RL Tutorial 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 we have a machine that is either running or is broken down. If it runs throughout one week, it makes a gross profit of $100. If it fails during the week, gross profit is zero. If it is running at the start of the week and we perform preventive maintenance, the probability that it will fail during the week is 0.4. If we do not perform such maintenance, the probability of failure is 0.7. However, maintenance will cost $20. When the machine is broken down at the start of the week, it may either be repaired at a cost of $40, in which case it will fail during the week with a probability of 0.4, or it may be replaced at a cost of $150 by a new machine that is guaranteed to run through its first week of operation. Find the optimal repair, replacement, and maintenance policy that maximizes total profit over four weeks, assuming a new machine at the start of the first week (that is guaranteed to run during the first week of operation).

3. 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+1} given by

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

The scalars W_k are independent random variables with identical probability distributions that do not depend either on X_k or U_k. Furthermore, $\mathbb{E}[W_k] = \bar{W} \geq 0$. The problem is to find the optimal investment policy that maximizes the total expected product stored over N years.

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

Show the following
(a) If $\bar{W} > 1$,

$$\mu_i^*(X_i) = 1, 0 \leq i \leq N-1$$

(b)If $0 < \bar{W} < \frac{1}{N}$,

$$\mu_i^*(X_i) = 0, 0 \leq i \leq N-1$$

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

$$\mu_i^*(X_i) = 1, 0 \leq i \leq N-\bar{k}-1$$
$$\mu_i^*(X_i) = 0, N-\bar{k} \leq i \leq N-1$$

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