Northeastern University
College of Engineering

EECE 5614 - Reinforcement Learning and Decision Making Under Uncertainty

Project #1 Spring 2025

Think of an agent that plays a 2-armed bandit, trying to maximize its total reward. In each step, the agent selects one of the levers and is given some reward according to the reward distribution of that lever. Assume that reward distribution for the first lever is a Gaussian with $\mu_1 = 6, \sigma_1^2 = 15$, and for the second lever is a binomial Gaussian with $\mu_{21} = 11, \sigma_{21}^2 = 16, \mu_{22} = 3, \sigma_{22}^2 = 8$, which means that the resulting output will be uniformly probable from these two Gaussian distributions (See http://en.wikipedia.org/wiki/Mixture_distribution).
If these distributions were known (which in practice are not), we could compute the optimal/true action values as:

$$Q^*(a^1) = E[R^1] = \mu_1 = 6$$
$$Q^*(a^2) = E[R^2] = \frac{1}{2} \times \mu_{21} + \frac{1}{2} \times \mu_{22} = 7$$

However, in this problem, we assume the reward distributions are unknown, and the agent only sees a realization of reward after selecting an action. The agent takes action according to the ϵ-greedy action selection policy with parameter ϵ:

$$\pi^{\epsilon-greedy} = \begin{cases} \arg max\, Q(a) & \text{w. p. } 1-\epsilon \\ Random & \text{w. p. } \epsilon \end{cases}$$

We consider the agent selects 1000 actions, which is referred to as step/time. In order to have smooth results, we repeat 1000 steps for 100 independent runs.

a) In this part, set the initial Q values at the beginning of each run as $Q(a^1) = Q(a^2) = 0$. Assuming action a is selected at time step k and the reward r_k is observed, the Q-value for the corresponding action will be updated according to: $Q(a) = Q(a) + \alpha(\text{ r} - Q(a)\text{ })$. For the learning rates, consider the following values: $\alpha = 1, \alpha = 0.9^k$, $\alpha = \frac{1}{1+\text{Ln}(1+k)}$ and $\alpha = \frac{1}{k}$ and for the ϵ-greedy policy, use $\epsilon = 0,\ 0.1,\ 0.2, 0.5$. You need to provide your results in terms of average accumulated reward with respect to time/step (see the following plot).
Here is a brief guideline: For the ith independent run, you need to keep track of accumulated rewards $AccR^i$ (a vector of size $\mathbb{R}^k$) as:

$$AccR^i(k) = \frac{1}{k}\sum_{j=1}^{k} r_j \quad , \qquad k = 1, \dots, 1000$$

where $AccR^i(k)$ denotes the average reward per step obtained by the agent up to the time step k in the ith independent run. Then the average over 100 independent runs of accumulated rewards $\overline{AccR}$ (a vector of size $\mathbb{R}^k$) can be obtained at any given step/time as:

$$\overline{AccR(k)} = \frac{1}{100}\sum_{i=1}^{100} AccR^i(k) \quad , \qquad k = 1, \dots, 1000$$

Therefore, in the example of the plot shown on the next page, $\overline{AccR(k)}$ is in the y-axis and k in the x-axis.
You expect to have four plots, which each one is associated with a learning rate and includes four curves for four different ϵ values. For all pairs of learning rate and the policy parameter (i.e., each curve), you also need to include the average action values $Q(a^1)$ and $Q(a^2)$ after finishing 1000 steps over 100 runs. An example of the table for this result is shown below.

Expected Results: Four Plots and Four Tables.

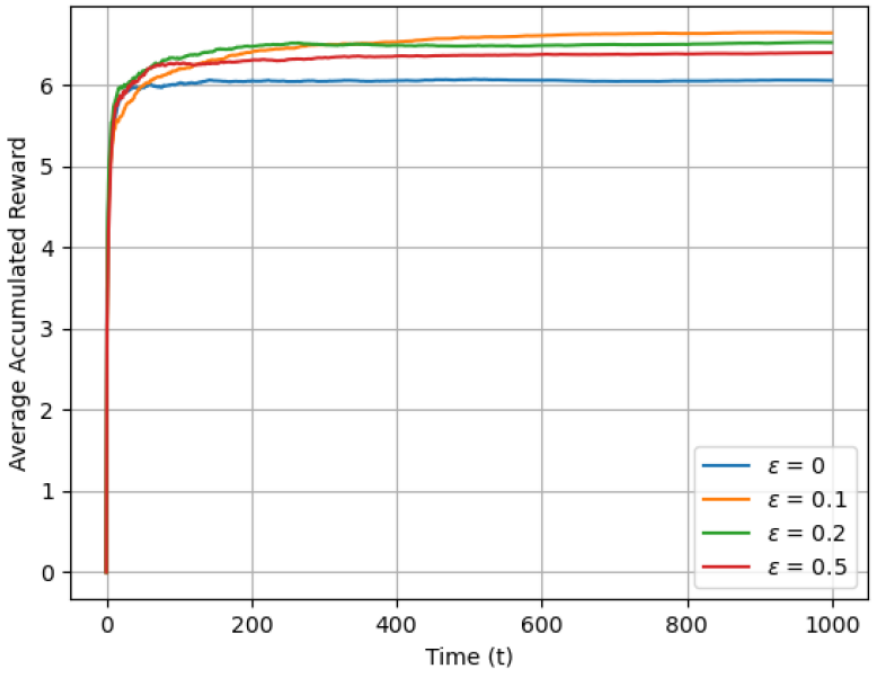

Figure 1 Average Accumulated Reward for $\alpha = \frac{1}{1+ln\,(1+k)}$

Table 1 Average Q-values for $\alpha = \frac{1}{1+ln\,(1+k)}$

Epsilon-greedy	Average of action value $Q(a^1)$ of 100 runs	**True action value $Q^*(a^1)$**	Average of action value $Q(a^2)$ of 100 runs	**True action value $Q^*(a^2)$**
ε = 0 (greedy)		**6**		**7**
ε = 0.1		**6**		**7**
ε = 0.2		**6**		**7**
ε = 0.5 (random)		**6**		**7**

b) For a fixed $\alpha = 0.1$ and $\epsilon = 0.1$, use the following optimistic initial values and compare the results: $Q = [0\ 0]$, $Q = [6\ 7]$, $Q = [15\ 15]$ (note that $Q = [Q(a^1)\ \ Q(a^2)]$.) Plot the average accumulated reward with respect to step/time in a single plot with four curves, where each curve is associated with a single initial Q-values. The average action values should be reported in the following table.

Expected Results: One Plot and One Table.

Initial Q values	Average of action value $Q(a^1)$ of 100 runs	**True action value $Q^*(a^1)$**	Average of action value $Q(a^2)$ of 100 runs	**True action value $Q^*(a^2)$**
Q =[0 0]		**6**		**7**
Q =[6 7]		**6**		**7**
Q =[15 15]		**6**		**7**

c) For a fixed $\alpha = 0.1$, use the Gradient-Bandit policy with $H_1(a^1) = H_1(a^2) = 0$. Plot the average accumulated reward with respect to step/time. How the results are different from ϵ-greedy results with $Q(a^1) = Q(a^2) = 0, \alpha = 0.1$ and $\epsilon = 0.1$? You might choose to plot both curves on top of each other for comparison purposes.

Expected Results: One Plot.

$$\pi_t(a^1) = \frac{e^{H_t(a^1)}}{e^{H_t(a^1)} + e^{H_t(a^2)}} \qquad \pi_t(a^2) = \frac{e^{H_t(a^2)}}{e^{H_t(a^1)} + e^{H_t(a^2)}}$$

$H_{t+1}(a_t) = H_t(a_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(a_t)),$ for selected action at time t

$H_{t+1}(a) = H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a),$ for the action that is not chosen at time t

$$\bar{R}_t = (r_1 + \cdots + r_t)/t$$

Important Note: For all results, you need to interpret/explain your findings. For instance, in part (a), you need to explain which of greedy, random, in-between policies performed the best, which learning rate was the best, which pair of α and ϵ led to the maximum average accumulated reward, etc. For part (b), you also need to explain how optimistic initial values impact the overall performance of the selection process and which choice is the best. For part (c), you need to compare the results of gradient-based policy with previously computed ϵ-greedy policy.

Questions about the project should be directed to TA, Hamid, at
hosseini.ha@northeastern.edu