Model questions for RL Quiz 1

1. Consider a k-armed bandit problem with $k = 4$ actions,denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = 1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = 2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

2. Suppose ν is a finite-armed stochastic bandit and π be a policy. The regret of policy π in bandit ν is defined by:

$$R_n(\pi, \nu) = n\mu^*(\nu) - \mathbb{E}\Big[\sum_{t=1}^{n} X_t\Big]$$

It also satisfies the following:

$$\lim_{n\to\infty} \frac{R_n(\pi, \nu)}{n} = 0$$

Let $T^*(n) = \sum_{t=1}^{n} \mathbb{I}_{\{\mu_{A_t}=\mu^*\}}$ be the number of times optimal arm is chosen. Prove or disprove each of the following statements:

   (a) $\lim_{n\to\infty} \mathbb{E}[T^*(n)]/n = 1$

   (b) $\lim_{n\to\infty} P(\mu^* - \mu_{A_t} > 0) = 0$

3. An unscrupulous innkeeper charges a different rate for a room as the day progresses, depending on whether he has many or few vacancies.His objective is tomaximize his expected total income during the day.Let x be the number of empty rooms at the start of the day, and let y be the number of customers that will ask for a room in the course of the day. We assume(somewhat unrealistically) that the innkeeper knows y with certainty, and upon arrival of a customer, quotes one of m prices r_i, $i = 1, ..., m$, where $0 < r_1 \leq r_2 \leq ... \leq r_m$. A quote of a rate T_i is accepted with probability p_i and is rejected with probability $1 - p_i$,in which case the customer departs, never to return during that day.

(a) Formulate this as a DP problem with y stages to find the maximum expected income. Assuming that the product $p_i r_i$ is monotonically nondecreasing with i, and that p_i is monotonically nonincreasing with i, show that the innkeeper should always charge the highest rate r_m

(b) Consider a variant of the problem where each arriving customer, with probability p_i, offers a price r_i for a room, which the innkeeper may accept or reject, in which case the customer departs, never to return during that day. Formulate this as a DP problem to find the maximum expected income. Show also it is optimal to accept a customer's offer if it is larger than some threshold $\bar{r}$ depending on the state and stage.

4. A farmer annually producing X_k units of a certain crop stores $(1-U_k)X_k$ units of his production, where $0 \leq U_k \leq 1$, and invests the remaining $U_k X_k$ units, thus increasing the next year's production to a level X_{k+l} given by

$$X_{k+1} = X_k + W_k U_k X_k, k = 0, 1, \ldots, N-1$$

The scalars W_k are independent random variables with identical probability distributions that do not depend either on X_k or U_k. Furthermore, $\mathbb{E}[W_k] = \bar{W} > O$. The problem is to find the optimal investment policy that maximizes the total expected product stored over N years.

$$\mathbb{E}_{w_k, k=0\ldots, N-1}\big[X_N + \sum_{k=0}^{N-1}(1-U_k)X_k\big]$$

Show the optimality of the following policy are constant functions:

(a) If $\bar{W} > 1$, $\mu_0^*(X_0) = \ldots = \mu_{N-1}^*(X_{N-1}) = 1$

(b) If $0 < \bar{W} < \frac{1}{N}$, $\mu_0^*(X_0) = \ldots = \mu_{N-1}^*(X_{N-1}) = 0$

(c) If $\frac{1}{N} \leq \bar{W} \leq 1$,

$$\mu_0^*(X_0) = \ldots = \mu_{N-\bar{k}-1}^*(X_{N-\bar{k}-1}) = 1$$

$$\mu_{N-\bar{k}}^*(X_{N-\bar{k}}) = \ldots = \mu_{N-1}^*(X_{N-1}) = 0$$

where $\bar{k}$ is such that $\frac{1}{\bar{k}+1} < \bar{W} \leq \frac{1}{\bar{k}}$