Practice Problems: Reinforcement Learning

1. An important quantity in the policy gradient setting (as with Chapter 13 of the Text Book) that appears in many algorithms is $\nabla \ln \pi_\theta(a \mid s) = \dfrac{\nabla \pi_\theta(a \mid s)}{\pi_\theta(a \mid s)}$.

   (a) Suppose $\pi_\theta(a \mid s)$ has the form $\pi_\theta(a \mid s) = \dfrac{\theta^T x_{s,a}}{\sum_b \theta^T x_{s,b}}$ where $\theta \geq 0$. Also, $x_{s,a} \geq 0$ are associated state-action features. Find $\nabla \ln \pi_\theta(a \mid s)$ in this case?

   (b) Let $\pi_\theta(a \mid s)$ have a Gaussian parameterization of the form

   $$\pi_\theta(a \mid s) = \frac{1}{\sigma(s,\theta)\sqrt{2\pi}} \exp\left(-\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}\right),$$

   where μ and σ are parameterized mean and standard deviation of the Gaussian. Here, the action space is the whole of $\mathcal{R}$. Here, the action space is the whole of $\mathcal{R}$.
   Let us assume that the parameter θ has two components such that $\theta = (\theta_\mu, \theta_\sigma)^T$ where θ_μ is the parameter for the mean while θ_σ is the parameter for the standard deviation, respectively. In particular, let

   $$\mu(s,\theta) = \theta_\mu^T x_\mu(s) \text{ and } \sigma(s,\theta) = \exp(\theta_\sigma^T x_\sigma(s)),$$

   where $x_\mu(s)$ and $x_\sigma(s)$ are certain state features. Let ∇_{θ_μ} and ∇_{θ_σ} respectively denote the gradients under parameters θ_μ and θ_σ, respectively. Show the following:

   i. $\nabla_{\theta_\mu} \ln \pi_\theta(a \mid s) = \dfrac{1}{\sigma(s,\theta)^2}(a - \mu(s,\theta))x_\mu(s)$,

   ii. $\nabla_{\theta_\sigma} \ln \pi_\theta(a \mid s) = \left(\dfrac{(a-\mu(s,\theta))^2}{\sigma(s,\theta)^2} - 1\right) x_\sigma(s)$.

2. Consider a recycling robot whose job is to search and pick empty cold drink cans in an office environment. The battery-operated robot can be in either H (high) or L (low) state depending on it's charge level. When the state is H, the actions allowed are S (search) or W (wait). The S action corresponds to the robot actively looking for cans while the W action corresponds to it standing at one point in the hope that people will come and give it cans. When the robot takes action S in the H state, it receives a reward of r_S and remains in the H state at the next instant w.p. α but goes to the L state w.p. $1-\alpha$. When taking the W action in the H state, it remains in the H state w.p. 1 and receives a reward of r_W.

   When the robot is in the L state, the feasible actions are S, W and Re (recharge). When it decides to take the S action, in the L state, it remains in the L state at the next instant w.p. β and receives a reward of r_S. However, while doing the S operation in the L state, it completely loses it's charge w.p. $1-\beta$. In this case, it's manager needs to carry the robot to a plug point where it gets recharged and returns to the H state at the next instant. This, however, comes at a negative reward of -5 (that can be interpreted as a penalty). Finally, when in the L state, the robot decides to take action Re, it gets to the H state at the next instant and the single-stage reward in this case is 0.

   The state transition diagram depicting the states, actions, rewards and state-transitions is given in the Figure 1.

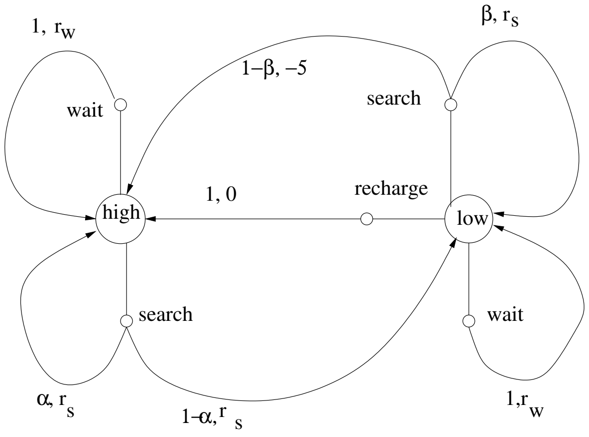

Figure 1: The Recycling Robot

Suppose for this example, when state = high (H), the robot chooses the search (S) action w.p. 0.6 and wait (W) w.p. 0.4. When state = low (L), the robot chooses the search (S) action w.p. 0.3, wait (W) action w.p. 0.3 and recharge (Re) action w.p. 0.4, respectively. We define an episode here as follows: Each episode starts in the high state and ends when either the system transits from the state low to high upon choosing the search action or else the episode is terminated after ten transitions whichever occurs earlier. Assume $\alpha = \beta = 0.5$, $r_S = 1$ and $r_W = 0.5$, respectively. Let the discount factor $\gamma = 1$ (i.e., this is an undiscounted task). Also, assume that all Q-value updates associated with 'terminal state-action tuples' are zero.

Consider the following episode that is observed with the above policy:

$$H,\ W,\ r_W,\ H,\ S,\ r_S,\ L,\ Re,\ 0,\ H,\ S,\ r_S,\ L.$$

Consider now the SARSA, expected SARSA and Q-learning algorithms but which are applied to the above fixed policy on the above episode. Let the step-size $\alpha = 0.1$ for all three algorithms. Assume that the initial value of $Q(S, A) = 0.5$ for all (S, A) tuples in all algorithms.

Note that in the original algorithms, as described in the Text Book, the policy used to sample the actions is updated based on the Q-values. However, here, our policy is fixed and under that policy, we observe the above episode. Thus, in the Q-value updates for the three algorithms, the action A_t chosen in state S_t to decide the (S_t, A_t) tuple whose Q-value is to be updated, is given by the above episode. The action A_{t+1} in S_{t+1} used by the SARSA update is again decided by the above episode. However, expected SARSA and Q-learning will still use $\sum_b \pi(b \mid S_{t+1})Q(S_{t+1}, b)$ and $\max_b Q(S_{t+1}, b)$, respectively, in their updates.

Answer now the following questions.

(a) Identify all the (S, A) tuples visited during the episode?

(b) Find the final values of $Q(S, A)$ for all (S, A) tuples at the end of the above episode when SARSA is used?

(c) Find the final values of $Q(S, A)$ for all (S, A) tuples at the end of the above episode when expected SARSA is used?

(d) Find the final values of $Q(S, A)$ for all (S, A) tuples at the end of the above episode when q-learning is used?

3. Consider an infinite horizon MDP with two states 1 and 2. A transition to state 1 (from any state) leads to a reward of 1, whereas a transition to state 2 gives a reward of 0. Let the discount factor be γ. Suppose we have obtained following two infinite length trajectories by following a known policy π:

$$1, 2, 1, 2, 1, 2, ...$$
$$1, 1, 1, 1, 1, 1, ...$$

   (a) Write down the first visit MC estimate of the value function of 1?

   (b) Does this estimate change, if every visit MC estimate is used? Prove or Disprove.

4. By a Markov reward process (MRP) we mean a Markov chain with a reward structure. Thus on each transition from one state to another, a single-stage reward is obtained. One may alternatively view an MRP as MDP wherein each state only one action is feasible and which one may assume the agent will anyway take. Let's call this policy as π. Consider now an MRP as shown in Fig.2. There are five non-terminating states labelled A,B,...,E. States F and G are assumed to be terminal states. All transitions give a single-stage reward of 0 except the transition to state F from D that gives a reward of 1. The weights shown on the various links connecting the nodes in the figure are the probabilities of transition. Thus,

$$P(A, B) = P(A, C) = P(B, D) = P(B, E) = P(C, D) = P(C, E) = P(D, F) = P(D, B)$$
$$= P(E, G) = P(E, C) = 0.5.$$

Let $\gamma = 1$, i.e., no discounting.

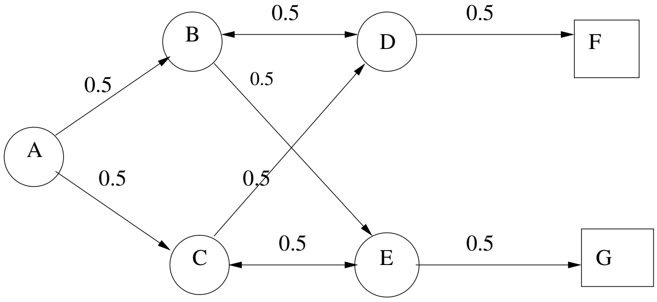

Figure 2: The MRP Model

   (a) For the problem above, compute the values $v_\pi(A), v_\pi(B), v_\pi(C), v_\pi(D), v_\pi(E)$?

   (b) Suppose now that we apply the TD(0) algorithm (for the look-up table case) with state A as the starting state. Assume the initial values $V_0(A) = \cdots = V_0(E) = 0.5$, $V_0(F) = V_0(G) = 0$ and that a constant step-size of 0.1 is used. Suppose also that two episodes are obtained and both episodes correspond to

$$A \longrightarrow C \longrightarrow E \longrightarrow G.$$

   Compute $V_1(\cdot)$ and $V_2(\cdot)$ for all states $A, \ldots, G$.