# Assignment 1

## Reinforcement Learning
## Prof. B. Ravindran

1. Which of the following is not a useful way to approach a standard multi-armed bandit problem with $n$ arms? Assume bandits are stationary.

   (a) "How can I ensure the best action is the one which is mostly selected as time tends to infinity?"

   (b) "How can I ensure the total regret as time tends to infinity is minimal?"

   (c) "How can I ensure an arm which has an expected reward within a certain threshold of the optimal arm is chosen with a probability above a certain threshold?"

   (d) "How can I ensure that when given any 2 arms, I can select the arm with a higher expected return with a probability above a certain threshold?"

   **Sol.** (d)
   Options a,b and c refer to asymptotic correctness, regret optimality and PAC optimality respectively. Option d, i.e, being able to choose the better arm given 2 arms is not a useful way to look at the standard multi-armed bandit problem, since it is not necessary to ensure any 2 arms can be compared with a high degree of success. For example, if two arms had very similar and very low expected returns as compared to the optimal arm, it would not be useful to pick those arms again for the purpose of gaining more accurate estimates of the expected reward obtained for picking that arm (which would help in finding the better arm among the two).

2. What is the decay rate of the weightage given to past rewards in the computation of the $Q$ function in the stationary and non-stationary updates in the multi-armed bandit problem?

   (a) hyperbolic, linear

   (b) linear, hyperbolic

   (c) hyperbolic, exponential

   (d) exponential, linear

   **Sol.** (c)
   In the stationary case, the weightage of all rewards given so far, is $\frac{1}{n}$, where $n$ is the number of rewards obtained for that arm/action so far. Therefore it is hyperbolic. In the non-stationary update, the weightage given to the most recent reward is $\alpha$, for the pre-update $Q$ value it is $1 - \alpha$. Upon expansion of of the pre-update $Q$ value, we can see that the weightage of the reward in the last but one time step is $(\alpha)(1 - \alpha)$. We can continue to expand in a similar fashion to get that a reward obtained $t$ time steps ago has a weightage of $\alpha(1 - \alpha)^t$.

3. In the update rule $Q_{t+1}(a) \leftarrow Q_t(a)+\alpha(R_t - Q_t(a))$, select the value of $\alpha$ that we would prefer to estimate Q values in a non-stationary bandit problem.

   (a) $\alpha = \frac{1}{n_a+1}$

   (b) $\alpha = 0.1$

   (c) $\alpha = n_a + 1$

(d) $\alpha = \frac{1}{(n_a+1)^2}$

**Sol.** (b)
By using a constant value of $\alpha$, we decrease the importance of past samples exponentially as time progresses, so that our Q value estimates are able to shift as the true action values change.

Option (a) weights each sample equally (computing a simple average), and after a large number of time steps, the importance of new samples in the average will be very low - preventing meaningful update of Q values when the distribution of action-values change.

Option (c) increases the importance of new samples as training progresses and creates an unbounded sum. Action values estimates will not converge, but grow to infinity.

Option (d) weights newer samples lower, and the importance of newer samples for action-value estimates will fall quicker than in case (a).

4. Assertion: Taking exploratory actions is important for RL agents
Reason: If the rewards obtained for actions are stochastic, an action which gave a high reward once, might give lower reward next time.

(a) Assertion and Reason are both true and Reason is a correct explanation of Assertion

(b) Assertion and Reason are both true and Reason is not a correct explanation of Assertion

(c) Assertion is true and Reason is false

(d) Both Assertion and Reason are false

**Sol.** (b)
An RL agent needs to take exploratory actions because it needs to estimate the advantage for each action correctly. So, the Assertion is true. Reason is also true because if the rewards are stochastic, the agent might get high reward once and low reward next time for the same action in the same state. However, even if the rewards obtained for actions were deterministic, the agent still needs to take exploratory actions to figure out more advantageous actions. Hence, Reason does not explain the Assertion correctly.

5. We are trying different algorithms to find the optimal arm for a multi arm bandit. We plot expected payoff vs time graph for each algorithm for which the expected payoff satisfy some function with respect to time (staring from 0). Which among the following functions will have the least regret. (We know that the optimal expected pay off is 1) (Hint: Plot the functions)

(a) $tanh(t)$

(b) $1 - 2^{-t}$

(c) $x/20$ if $x < 20$ and 1 after that

(d) Same regret for all the above functions.

**Sol.** (a)
If we plot the functions we can clearly see the area between $y = 1, f(t)$. The exact regret will be equal to $\int_0^\infty 1 - f(t)dt$.

6. Consider the following statements for $\epsilon$-greedy approach in a non-stationary environment:

i Keeping a small constant $\epsilon$ is a good approach if the environment is non-stationary..

ii Large values of $\epsilon$ will lead to unnecessary exploration in the long run.

iii For a stationary environment, decaying $\epsilon$ value to zero is a good approach, as after reaching optimality, we would like to reduce exploration.

Which of the above statements is/are correct?

(a) ii, iii

(b) only iii

(c) only ii

(d) i, ii

**Sol.** (d)
In non-stationary environment since optimal state changes, hence it requires constant exploration to find the optimal state.
Decaying epsilon to zero may not find an optimal state; hence, iii is incorrect.

7. Following are two ways for defining the probability of selecting an action/arm. Select the option regarding better choice among the following

i $Pr(a_t = a) = \frac{Q_t(a)}{\sum_a Q_t(a)}$.

ii $Pr(a_t = a) = \frac{e^{Q_t(a)}}{\sum_{b=1}^{n} e^{Q_t(b)}}$.

(a) Both are good as both formulas can bound probability in range 0 to 1.

(b) (i) is better because it is differentiable and requires less complex computation.

(c) None of the above

**Sol.** (c)
(a), (b) are incorrect as (i) cannot handle negative values in some cases.

8. Which of the following best refers to $PAC$-optimality solution to bandit problems?
$\epsilon$ – is the difference between the reward of the chosen arm and true optimal reward
$\delta$ – is the probability that chosen arm is not optimal
$N$ – is the number of steps to reach PAC-optimality

(a) Given $\delta$ and $\epsilon$, minimize the number of steps to reach PAC-optimality(i.e. N)

(b) Given $\delta$ and $N$, minimize $\epsilon$.

(c) Given $\epsilon$ and $N$, maximize the probability of choosing optimal arm(i.e. minimize $\delta$)

(d) none of the above is true about $PAC$-optimality

**Sol.** (a)
refer to the definition of $PAC$-optimality in Bandit Optimalities video.

9. Suppose we have a 10-armed bandit problem where the rewards for each of the 10 arms is deterministic and in the range (0, 10). Which among the following methods will allow us to accumulate maximum reward in the long term?

(a) $\epsilon$-greedy with $\epsilon = 0.1$.

(b) $\epsilon$-greedy with $\epsilon = 0.01$.

(c) greedy with initial reward estimates set to 0.

(d) greedy with initial reward estimates set to 10.

**Sol.** (d)
Since the rewards are deterministic, we only need to select each arm once to identify an optimal arm. The greedy method with initial reward higher than all possible rewards ensures that each arm is selected at least once, since on selecting any arm, the resultant reward estimate will necessarily be lower than the initial estimates of the other arms. Once each arm has been selected, the greedy method will settle on the arm with the maximum reward.

10. Which of the following is/are correct and valid reasons to consider sampling actions from a softmax distribution instead of using an $\epsilon$-greedy approach?

i Softmax exploration makes the probability of picking an action proportional to the action-value estimates. By doing so, it avoids wasting time exploring obviously 'bad' actions.

ii We do not need to worry about decaying exploration slowly like we do in the $\epsilon$-greedy case. Softmax exploration gives us asymptotic correctness even for a sharp decrease in temperature.

iii It helps us differentiate between actions with action-value estimates (Q values) that are very close to the action with maximum Q value.

Which of the above statements is/are correct?

(a) i, ii, iii

(b) only iii

(c) only i

(d) i, ii

(e) i, iii

**Sol.** (e)
ii is incorrect. If we decrease temperature to quickly, we could fail to do enough exploration to make correct action-value estimates, just as we would by decaying $\epsilon$ too quickly in $\epsilon$-greedy exploration.
i and iii are correct. Softmax encourages exploration of actions that have action-value estimates close to the action with maximum Q value. Further concentrated exploration of these actions improve our action-value estimates for them, and this allows us to differentiate between them better.