Assignment 3
Reinforcement Learning
Prof. B. Ravindran

1. Which of the following is true for an MDP?

   (a) $Pr(s_{t+1}, r_{t+1}|s_t, a_t) = Pr(s_{t+1}, r_{t+1})$

   (b) $Pr(s_{t+1}, r_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, s_{t-2}, a_{t-2}, ..., s_0, a_0) = Pr(s_{t+1}, r_{t+1}|s_t, a_t)$

   (c) $Pr(s_{t+1}, r_{t+1}|s_t, a_t) = Pr(s_{t+1}, r_{t+1}|s_0, a_0)$

   (d) $Pr(s_{t+1}, r_{t+1}|s_t, a_t) = Pr(s_t, r_t|s_{t-1}, a_{t-1})$

   Sol. (b)
   (b) is true for any MDP. (a),(c) and (d) are not true.

2. The baseline in the REINFORCE update **should not** depend on which of the following (without voiding any of the steps in the proof of REINFORCE)?

   (a) r_{n-1}

   (b) r_n

   (c) Action taken(a_n)

   (d) None of the above

   Sol. (c)
   The baseline must not depend on any action. The baseline can depend on current and past rewards. An example baseline is given in the videos where the average of rewards obtained so far is considered to be the baseline.

3. 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

   Sol. (a), (c)
   The feedback in most supervised learning algorithms is an error signal which we wish to minimise. Hence, we would look to perform gradient descent. In policy gradient, we are trying to maximise the reward signal, hence gradient ascent.

4. In case of linear bandits, let's consider we have 2 actions - a_1 and a_2. The policy π to be followed when encountering a state s is given by

$$\pi(a_1|s) = \frac{e^{w_1^\top s}}{e^{w_1^\top s} + e^{w_2^\top s}}$$

$$\pi(a_2|s) = \frac{e^{w_2^\top s}}{e^{w_1^\top s} + e^{w_2^\top s}}$$

where w_1 and w_2 are weight vectors associated with a_1 and a_2 respectively. If we are using REINFORCE to learn the parameters, what will be the update for w_1 and w_2, when we pull action a_1?

(i) $w_1 \rightarrow w_1 + \alpha r(s - \frac{e^{w_1^\top s}}{e^{w_1^\top s}+e^{w_2^\top s}} s)$

(ii) $w_1 \rightarrow w_1 + \alpha r(-\frac{e^{w_1^\top s}}{e^{w_1^\top s}+e^{w_2^\top s}} s)$

(iii) $w_2 \rightarrow w_2 + \alpha r(s - \frac{e^{w_2^\top s}}{e^{w_1^\top s}+e^{w_2^\top s}} s)$

(iv) $w_2 \rightarrow w_2 + \alpha r(-\frac{e^{w_2^\top s}}{e^{w_1^\top s}+e^{w_2^\top s}} s)$

Which of the above updates are correct?

(a) (i), (iii)
(b) (i), (iv)
(c) (ii), (iii)
(d) (ii), (iv)

Sol. (b)
Derive the update using the REINFORCE formula.

5. The update in REINFORCE is given by $\theta_{t+1} = \theta_t + \alpha r_t \frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}$, where $r_t \frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}$ is an unbiased estimator of the true gradient of the performance function. However, there was another variant of REINFORCE, where a baseline b, that is independent of the action taken, is subtracted from the obtained reward, i.e, the update is given by $\theta_{t+1} = \theta_t + \alpha(r_t - b)\frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}$. How are $\mathbb{E}[(r_t - b)\frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}]$ and $\mathbb{E}[r_t \frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}]$ related?

(a) $\mathbb{E}[(r_t - b)\frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}] = \mathbb{E}[r_t \frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}]$
(b) $\mathbb{E}[(r_t - b)\frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}] < \mathbb{E}[r_t \frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}]$
(c) $\mathbb{E}[(r_t - b)\frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}] > \mathbb{E}[r_t \frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}]$
(d) Could be either of a, b or c, depending on the choice of baseline

Sol. (a)

$$\begin{aligned}
\mathbb{E}[b\frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}] &= \mathbb{E}[b\frac{1}{\pi(a_t;\theta_t)}\frac{\partial \pi(a_t;\theta_t)}{\partial \theta_t}] \\
&= \sum_a [b\frac{1}{\pi(a;\theta_t)}\frac{\partial \pi(a;\theta_t)}{\partial \theta_t}]\pi(a;\theta_t) \\
&= \sum_a b\frac{\partial \pi(a;\theta_t)}{\partial \theta_t} \\
&= b\frac{\partial 1}{\partial \theta_t} \\
&= 0
\end{aligned}$$

Thus, $\mathbb{E}[(r_t - b)\frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}] = \mathbb{E}[r_t \frac{\partial \ln \pi(a_t;\theta_t)}{\partial \theta_t}]$

6. Consider the following policy-search algorithm for a multi-armed binary bandit:

$$\forall a, \quad \pi_{t+1}(a) = \pi_t(a)(1-\alpha) + \alpha(\mathbb{1}_{a=a_t} r_t + (1 - \mathbb{1}_{a=a_t})(1 - r_t))$$

where $\mathbb{1}_{a_t=a}$ is 1 if $a = a_t$ and 0 otherwise. Which of the following is true for the above algorithm?

(a) It is L_{R-I} algorithm.

(b) It is $L_{R-\epsilon P}$ algorithm.

(c) It would work well if the best arm had probability of 0.9 of resulting in +1 reward and the next best arm had probability of 0.5 of resulting in +1 reward

(d) It would work well if the best arm had probability of 0.3 of resulting in +1 reward and the worst arm had probability of 0.25 of resulting in +1 reward

Sol. (c)
The given algorithm is $L_{R=P}$ algorithm. It would work well for the case described in (c) as it gives equal weightage to penalties and rewards, and as the gap between best arm and next best arm's probability of giving +1 reward is significant, it would easily figure out the best arm.

7. The actions in contextual bandits do not determine the next state, but typically do in full RL problems. True or false?

(a) True

(b) False

Sol. (a)
Refer to the contextual bandits lecture.

8. In a continuous action space environment, we can employ any value function-based algorithm to discover an optimal policy.

(a) True

(b) False

Sol. (b)
When the action space is continuous, storing values for each action is hard in the case of value function-based methods. Due to the action approximation for the continuous actions, it is necessary to optimize the action space for every action chosen from the policy, to find the estimated optimal action. This approach is slow, and policy gradient-based methods are likely to generalize better.

9. Let's assume for some full RL problem we are acting according to a policy π. At some time t, we are in a state s where we took action a_1. After few time steps, at time t', the same state s was reached where we performed an action $a_2 (\neq a_1)$. Which of the following statements is true?

(a) π is definitely a stationary policy

(b) π is definitely a non-stationary policy

(c) π can be stationary or non-stationary.

Sol. (c)
A stationary policy can be stochastic and thus the for same state different actions can be chosen at different time steps. Thus π can be stationary or non-stationary policy.

10. 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

    Sol. (a)
    Depending upon the properties of the function whose gradient is being ascended, the policy gradient approach may converge to a local optimum.