Stanford CME 241 (Winter 2019) - Final Exam

1. **5 points**: Assume you have data in the form of just the following 5 complete episodes for an MRP. Non-terminal *State*s are labeled A and B, the numbers in the episodes denote *Reward*s, and all states end in a terminal state T.

   - A 2 A 6 B 1 B 0 T
   - A 3 B 2 A 4 B 2 B 0 T
   - B 3 B 6 A 1 B 0 T
   - A 0 B 2 A 4 B 4 B 2 B 0 T
   - B 8 B 0 T

   Given only this data and experience replay (repeatedly and endlessly drawing an episode at random from this pool of 5 episodes), what is the Value Function *Every-Visit Monte-Carlo* will converge to, and what is the Value Function TD(0) (i.e., one-step TD) will converge to? Assume discount factor $\gamma = 1$. Note that your answer (Value Function at convergence) should be independent of step size.

2. **5 points**: Consider an MDP with an infinite set of states $\mathcal{S} = \{1, 2, 3, \ldots\}$. The start state is $s = 1$. Each state s allows a continuous set of actions $a \in [0, 1]$. The transition probabilities are given by:

   $$\Pr[s+1 \mid s, a] = a, Pr[s \mid s, a] = 1 - a \text{ for all } s \in \mathcal{S} \text{ for all } a \in [0,1]$$

   For all states $s \in \mathcal{S}$ and actions $a \in [0, 1]$, transitioning from s to $s+1$ results in a reward of $1 + a$ and transitioning from s to s results in a reward of $1 - a$. The discount factor $\gamma = 0.5$.

   - Calculate the Optimal Value Function $V^*(s)$ for all $s \in \mathcal{S}$
   - Calculate an Optimal Deterministic Policy $\pi^*(s)$ for all $s \in \mathcal{S}$

3. **3 points**: Tabular Monte-Carlo RL update for the n^{th} sample of a state s is given by:

   $$V_n(s) \leftarrow V_{n-1}(s) + \alpha(G_n - V_{n-1}(s))$$

   where G_n is the sample episode return following the n^{th} sample of state s, $V_n(s)$ is the Value Function estimate after the n^{th} update, and α is the step-size of the update. Assume that we initialize with $V_0(s) = 0$.

   Show that as $n \to \infty$, $V_n(s)$ can be formulated as an exponentially-decaying weighted average of the sample returns $G_n, G_{n-1}, \ldots G_1$. What are the precise weights associated with the sample returns in the weighted average? Show that the weights indeed sum to 1 as $n \to \infty$.

4. **7 points**: Consider a finite MDP with the set of states denoted as $\mathcal{S}$ and a set of actions denoted as $\mathcal{A}$. Let π be an ϵ-greedy policy. Let π' be the ϵ-greedy policy imputed from the Action-Value function Q_π (ϵ-greedy *Policy Improvement* from π to π'), i.e.,

$$\pi'(a \mid s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}|} & \text{if } a = \arg\max_{b \in \mathcal{A}} Q_\pi(s, b) \\ \frac{\epsilon}{|\mathcal{A}|} & \text{otherwise} \end{cases}$$

Prove that:

$$\sum_{a \in \mathcal{A}} \pi'(a \mid s) \cdot Q_\pi(s, a) \geq V_\pi(s) \text{ for all } s \in \mathcal{S}$$

where V_π is the State-Value function for policy π.

5. **3 points**: I've mentioned in class that RL with tabular Value Function is a special case of RL with linear function approximation for the Value Function. Linear function approximation can be expressed as: $V(s) = \sum_{i=1}^{n} \phi_i(s) \cdot w_i$ where $w_i, 1 \leq i \leq n$, are the parameters of the linear function approximation and $\phi_i(\cdot), 1 \leq i \leq n$, are the feature functions. For the case of RL with tabular Value Function, what are the values of parameters w_i and what are the feature functions $\phi_i(\cdot)$?

6. **5 points**: Assume we have a finite action space $\mathcal{A}$. Let $\phi(s, a) = (\phi_1(s, a), \phi_2(s, a), \ldots, \phi_n(s, a))$ be the features vector for any $s \in \mathcal{S}, a \in \mathcal{A}$. Let $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ be an n-vector of parameters. Let the action probabilities conditional on a given state s and given parameter vector θ be defined by the softmax function on the linear combination of features: $\theta^T \cdot \phi(s, a)$, i.e.,

$$\pi(a \mid s; \theta) = \frac{e^{\theta^T \cdot \phi(s, a)}}{\sum_{b \in \mathcal{A}} e^{\theta^T \cdot \phi(s, b)}}$$

- Evaluate the score function $\nabla_\theta \log \pi(a \mid s, \theta)$
- Construct the Action-Value function approximation $Q(s, a; w)$ so that the following key constraint of the Compatible Function Approximation Theorem (for Policy Gradient) is satisfied:

  $$\nabla_w Q(s, a; w) = \nabla_\theta \log \pi(a \mid s; \theta)$$

  where w defines the parameters of the function approximation of the Action-Value function.
- Show that $Q(s, a; w)$ has zero mean for any state s, i.e. show that

  $$\mathbb{E}_\pi[Q(s, a; w)] \text{ defined as } \sum_{a \in \mathcal{A}} \pi(s, a) \cdot Q(s, a; w) = 0$$

7. **12 points**: We want to develop a model to validate the classical Theory of Derivatives Pricing/Hedging empirically (using Reinforcement Learning) for the simple case of an European Derivative expiring at time T with payoff $g(S_T)$, where S_t is the underlying stock price at time t. Specifically, we want to identify the appropriate portfolio (at any time t) of the stock S_t (with holding α_t) and a risk-free asset R_t (with holding β_t) that would replicate the Derivative payoff. Formally, this replication requirement is:

$$\alpha_T S_T + \beta_T R_T = g(S_T) \text{ for all values of } S_T$$

Assume that $R_t = e^{rt}$ for a given constant risk-free rate r. Assume current stock price S_0 and expiration time T are given. Assume you don't have a formulaic description of the stochastic process for S_t, but you have a simulator for generating S_u, given S_t, for any $u > t \geq 0$. Assume the payoff function $g(\cdot)$ is given (eg: payoff for European Call Option is $g(S_T) = \max(S_T - K, 0)$ where K is the strike price).

We make a key assumption (from knowledge of Pricing Theory) that the Derivative can be replicated by a dynamic continuous-time rebalancing of holdings α_t, β_t (as the stock price evolves stochastically) without any addition or removal of wealth at any time $t > 0$, specified formally as the following *Balance Constraint*:

$$\alpha_t S_{t+dt} + \beta_t R_{t+dt} = \alpha_{t+dt} S_{t+dt} + \beta_{t+dt} R_{t+dt} \text{ for all } 0 \leq t < T$$

Our goal is to identify (using Reinforcement Learning):

- Initial Holdings (α_0, β_0), and
- Dynamic Rebalancing Rule $(\alpha_t, \beta_t) \rightarrow (\alpha_{t+dt}, \beta_{t+dt})$ for all $0 \leq t < T$ under satisfaction of the *Balance Constraint*

so that the option payoff $g(S_T)$ is replicated by $\alpha_T S_T + \beta_T R_T$ *for all values of* S_T. Achieving this goal means the Option Price is $\alpha_0 S_0 + \beta_0 R_0$ and the dynamic holdings (α_t, β_t) provide the Dynamic Hedging Strategy, that can be validated against the results from Derivatives Pricing/Hedging Theory (up to a time-discretized approximation).

For the purposes of this exam question, you have the following two tasks:

- **MDP Modeling**: Provide a precise description of a continuous-time, continuous-states, continuous-actions MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, whose Optimal Policy would yield the above-mentioned Initial Holdings and Dynamic Rebalancing Rule.
- **Algorithm for Optimal Policy**: Describe the technical details of an Optimal Control RL algorithm customized to a time-discretized version of this MDP (you don't need to write Python code, but provide sufficient details of the algorithm).