Assignment 3

Reinforcement Learning

Prof. B. Ravindran

1. In solving a multi-arm bandit problem using the policy gradient method, are we assured of converging to the optimal solution?

   (a) no

   (b) yes

2. In many supervised machine learning algorithms, such as neural networks, we rely on the gradient descent technique. However, in the policy gradient approach to bandit problems, we made use of gradient ascent. This discrepancy can mainly be attributed to the differences in

   (a) the objectives of the learning tasks

   (b) the parameters of the functions whose gradient are being calculated

   (c) the nature of the feedback received by the algorithms

3. Consider a bandit problem in which the parameters on which the policy depends are the preferences of the actions and the action selection probabilities are determined by the softmax relationship as $\pi(a_i) = \frac{e^{\theta_i}}{\sum_{j=1}^{k} e^{\theta_j}}$, where k is the total number of actions and θ_i is the preference value of action a_i. Derive the parameter update conditions according to the REINFORCE procedure considering the above described parameters and where the baseline is the reference reward defined as the average of the rewards received for all arms.

   (a) $\Delta\theta_i = \alpha(R - b)(1 - \pi(a_i; \theta))$

   (b) $\Delta\theta_i = \alpha(R - b)\frac{e^{\theta_i}}{\sum_{j=1}^{k} e^{\theta_j}}$

   (c) $\Delta\theta_i = \alpha(R - b)(\pi(a_i; \theta) - 1)$

   (d) $\Delta\theta_i = \alpha(R - b)\frac{1 - e^{\theta_i}}{\sum_{j=1}^{k} e^{\theta_j}}$

4. Repeat the above problem for the case where the parameters are the mean and variance of the normal distribution according to which the actions are selected and the baseline is zero.

   (a) $\Delta\mu_n = \alpha_n r_n \left(\frac{a_n - \mu_n}{\sigma_n}\right); \Delta\sigma_n = \alpha_n r_n \left\{\frac{(a_n - \mu_n)^2}{\sigma_n^2} - 1\right\}$

   (b) $\Delta\mu_n = \alpha_n r_n \left(\frac{a_n - \mu_n}{\sigma_n^2}\right); \Delta\sigma_n = \alpha_n r_n \left\{\frac{(a_n - \mu_n)^2 - \sigma_n^2}{\sigma_n^2}\right\}$

   (c) $\Delta\mu_n = \alpha_n r_n \left(\frac{a_n - \mu_n}{\sigma_n^2}\right); \Delta\sigma_n = \alpha_n r_n \left\{\frac{(a_n - \mu_n)^2 - \sigma_n^2}{\sigma_n^3}\right\}$

(d) $\Delta\mu_n = \alpha_n r_n \left(\frac{a_n - \mu_n}{\sigma_n}\right); \Delta\sigma_n = \alpha_n r_n \left\{\frac{(a_n - \mu_n)^2 - \sigma}{\sigma_n^2}\right\}$

5. Which among the following is/are differences between contextual bandits and full RL problems?

   (a) the actions and states in contextual bandits share features, but not in full RL problems
   (b) the actions in contextual bandits do not determine the next state, but typically do in full RL problems
   (c) full RL problems can be modelled as MDPs whereas contextual bandit problems cannot
   (d) no difference

6. Given a stationary policy, is it possible that if the agent is in the same state at two different time steps, it can choose two different actions?

   (a) no
   (b) yes

7. In class we saw that it is possible to learn via a sequence of stationary policies, i.e., during an episode, the policy does not change, but we move to a different stationary policy before the next episode begins. Does the temporal difference method encountered when discussing the tic-tac-toe example follow this pattern of learning?

   (a) no
   (b) yes

8. We saw the following definition of the action-value function for policy π: $q_\pi(s, a) = E_\pi[G_t | S_t = s, A_t = a]$. Suppose that the action selected according to the policy π in state s is a_1. For the same state, will the function be defined for actions other than a_1?

   (a) no
   (b) yes

9. Consider a 100x100 grid world domain where the agent starts each episode in the bottom-left corner, and the goal is to reach the top-right corner in the least number of steps. To learn an optimal policy to solve this problem you decide on a reward formulation in which the agent receives a reward of +1 on reaching the goal state and 0 for all other transitions. Suppose you try two variants of this reward formulation, (P_1), where you use discounted returns with $\gamma \in (0, 1)$, and (P_2), where no discounting is used. Which among the following would you expect to observe?

   (a) the same policy is learned in (P_1) and (P_2)
   (b) no learning in (P_1)
   (c) no learning in (P_2)
   (d) policy learned in (P_2) is better than the policy learned in (P_1)

10. Given an MDP with finite state set, S, and an arbitrary function, f, that maps S to the real numbers, the MDP has a policy π such that $f = V^\pi$.

    (a) false
    (b) true