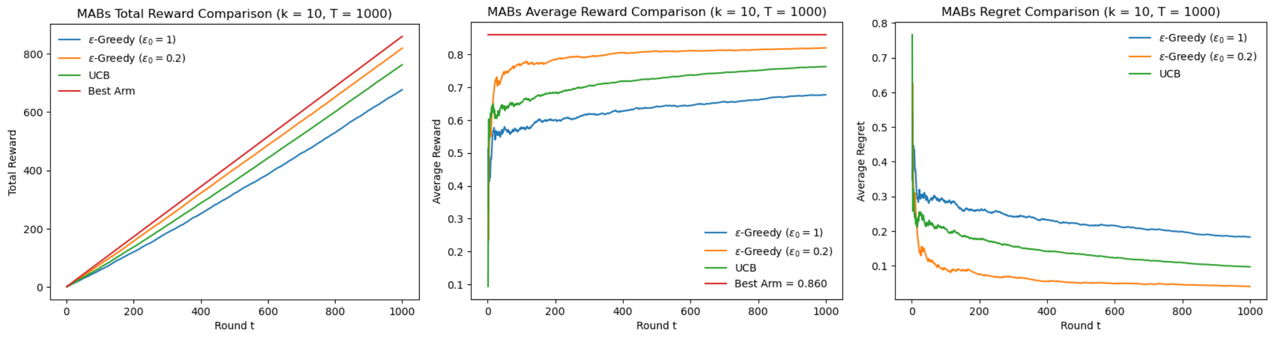
COMP423 - Reinforcement Learning and Dynamic Optimization 1st Programming Assignment Report

Pantourakis Michail AM 2015030185
School of Electrical and Computer Engineering
Technical University of Crete

Date submitted: 18 March 2023

In this report I present and analyze the results of realizations of two Multi-Armed Bandit (MAB) algorithms: 1) the ϵ-Greedy algorithm, and 2) the Upper Confidence Bound (UCB) algorithm. As shown in theory, both algorithms achieve sublinear regret rates. More specifically, ϵ-Greedy achieves sublinear regret rates when its ϵ parameter decreases as a function of time. In the accompanying Python script, the following function was used: $\epsilon(t) = \epsilon_0 t^{-1/3}(k \log t)^{1/3}$, where ϵ_0 is a constant.

All results assumed k stochastic bandits, with a reward distribution of $U(a_i, b_i)$ for each arm i. In each problem scenario, a_i and b_i were drawn randomly from distribution $U(0, 1)$, and they were kept the same for both algorithms. In total, three scenarios are shown here: 1) $k = 10$ arms played for $T = 1000$ rounds, 2) $k = 10$ for $T = 50000$, and 3) $k = 100$ for $T = 50000$.

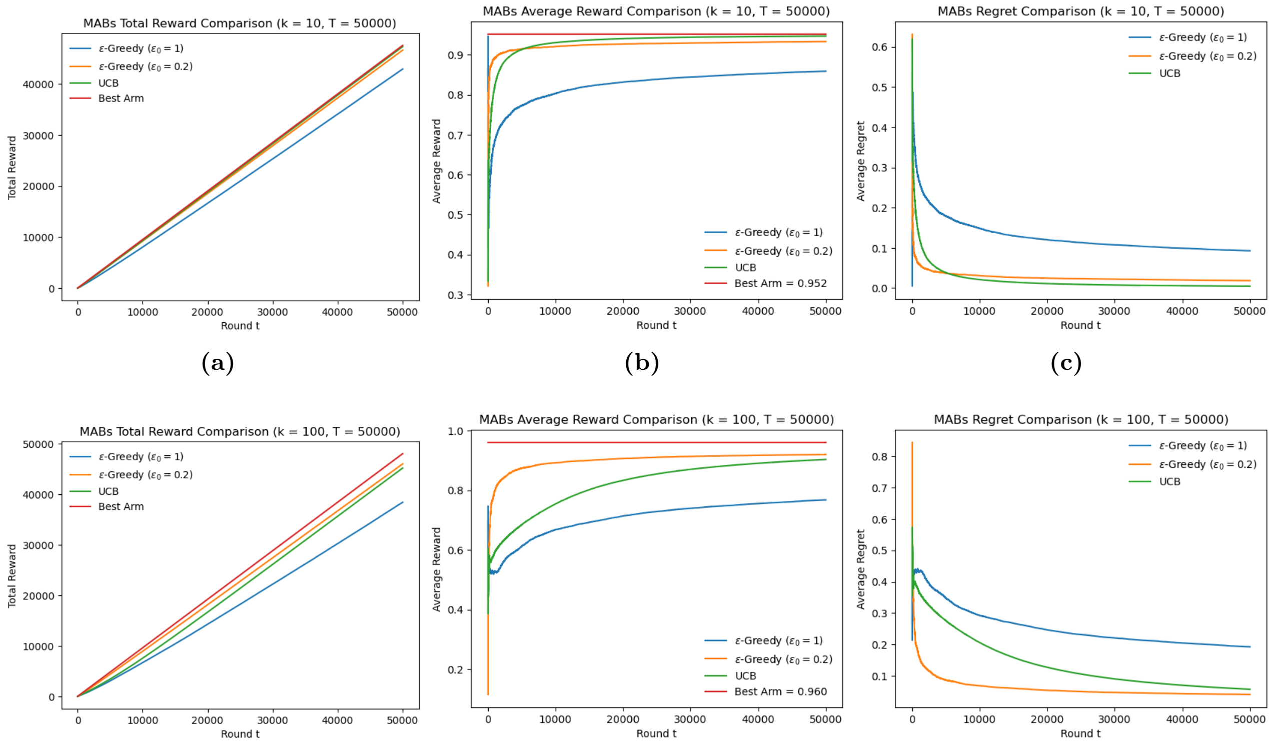


(a) **(b)** **(c)**

Figure 1: Performance of ϵ-Greedy and Upper Confidence Bound (UCB) algorithms for the Multi-Armed Bandit (MAB) problem, in one realization for $k = 10$ arms and $T = 1000$ rounds.

Figure 1 shows the results of just one realization per algorithm for the first scenario. Subfigure (a) shows the total reward accumulated at each round t, whereas (b) shows the average over t. In both subfigures, the reward provided by picking the best arm is also shown. Clearly, all algorithms select suboptimal arms at the beginning, akin to an "exploratory" phase. Nevertheless, as illustrated by Subfigure (3), all schemes achieve a sublinear regret rate, as indicated by approaching 0 as time passes.

Interestingly, the UCB algorithm seems to converge to the optimal reward faster than the ϵ-Greedy algorithm with $\epsilon_0 = 1$, and overall outperforms it. Viewing on the evolution of ϵ over time, this value of ϵ_0 allows a high probability of picking arms randomly even on later rounds, which explains why the algorithm converges much slower. To highlight the effect of ϵ_0 on convergence, I also investigated the algorithm's behavior for a smaller $\epsilon_0 = 0.2$. With such a decreased value, ϵ-Greedy starts picking good arms even faster (since the probability of picking random arms was 5 times smaller across all rounds). However, although not entirely clear in this time frame, its subsequent rate of convergence still seems worse than UCB. To better illustrate the asymptotic behavior of the algorithms, I generated another realization per algorithm for $T = 50000$ rounds. As seen in Figure 2(a-c), UCB clearly exhibits better learning rate than both ϵ-Greedy realizations in the long run.



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure 2: Performance of ϵ-Greedy and Upper Confidence Bound (UCB) algorithms for the Multi-Armed Bandit (MAB) problem, in one realization for $k = 10$ arms and $T = 50000$ rounds (subfigures a-c), and another for $k = 100$ bandits and $T = 50000$ rounds (subfigures d-f).

Finally, the effect of increasing k to 100 was investigated in scenario three (Figure 2(d-f)). Subfigures (e-f) clearly show that UCB now required more rounds to reach low regret levels compared to $k = 10$. This behavior is expected since the algorithm needs more rounds to explore all arms and reduce their confidence bounds before converging. In contrast, the effect of k to ϵ-Greedy for $\epsilon = 0.2$ is less severe on the initial ramp.

Due to its limited scope, this report only provides a glimpse of MAB algorithms in action. A more systematic investigation and statistical analysis out of multiple realizations would provide safer conclusions. In spite of this limitation, these results indicate that given enough rounds, the UCB algorithm shows consistently better regret rate than ϵ-Greedy. As expected from a “greedy” approach, it seems the best method when only a limited number of rounds T (in respect to k) is allowed, since this limit means that time is not enough for substantial exploration and good arms must be “exploited” fast. As experienced by subsequent tries, one could pick even smaller values of ϵ_0 to get lower regret as fast as possible. However, such a parameter setup presents the risk of converging fast to a suboptimal arm, which can be thought as the equivalent of committing to the best -yet suboptimal- estimated arm so far, after a short and inadequate exploration phase of an explore-then-exploit algorithm.