EE 556
The Exam

In class.
80 minutes. 50 points.

Your Name: ______________________

Instructions: Please write your name on each page of your answer script. No collaboration is allowed. Closed book except for 1 page cheat sheet. Please state your reasoning clearly.

1. Consider a Markov decision process on state space $\mathcal{X} = \{1, ..., N\}$, action space $\mathcal{U} = \{1, ..., M\}$, and reward function $r(x, u)$. π will denote a stationary policy and state transition probability kernel is $P_{xy}(u)$. We consider infinite horizon cost with discount factor $\gamma \in (0, 1)$. Define an operator G:
$$(Gz)(x, u) = r(x, u) + \gamma \sum_y P_{xy}(u) \max_{u'} z(y, u'),$$
where $z : X \times U \to \Re$.
(i) [8] Show that G is a contraction map w.r.t. $|| \cdot ||_\infty$.
(ii) [4] Suppose that the optimal Q^* function,
$$Q^*(x, u) = \max_\pi \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(x_t, u_t) | x_0 = x, u_0 = u],$$
satisfies $GQ^* = Q^*$. Will a solution exist, and will it be unique? Give an algorithm to find it.
(iii) [8] Now suppose that there is another operator $\tilde{G}$, also a contraction, such that
$$||Gz - \tilde{G}z|| \leq \epsilon, \ \forall z.$$
Can we say something about the relationship between the fixed points of G and $\tilde{G}$?

.

2. (i) [6] Consider the γ-discounted-reward infinite horizon MDP defined above with $r(x,u) \leq \bar{R}$. Determine a K such that if the Q-Value Iteration algorithm, $Q_{k+1} = GQ_k$ is stopped after K iterations, $||Q_K - Q^*||_\infty < \epsilon$, where $\epsilon > 0$ is given.
(ii) [4] Suppose that in the QVI algorithm, the expectation is replaced by $\frac{1}{n}\sum_{l=1}^{n} \max_{u'} Q(y'_l, u')$ where $(y'_1, \cdots, y'_n)$ are independent samples of the next state from state x under action u, and we get an operator $\hat{G}_n$:

$$(\hat{G}_n z)(x,u) = r(x,u) + \gamma \frac{1}{n} \sum_{l=1}^{n} \max_{u'} z(y_l, u').$$

Show that $\hat{G}Q$ provides an unbiased estimate of GQ, i.e., $\mathbb{E}[\hat{G}_n Q] = GQ$.

.

3. Consider a robot that starts at the origin $(0, 0)$, and is to reach the goal at $(10, 10)$. The state space is discrete, $\mathbb{S} = [-10 : 10, -10 : 10]$. At each point (other than the goal state), the robot can take one of four actions $\mathbb{A} = \{U, D, L, R\}$ for up, down, left and right respectively. The robot moves in the direction of intended motion with probability $p > 1/2$, and with probability $(1 - p)/3$, it moves in one of the other three directions. We note that if the robot is at one of the edges, and the action causes the robot to move off the edge, it will end up on the opposite edge. For example, if the robot is at (-10,5), an L action will move it to (10,5). Similarly, if the robot is at (5,10), a U action will cause it to move to (5,-10). Once the robot reaches the goal state $G = (10, 10)$, it gets a large reward $\bar{R}$. Otherwise, the reward is zero. Consider a discounted criterion.
(i) [8] Model this problem as an MDP: identify the state space, the control space, the transition kernel, observations, objective, etc.
(ii) [6] Write down a DP algorithm to find the optimal policy.
(iii) [6] Identify the optimal policy using your intuition as best as you can.

.