\infty} \alpha^t r(x_t).$$

Define the value function $J(\cdot)$ and the variance of the cumulative discounted reward as follows:

$$J(x) = \mathbb{E}[R \mid x_0 = x], \text{ and } V(x) = \text{Var}[R \mid x_0 = x], \forall x \in \mathcal{X}.$$

Also, define the second moment of the cumulative discounted reward as

$$M(x) = \mathbb{E}[R^2 \mid x_0 = x], \forall x \in \mathcal{X}.$$

Answer the following:

[5] (a) For any $x \in X$, express $J(x)$ and $M(x)$ in terms of a fixed-point equation.

[3] (b) Show that

$$\forall x \in X, V(x) = \psi(x) + \alpha^2 \sum_{x' \in \mathcal{X}} \mathbb{P}(x' \mid x) V(x'), \text{ where}$$

$$\psi(x) = \alpha^2 \left(\sum_{x' \in \mathcal{X}} \mathbb{P}(x' \mid x) J(x')^2 - \left(\sum_{x' \in \mathcal{X}} \mathbb{P}(x' \mid x) J(x') \right)^2 \right).$$

4. **Discounted MDP**
Consider a machine that can be one of the following two states: 'good' and 'bad'. If the machine is in a good state in the current period, then it will transition to a bad state in the next period with probability (w.p.) p_1. On the other hand, a machine in 'bad' state in the current period has to undergo maintenance, and transitions to a 'good' state in the next period w.p. p_2, and remains in the 'bad' state w.p. $(1 - p_2)$. Suppose the machine in 'good' state earns A INR (i.e., a cost of $-A$ INR) per-period, while the per-period maintenance cost for a machine in 'bad' state is B INR.

Answer the following:

[4] (a) Formulate this problem in the infinite horizon discounted cost framework. What is the expected discounted cost of the machine for each possible initial state, i.e., 'good' and 'bad'?

[1] (b) Suppose a new machine starts in the 'good' state, and costs M INR. Compare the purchase cost to the expected discounted cost to infer when it is optimal to a buy a new machine.

5. Consider the problem of approximate policy evaluation using the LSTD algorithm in the context of a discounted MDP. Fix a policy π, and obtain a single sample path of the underlying MDP using this policy. Let $i_0, \ldots, i_k$ denote this sample path. With notation as in the class notes, the LSTD algorithm solves the following linear system:

$$C_k r_k = d_k, \text{ where } C_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)(\phi(i_t) - \alpha\phi(i_{t+1}))^\intercal, d_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t) g(i_t, i_{t+1}).$$

Let r^* denote the projected fixed point, i.e.,

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

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

Answer the following:

[5] (a) Show that r_k converges asymptotically to r^*, i.e.,

$$r_k \to r^* \quad \text{almost surely as} \quad k \to \infty.$$

Provide a rigorous proof with all the necessary details.

[3] (b) Consider the following LSTD variant:

$$\tilde{r}_k = \tilde{C}_k^{-1} d_k, \text{ where } \tilde{C}_k = (k+1) \left(\sum_{t=0}^{k} \phi(i_t)(\phi(i_t) - \alpha\phi(i_{t+1}))^\intercal + \mathcal{I} \right)^{-1},$$

where $\mathcal{I}$ is the d-dimensional identity matrix, and d_k is as in regular LSTD.
Does $\tilde{r}_k$ converge to r^* asymptotically? Justify your answer.