Stanford CME 241 (Winter 2019) - Final Exam Solutions

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.

   Answer:

   Every-Visit Monte Carlo averages the returns starting from each occurrence of each of the states across all episodes. Therefore, the Every-Visit Monte Carlo Value Function estimate (with experience replay) would converge to:

   $$V(A) = \frac{9 + 7 + 11 + 6 + 1 + 12 + 10}{7} = 8$$

   $$V(B) = \frac{1 + 8 + 2 + 10 + 7 + 12 + 6 + 2 + 8 + 5 * 0}{9 + 5} = 4$$

   TD(0) essentially constructs an MDP with transition probabilities and reward function estimated from the one-step transitions and sample rewards seen in the data, and its Value Function estimate is the Value Function of that estimated MDP.

   The transition probabilities would be estimated as:

   $$P(A \to A) = \frac{1}{7}, P(A \to B) = \frac{6}{7}$$

   $$P(B \to A) = \frac{3}{14}, P(B \to B) = \frac{6}{14}, P(B \to T) = \frac{5}{14}$$

   The reward function would be estimated as:

   $$R(A) = \frac{2 + 6 + 3 + 4 + 1 + 0 + 4}{7} = \frac{20}{7}$$

   $$R(B) = \frac{1 + 2 + 2 + 3 + 6 + 2 + 4 + 2 + 8 + 5 * 0}{14} = \frac{15}{7}$$

The MDP with these transition probabilities and reward function leads to the following Bellman Equations:

$$V(A) = \frac{20}{7} + \frac{1}{7}V(A) + \frac{6}{7}V(B) \Rightarrow 3V(A) - 3V(B) = 10$$

$$V(B) = \frac{15}{7} + \frac{3}{14}V(A) + \frac{6}{14}V(B) \Rightarrow -3V(A) + 8V(B) = 30$$

This yields:

$$V(A) = \frac{34}{3}$$

$$V(B) = 8$$

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 $s \in \mathcal{S}$

Answer:

Since this is an infinite horizon MDP and since each state has identical state transition probabilities and identical reward function, each state would have the same value for the optimal state-value function. Let us refer to this common value as V^*. Invoking Bellman Optimality Equation, we get:

$$V^* = \max_{a \in [0,1]} \{a((1+a) + \frac{V^*}{2}) + (1-a)((1-a) + \frac{V^*}{2})\}$$

Moving V^* from the RHS to the LHS, we get:

$$V^* - \frac{V^*}{2} = \max_{a \in [0,1]} \{2a^2 - a + 1\}$$

$$\Rightarrow V^* = \max_{a \in [0,1]} \{4a^2 - 2a + 2\}$$

For $a \in [0, 1]$, the RHS maximizes for $a = 1$. So the Optimal Policy is $\pi^*(s) = 1$ for all states $s \in \mathcal{S}$. Substituting for $a = 1$, the Optimal Value Function is given by:

$$V^* = 4(1^2) - 2(1) + 2 = 4$$

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 $n \to \infty$.

Answer:

$$V_n(s) \leftarrow V_{n-1}(s) + \alpha(G_n - V_{n-1}(s)) = \alpha G_n + (1-\alpha)V_{n-1}(s)$$
$$= \alpha G_n + (1-\alpha)(\alpha G_{n-1} + (1-\alpha)V_{n-2}) = \alpha G_n + \alpha(1-\alpha)G_{n-1} + (1-\alpha)^2 V_{n-2}$$
$$\ldots$$
$$\ldots$$
$$= \alpha G_n + \alpha(1-\alpha)G_{n-1} + \ldots + \alpha(1-\alpha)^{n-1}G_1 + (1-\alpha)^n V_0$$
$$= \alpha G_n + \alpha(1-\alpha)G_{n-1} + \ldots + \alpha(1-\alpha)^{n-1}G_1$$
$$= \sum_{i=0}^{n-1} \alpha(1-\alpha)^i G_{n-i}$$
$$= \sum_{i=0}^{n-1} w_i G_{n-i}$$

where $w_i, 0 \le i \le n-1$ are the weights associated with the returns $G_n, G_{n-1}, \ldots, G_1$. As $n \to \infty$, we see that these are exponentially decaying weights that sum to 1.

$$\sum_{i=0}^{\infty} w_i = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i = \alpha \sum_{i=0}^{\infty} (1-\alpha)^i = \alpha \cdot \frac{1}{1-(1-\alpha)} = 1$$

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:

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

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

Answer:

$$\sum_{a\in\mathcal{A}} \pi'(a \mid s) \cdot Q_\pi(s,a) = \frac{\epsilon}{m} \sum_{a\in\mathcal{A}} Q_\pi(s,a) + (1-\epsilon) \max_{a\in\mathcal{A}} Q_\pi(s,a)$$

Now we make the crucial observation that a max over choices of $\mathcal{A}$ is greater than or equal to a weighted average over choices of $\mathcal{A}$. Specifically,

$$\max_{a\in\mathcal{A}} Q_\pi(s,a) \geq \sum_{a\in\mathcal{A}} w_a \cdot Q_\pi(s,a)$$

for any choice of weights $w_a \geq 0, a \in \mathcal{A}$ constrained by $\sum_{a\in\mathcal{A}} w_a = 1$. We will make a specific choice of w_a as follows:

$$w_a = \frac{\pi(a \mid s) - \frac{\epsilon}{m}}{1-\epsilon}$$

We note that $w_a \geq 0$ for all $a \in \mathcal{A}$ because $\pi(a \mid s) \geq \frac{\epsilon}{m}$ (since $\pi(a \mid s)$ is an ϵ-greedy policy). We also note that

$$\sum_{a\in\mathcal{A}} w_a = \frac{\sum_{a\in\mathcal{A}} \pi(a \mid s) - \sum_{a\in\mathcal{A}} \frac{\epsilon}{m}}{1-\epsilon} = \frac{1-\epsilon}{1-\epsilon} = 1$$

Having established that

$$\max_{a\in\mathcal{A}} Q_\pi(s,a) \geq \sum_{a\in\mathcal{A}} \frac{\pi(a \mid s) - \frac{\epsilon}{m}}{1-\epsilon} \cdot Q_\pi(s,a)$$

we can go back to the initial equation and state that:

$$\frac{\epsilon}{m} \sum_{a\in\mathcal{A}} Q_\pi(s,a) + (1-\epsilon) \max_{a\in\mathcal{A}} Q_\pi(s,a) \geq \frac{\epsilon}{m} \sum_{a\in\mathcal{A}} Q_\pi(s,a) + (1-\epsilon) \sum_{a\in\mathcal{A}} \frac{\pi(a \mid s) - \frac{\epsilon}{m}}{1-\epsilon} \cdot Q_\pi(s,a)$$

Therefore,

$$\begin{aligned}
\sum_{a\in\mathcal{A}} \pi'(a \mid s) \cdot Q_\pi(s,a) &\geq \frac{\epsilon}{m} \sum_{a\in\mathcal{A}} Q_\pi(s,a) + (1-\epsilon) \sum_{a\in\mathcal{A}} \frac{\pi(a \mid s) - \frac{\epsilon}{m}}{1-\epsilon} \cdot Q_\pi(s,a) \\
&= \frac{\epsilon}{m} \sum_{a\in\mathcal{A}} Q_\pi(s,a) + \sum_{a\in\mathcal{A}} \pi(a \mid s) \cdot Q_\pi(s,a) - \frac{\epsilon}{m} \sum_{a\in\mathcal{A}} Q_\pi(s,a) \\
&= \sum_{a\in\mathcal{A}} \pi(a \mid s) \cdot Q_\pi(s,a) \\
&= V_\pi(s)
\end{aligned}$$

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 $\phi_i(\cdot)$?

   Answer: If the state space is $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ and the Value Function for state s_i is denoted by $V(s_i)$ for all $1 \leq i \leq n$, we have n feature functions $\phi_1(\cdot), \phi_2(\cdot), \ldots, \phi_n(\cdot)$ and n weights $w_1, w_2, \ldots w_n$ given by: $\phi_i(s_j) = \mathbf{1}_{i=j}$ and $w_i = V(s_i)$ for all $1 \leq i \leq n$

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

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

Answer:

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

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

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

$$= \phi_i(s,a) - \sum_{b \in \mathcal{A}} \pi(b \mid s;\theta) \cdot \phi_i(s,b)$$

$$= \phi_i(s,a) - \mathbb{E}_\pi[\phi_i(s,\cdot)]$$

Therefore,

$$\nabla_\theta \log \pi(a \mid s,\theta) = \phi(s,a) - \mathbb{E}_\pi[\phi(s,\cdot)]$$

To satisfy the key constraint $\nabla_w Q(s,a;w) = \nabla_\theta \log \pi(a \mid s;\theta)$ of the Compatible Function Approximation Theorem, we let the features of $Q(s,a;w)$ be $\nabla_\theta \log \pi(a \mid s,\theta)$ and we set $Q(s,a;w)$ to be linear in these features:

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

Finally,

$$\sum_{a \in \mathcal{A}} \pi(a \mid;\theta) \cdot Q(s,a;w) = \sum_{a \in \mathcal{A}} \pi(a \mid s;\theta) \cdot w^T \cdot \nabla_\theta \log \pi(a \mid s,\theta)$$

$$= \sum_{a \in \mathcal{A}} w^T \cdot \nabla_\theta \pi(a \mid s,\theta)$$

$$= w^T \cdot \nabla_\theta (\sum_{a \in \mathcal{A}} \pi(a \mid s,\theta))$$

$$= w^T \cdot \nabla_\theta 1$$

$$= 0$$

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).

Answer:

MDP Modeling:

The start state (denoted s_0) is a special state defined by the current time t_0 and the current stock price S_0, i.e., $s_0 = (t_0, S_0)$. The actions allowable for this start state s_0 are all pairs (α_0, β_0) with both $\alpha_0, \beta_0 \in \mathbb{R}$.

The remaining set of states (non-start states) are defined by the triple (t, S_t, W_t) for all $0 < t \leq T$, where S_t is the stock price at time t and W_t is the portfolio value at time t. The actions allowable

for these non-start states are defined by the stock holdings $\alpha_t \in \mathbb{R}$ that we rebalance the portfolio to, at time t. The risk-free asset holdings β_t are automatically given to us by the *Balance Constraint* as:

$$\beta_t = \frac{W_t - \alpha_t S_t}{R_t} \text{ where } R_t = e^{rt}$$

The states for which $t = T$ will be the terminal states.

The state transition probabilities for $0 \leq t < T$ are given by:

$$\begin{aligned}
&Pr[(t + dt, S_{t+dt}, W_{t+dt}) \mid (t, S_t, W_t), \alpha_t] \\
&= Pr[(t + dt, S_{t+dt}, \alpha_t S_{t+dt} + \beta_t R_{t+dt}) \mid (t, S_t, W_t), \alpha_t] \\
&= Pr[(t + dt, S_{t+dt}, \alpha_t S_{t+dt} + \frac{W_t - \alpha_t S_t}{R_t} R_{t+dt}) \mid (t, S_t, W_t), \alpha_t] \\
&= Pr[(t + dt, S_{t+dt}, \alpha_t S_{t+dt} + (W_t - \alpha_t S_t) e^{r \cdot dt} \mid (t, S_t, W_t), \alpha_t]
\end{aligned}$$

So the above conditional probability for next state, given current state and action, is simply given by the stochastic process of the stock price: $Pr[S_{t+dt} \mid S_t]$.

Since we'd like for the portfolio to replicate the derivative payoff $g(S_T)$ for all S_T, our optimization goal is to minimize the squared difference between the portfolio value W_T and the derivative payoff $g(S_T)$. So we will set the Reward function to be 0 for the start state s_0 as well as for all states (t, S_t, W_t) where $0 < t < T$, and set the Reward function to be $-(W_T - g(S_T))^2$ for the terminal states (T, S_T, W_T). Note that maximizing $-(W_T - g(S_T))^2$ will minimize $(W_T - g(S_T))^2$, which is what we want.

We set the discount factor arbitrarily to $\gamma = 1$ since there is only one reward at $t = T$ and hence, γ has no effect.

Algorithm for Optimal Policy:

We discretize time into n intervals of width $\frac{T}{n}$ with the discrete time points labeled as $i = 0, 1, \ldots n$ such that $t = \frac{iT}{n}$. Henceforth, we will time-index with i instead of t. Since we have discretized in time, the optimal value function will not be 0 (as would be the case in continuous time). Instead, the optimal value function will always produce a negative value that minimizes $(W_T - g(S_T))^2$. This also means that the optimal policy will only be an approximation to the formulas for the portfolio holdings from Derivatives Pricing/Hedging Theory. However, we should be able to observe how the optimal policy (and hence, the price and hedging strategy) converge to the theoretical formulas as we keep increasing n.

Since the action space is continuous, we will employ the Policy Gradient algorithm. We represent the Stochastic Policy with Actor function approximations for the mean and variance of a gaussian distribution for the action $\alpha_i \mid (i, S_i)$ for all $0 \leq i < n$. The parameters of the function approximations for the action mean (call it μ_i at time step $0 \leq i < n$) and the action variance (call it σ_i^2 at time step $0 \leq i < n$) will be updated based on the Policy Gradient Theorem: "Step Size times Score times Q-value Critic". Let us call their respective set of parameters θ_1 (for $\mu_i \mid (i, S_i)$) and θ_2 (for $\sigma_i^2 \mid (i, S_i)$). The Q-value Critic will also have a function approximation (let us call it's parameters as w, which simultaneously get updated) and the Critic's inputs at time step i will be the State at time step i, i.e., (i, S_i, W_i). We can use $TD(\lambda)$ to update both the Actor (θ_1, θ_2) and Critic (w) parameters.

Note that there is one more free parameter to solve for: β_0. We will model this also with a gaussian distribution with mean μ and variance σ^2. We update μ and σ^2 based on Policy Gradient Theorem. The score will be with respect to μ and σ^2 (i.e., analytical gradient of gaussian distribution probabilities of β_0 with respect to mean μ and variance σ^2). The Q-value Critic's inputs in this case will be $(0, S_0, \alpha_0 S_0 + \beta_0)$.

For each episode, we do the following:

- Start the episode with current values of $\theta_1, \theta_2, w, \mu, \sigma^2$
- The stock prices for the episode are simulated as $S_0, S_1, \ldots, S_n$.
- β_0 is sampled from $\mathcal{N}(\mu, \sigma^2)$
- At time step i $(0 \leq i < n)$:
  - μ_i is obtained from the Actor function approximation with parameters θ_1 with input (i, S_i)
  - σ_i^2 is obtained from the Actor function approximation with parameters θ_2 with input (i, S_i)
  - α_i is sampled from $\mathcal{N}(\mu_i, \sigma_i^2)$
  - Reward $R_i = 0$
  - $W_{i+1} = \alpha_i S_{i+1} + (W_i - \alpha_i S_i) e^{\frac{rT}{n}}$
- Reward R_n at time step n is: $-(W_n - g(S_n))^2$
- Using Policy Gradient with $TD(\lambda)$, we update the parameters $\theta_1, \theta_2, w, \mu, \sigma^2$