Final Project:
Comparative Analysis of Algorithms for Solving the Stochastic Multi-Armed Bandit

Thi Thu Ha Vo

May 10, 2025

1 Introduction

The multi-armed bandit (MAB) is an interesting problem in reinforcement learning, where an algorithm must balance the dual tasks of acquiring new knowledge through exploration and optimizing decisions based on existing knowledge through exploitation. The goal is to maximize the cumulative reward over time by effectively managing this exploration-exploitation trade-off. MAB is a practical problem, as many real-world scenarios can be modeled as instances of it. It has been shown that MAB has a wide range of applications, including clinical trials, recommendation systems, information retrieval, and dynamic pricing.

There are various versions of the multi-armed bandit problem, but in this project, we focus on the simplest formulation, which we refer to as the stochastic multi-armed bandit problem. In this setting, the problem has K possible actions, known as "arms." At each step, the algorithm selects an arm and receives a reward based on that arm. The reward is drawn independently from a fixed distribution associated with the chosen arm, but the algorithm has no prior knowledge of these reward distributions. The objective is to identify the arm with the highest expected reward and maximize the total reward over time. In this project, we are interested in comparing three well-known algorithms: ϵ-greedy, Softmax, and upper confidence bound (UCB).

2 Background

The problem was first introduced in [9]. In the basic setting, the problem can be described as follows,

- A bandit problem consists of K arms. Each arm a has a reward drawn from an unknown distribution D_a with mean $\mu(a)$. For simplicity, we suppose the support of the reward distribution D_a is $[0, 1]$ (or a subset of $[0, 1]$). However, this choice of reward range is not important, as any unbounded real interval can be appropriately scaled.

- At each round t, an arm a is selected, and a reward is independently drawn from the distribution D_a. We denote a_t as the arm chosen at round t, and $\mu^* = \mu(a^*)$ as the expected reward of the best arm a^*. Moreover, let $\hat{\mu}_t(a)$ denote the average reward for arm a at round t, and let $n_t(a)$ be the number of times arm a has been chosen up to round t.

The purpose of the MAB is to maximize the total reward over T rounds. However, many research papers prefer to use the total expected regret after T rounds as the performance measure. Here, we aim to minimize the total expected regret. The expected regret can be understood as the difference between the total reward one could obtain by always choosing the arm with the maximum expected reward over T rounds, and the expected total reward obtained from the actual sequence of T rounds.

$$R(T) = \mu^* \cdot T - \sum_{t=1}^{T} \mu(a_t) \tag{1}$$

The main idea is that if we choose a suitable setting for the algorithm, we can bound the expected regret, ensuring that it does not grow linearly - which would be the worst-case scenario.

3 Algorithms

3.1 ϵ-greedy algorithm

The ϵ-greedy strategy is widely used due to its simplicity. It was first proposed by Watkins [13], and later discussed in [11].

Algorithm 1 Epsilon-Greedy Algorithm

1: **for** round $t = 1, 2, ...T$ **do**
2: Explore: choose an arm uniformly at random with the probability ϵ
3: Exploit: choose the arm with the higherst emperical mean so far with the probability $1 - \epsilon$
4: **end for**

Notice that in this simplest form, the value of ϵ is constant, which prevents the algorithm from getting arbitrarily close to the optimal result. This is because, regardless of the value of ϵ, when the number of rounds T becomes sufficiently large, the exploration phase will continue to occur with probability ϵ, contributing asymptotically at least $\epsilon T\Delta$ to the expected regret, where Δ is the minimum difference between the true expected reward of the non-optimal arms and that of the optimal arm. Therefore, a constant exploration probability ϵ leads to linear growth in regret.

Since the linear bound on the expected regret with a constant value of ϵ is suboptimal, one can consider its adaptive variant, where the value of ϵ decreases over time. It has been proven that using a decreasing ϵ can lead to regret that is arbitrarily close to the asymptotic optimal level. Cesa-Bianchi and Fischer (1998) [3] proved poly-logarithmic bounds for variants of the algorithm in which ϵ decreases over time by choosing an appropriate function for ϵ. However, for simplicity, in this project, I adopt the bound provided in [10].

The algorithm for the decayed version of ϵ-greedy is exactly the same as in the constant case. By choosing an appropriate schedule for ϵ, the expected regret is guaranteed to attain a sublinear bound.

Theorem 1. *Assume the rewards of each arm are in* $[0, 1]$. *Then the* ϵ*-greedy algorithm with exploration probabilities* $\epsilon_t = t^{\frac{-1}{3}}(K \log t)^{\frac{1}{3}}$ *achieves regret bound* $\mathrm{E}[R(t)] \leq O(t^{\frac{2}{3}}(K \log t)^{\frac{1}{3}})$ *for each round* t.

Proof. We analyze the regret bound of this algorithm by fixing a round t. After round t, for each arm a, by applying Hoeffding's inequality, we obtain,

$$P(|\hat{\mu}_t(a) - \mu(a)| \geq \delta) \leq 2\exp\left(\frac{-2\delta^2 n_t(a)}{(1-0)^2}\right) = 2\exp\left(-2\delta^2 n_t(a)\right) \tag{2}$$

The denominator $(1-0)^2$ arises from the fact that the reward distributions have support in $[0, 1]$. Here, $n_t(a)$ denotes the number of times arm a has been chosen up to round t. Note that among the first t rounds, exploration occurs only in $t\epsilon_t$ rounds. Hence, each arm is explored on average $\frac{t\epsilon_t}{K}$ times. Naturally, some arms will be selected more frequently during the exploitation phase. Let us denote such

an arm as a; then $n_t(a) > \frac{t\epsilon_t}{K}$, which leads to even tighter probability bounds. Therefore, if we are interested in the asymptotic bound, we can proceed further using 2,

$$P(|\hat{\mu}_t(a) - \mu(a)| \geq \delta) \leq 2\exp\left(\frac{-2\delta^2 t\epsilon_t}{K}\right)$$

i.e,

$$P(|\hat{\mu}_t(a) - \mu(a)| \leq \delta) \geq 1 - 2\exp\left(\frac{-2\delta^2 t\epsilon_t}{K}\right)$$

Ideally, we want the average reward after the exploration phase to be a good estimate of the true expected reward. Therefore, after t rounds, we want $|\hat{\mu}_t(a) - \mu(a)|$ to be small. If we define the confidence radius as $\delta := \sqrt{\frac{2K\log t}{t\epsilon_t}}$, then by applying Hoeffding's inequality, we can ensure that $|\hat{\mu}_t(a) - \mu(a)|$ is smaller than this radius with probability at least,

$$1 - 2\exp\left(-2\left(\frac{2K\log t}{t\epsilon_t}\right)\frac{t\epsilon_t}{K}\right) = 1 - 2\exp\left(-4\log t\right) = 1 - \frac{2}{t^4}$$

Suppose a^* is the arm with the highest expected reward. However, after the exploration phase, some other arm $a \neq a^*$ may be chosen because it has a higher empirical average reward, i.e., $\hat{\mu}_t(a) > \hat{\mu}_t(a^*)$. We know that, with probability at least $1 - \frac{2}{t^4}$,

$$|\hat{\mu}_t(a) - \mu(a)| \leq \sqrt{\frac{2K\log t}{t\epsilon_t}}$$

$$|\hat{\mu}_t(a^*) - \mu(a^*)| \leq \sqrt{\frac{2K\log t}{t\epsilon_t}}$$

which leads to,

$$\mu(a) + \sqrt{\frac{2K\log t}{t\epsilon_t}} \geq \hat{\mu}_t(a) > \hat{\mu}_t(a^*) \geq \mu(a^*) - \sqrt{\frac{2K\log t}{t\epsilon_t}}$$

Hence,

$$\mu(a^*) - \mu(a) \leq 2\sqrt{\frac{2K\log t}{t\epsilon_t}}$$

Notice that this bound holds for all a that appear in the exploitation phase when $a \neq a^*$. Hence, we can

bound the expected value of the regret at a particular round t as follows,

$$\begin{aligned}
\mathrm{E}[\tilde{R}(t)] &\leq (1-\epsilon_t).(\mu(a^*) - \mu(a_t)) + \epsilon_t.1 \\
&\leq (1-\epsilon_t).2\sqrt{\frac{2K\log t}{t\epsilon_t}} + \epsilon_t \\
&\leq 2\sqrt{\frac{2K\log t}{t\epsilon_t}} + \epsilon_t
\end{aligned}$$

Here, $\mu(a_t)$ denotes the expected reward of the arm chosen at time t. Moreover, we multiply $(1-\epsilon_t)$ by 1, since the regret can be at most 1 during the exploration phase for simplicity. This is due to the fact that the support of the reward distribution lies within $[0, 1]$.

Recall that we can specify ϵ_t as a function of t. Therefore, we can choose ϵ_t to approximately minimize the right-hand side. Since the two summands are, respectively, monotonically decreasing and monotonically increasing in ϵ_t, we can set ϵ_t such that the two terms are approximately equal. This implies,

$$\epsilon_t = \sqrt{\frac{K\log t}{t\epsilon_t}} \implies \epsilon_t^2 = \frac{K\log t}{t\epsilon_t} \implies \epsilon_t = t^{\frac{-1}{3}}\,(K\log t)^{\frac{1}{3}}$$

Hence,

$$\mathrm{E}[\tilde{R}(t)] \leq O(t^{\frac{-1}{3}}\,(K\log t)^{\frac{1}{3}})$$

By using this choice of ϵ_t, we can now compute the overall regret as follows,

$$\begin{aligned}
\mathrm{E}[R(t)] &= \mathrm{E}\left[\sum_{k=1}^{t}\tilde{R}(k)\right] \\
&\leq tO(t^{\frac{-1}{3}}\,(K\log t)^{\frac{1}{3}}) \text{ (since } \mathrm{E}[\tilde{R}(t)] \text{ is a non-decreasing function in terms of } t) \\
&\leq O(t^{\frac{2}{3}}\,(K\log t)^{\frac{1}{3}})
\end{aligned}$$

□

This provides a better bound for the ϵ-greedy strategy, as it allows the expected regret to grow sub-linearly. However, Cesa-Bianchi and Fischer [3] proposed an ϵ-decreasing greedy strategy with $O(\log^2 T)$ regret for certain reward distributions. Moreover, Auer et al. [1] achieved an $O(\log T)$ regret bound using an ϵ-decreasing strategy, under specific constraints on the choice of the initial value ϵ_0.

The decayed value of ϵ_t in Theorem 1 decreases relatively slowly with respect to time t. This has the advantage that the algorithm continues exploring long enough for the empirical mean of each arm to converge to its true mean, thereby reducing the risk of exploiting suboptimal arms.

3.2 Upper Confidence Bound (UCB) algorithm

The UCB family of algorithms was proposed by Auer et al. in [1]. These algorithms involve maintaining empirical confidence bounds for the mean reward of each arm. The strategy always selects the arm with the highest upper confidence bound (UCB). To implement this, we first pull each arm once to obtain an empirical mean $\hat{\mu}_t(a)$ for each arm a. Then, for any subsequent time t, we define the upper and lower confidence bounds for each arm a as follows,

$$UCB_t(a) := \hat{\mu}_t(a) + r_t(a), \quad LCB_t(a) := \hat{\mu}_t(a) - r_t(a)$$

where $\hat{\mu}_t(a)$ is the empirical mean reward for arm a observed up to time t, and the confidence radius $r_t(a)$ is defined by,

$$r_t(a) = \sqrt{\frac{2 \log T}{n_t(a)}}$$

Here, $n_t(a)$ denotes the number of times the algorithm has selected arm a up to time t. Note that the confidence radius $r_t(a)$ only changes when arm a is chosen. After pulling each arm once to obtain initial confidence intervals, the UCB strategy proceeds by selecting the arm with the largest UCB at each time step.

The motivation behind this strategy is that there are two possible reasons why an arm might have a high UCB: either its empirical mean reward is large, or its confidence radius is large, indicating that the arm has not been explored much yet. Both cases suggest that the arm is worth trying, until its UCB drops below that of another arm.

Algorithm 2 Upper Confidence Interval (UCB) Algorithm

1: Try each arm once
2: In each round t after the first phase, pick the $\arg\max_a UCB_t(a)$, which means the arm with the highest upper confidence bound.

Theorem 2. *Assume the rewards of each arm are in* $[0, 1]$. *Then the UCB algorithm achieves regret*

$\mathrm{E}[R(t)] \leq O\left(\sqrt{Kt \log T}\right)$ *for each round* t.

Proof. Consider a "clean" event A_t, which denotes that the true means of the reward distributions lie within their respective confidence intervals up to round t, i.e.,

$$\begin{aligned} A_t &= \{\forall i \leq t, \forall a : \mu(a) \in [LCB_i(a), UCB_i(a)]\} \\ &= \{\forall i \leq t, \forall a : |\hat{\mu}_t(a) - \mu(a)| \leq r_i(a)\} \end{aligned}$$

By the law of total expectation,

$$\mathrm{E}[R(t)] = \mathrm{E}[R(t)|A_t].P(A_t) + \mathrm{E}[R(t)|\bar{A}_t].P(\bar{A}_t)$$

Now consider the second term, which corresponds to the "unclean" event $\bar{A}_t$. Since the expected regret per time step is at most 1, we have $\mathrm{E}[R(t)|\bar{A}_t] \leq t$. It therefore remains to bound $P(\bar{A}_t)$. By Hoeffding's inequality, we again have that, for any given arm a and time t,

$$\begin{aligned} P(|\hat{\mu}_t(a) - \mu(a)| > r_t(a)) &\leq 2\exp\left(\frac{-2r_t^2(a).n_t(a)}{(1-0)^2}\right) \\ &= 2\exp\left(-2.\frac{2\log T}{n_t(a)}.n_t(a)\right) \\ &= 2\exp\left(-4\log T\right) \\ &= \frac{2}{T^4} \end{aligned}$$

Assume that a_t is the arm chosen at time t. The only confidence radius and empirical mean changing at time t are those of arm a_t. Therefore by union bound we have,

$$\begin{aligned} P(\bar{A}_t) &\leq \sum_{i=1}^{t} P\left(|\hat{\mu}_i(a_i) - \mu(a_i)| > r_i(a_i)\right) \\ &\leq t.\frac{2}{T^4} \\ &= O\left(\frac{t}{T^4}\right) \end{aligned}$$

Hence, the second term associated with the unclean event satisfies,

$$\mathrm{E}[R(t)|\bar{A}_t].P(\bar{A}_t) \leq O\left(\frac{t^2}{T^4}\right)$$

Now we need to bound the first term (clean event). Assume that a_t is the arm chosen at time t, i.e.,

$$UCB_t(a_t) \geq UCB_t(a^*)$$

Since we are considering the clean event

$$\mu(a_t) \in [LCB_t(a_t), UCB_t(a_t)]$$

i.e,

$$\hat{\mu}_t(a_t) - r_t(a_t) \leq \mu(a_t) \leq \hat{\mu}_t(a_t) + r_t(a_t)$$

$$\implies \mu(a_t) + 2r_t(a_t) >= \hat{\mu}_t(a_t) - r_t(a_t) + 2r_t(a_t) = UCB_t(a_t)$$

and

$$UCB_t(a^*) \geq \mu(a^*)$$

which always hold in the clean event. Hence,

$$\mu(a_t) + 2r_t(a_t) \geq \mu(a^*) \implies \mu(a^*) - \mu(a_t) \leq 2r_t(a_t)$$

Thus the contribution of any arm a to the regret up to round t is,

$$\left(\mu(a^*) - \mu(a)\right).n_t(a) \leq 2r_t(a).n_t(a) = 2\sqrt{\frac{2\log T}{n_t(a)}}.n_t(a) = 2\sqrt{2\log T.n_t(a)}$$

So the total regret for clean event $\mathrm{E}[R(t)|A_t]$ is,

$$2\sqrt{2\log T}\sum_a \sqrt{n_t(a)}$$

Since $\sqrt{.}$ is concave, we can use Jensen's inequality to bound the sum,

$$\sum_a \sqrt{n_t(a)} \leq K\sqrt{\frac{1}{K}\sum_a n_t(a)} = \sqrt{tK}$$

Thus the total regret in the clean event, $\mathrm{E}[R(t)|A_t]$ is bounded by,

$$2\sqrt{2\log T}.\sqrt{tK} = O\left(\sqrt{Kt\log T}\right)$$

Combining all the above results, we finally obtain the total expected regret as follows,

$$\begin{aligned}
\mathrm{E}[R(t)] &= \mathrm{E}[R(t)|A_t].P(A_t) + \mathrm{E}[R(t)|\bar{A}_t].P(\bar{A}_t) \\
&\leq O\left(\sqrt{Kt\log T}\right) + O\left(\frac{t^2}{T^4}\right) \quad (P(A_t) = 1 - \frac{2}{T^4} \text{ can be bounded by 1}) \\
&= O\left(\sqrt{Kt\log T}\right)
\end{aligned}$$

□

3.3 Softmax algorithm (Boltzmann Exploration)

The Softmax algorithm (also known as Boltzmann Exploration) is a strategy in which each arm is selected with a probability proportional to the exponential of its empirical mean reward. Since exponentials cause larger values to grow multiplicatively, the arm with the highest empirical mean (i.e., corresponding to the largest logit) tends to dominate the selection probability.

Algorithm 3 Softmax Algorithm

1: Given intial empirical mean $\hat{\mu}_0(a)$ for all arm a
2: The arm a is chosen at round t with probability $p_t(a) = \frac{e^{\eta_t \hat{\mu}_{t-1}(a)}}{\sum_a e^{\eta_t \hat{\mu}_{t-1}(a)}}$, here η_t is the learning rate.

The most basic case of this algorithm assumes a constant learning rate across all rounds. We now consider a theorem that provides a regret bound when an adaptive learning rate is used in the Softmax algorithm.

Theorem 3. *Assume the rewards of each arm are in* $[0, 1]$ *and let* $\tau = \frac{16eK\log T}{\Delta^2}$. *Then the regret of*

Softmax strategy with learning rate $\eta_t = \mathbf{1}_{\{t<\tau\}} + \frac{\log(t\Delta^2)}{\Delta}.\mathbf{1}_{\{t>=\tau\}}$*, satisfies,*

$$\mathrm{E}[R(T)] \leq \frac{16eK\log T}{\Delta^2} + \frac{9K}{\Delta^2} = O\left(\frac{K\log T}{\Delta^2}\right)$$

where $\Delta := \min_a(\Delta_a)$*, for* $\Delta_a := \mu(a^*) - \mu(a)$*.*

Proof. For any round t and arm a, since $\hat{\mu}_{t-1}(a) \in [0,1]$, hence,

$$\frac{e^{\eta_t \hat{\mu}_{t-1}(a)}}{\sum_a e^{\eta_t \hat{\mu}_{t-1}(a)}} \geq \frac{1}{Ke^{\eta_t}} \tag{3}$$

and,

$$\frac{e^{\eta_t \hat{\mu}_{t-1}(a)}}{\sum_a e^{\eta_t \hat{\mu}_{t-1}(a)}} \leq e^{\eta_t(\hat{\mu}_{t-1}(a) - \hat{\mu}_{t-1}(a^*))} \tag{4}$$

Now, for any arm rather than the optimal arm a^*, for a_t is the arm chosen at time t, $\Delta_a = \mu(a^*) - \mu(a)$, we can write,

$$\mathbf{1}_{\{a_t=a\}} = \mathbf{1}_{\{a_t=a, \hat{\mu}_{t-1}(a) - \hat{\mu}_{t-1}(a^*) < -\frac{\Delta_a}{2}\}} + \mathbf{1}_{\{a_t=a, \hat{\mu}_{t-1}(a) - \hat{\mu}_{t-1}(a^*) \geq -\frac{\Delta_a}{2}\}} \tag{5}$$

Apply the union bound,

$$\mathbf{1}_{\{\hat{\mu}_{t-1}(a) - \hat{\mu}_{t-1}(a^*) \geq -\frac{\Delta_a}{2}\}} \leq \mathbf{1}_{\{\hat{\mu}_{t-1}(a) \geq \mu(a) + \frac{\Delta_a}{4}\}} + \mathbf{1}_{\{\hat{\mu}_{t-1}(a^*) \leq \mu(a^*) - \frac{\Delta_a}{4}\}}$$

Hence, 5 becomes,

$$\mathbf{1}_{\{a_t=a\}} = \mathbf{1}_{\{a_t=a, \hat{\mu}_{t-1}(a) - \hat{\mu}_{t-1}(a^*) < -\frac{\Delta_a}{2}\}} + \mathbf{1}_{\{a_t=a, \hat{\mu}_{t-1}(a) \geq \mu(a) + \frac{\Delta_a}{4}\}} + \mathbf{1}_{\{a_t=a, \hat{\mu}_{t-1}(a^*) \leq \mu(a^*) - \frac{\Delta_a}{4}\}} \tag{6}$$

Now, we take the expectation of the three terms above and sum over $t = \lfloor \tau \rfloor + 1, \ldots, T$. Since $t > \tau$, we

have $\eta_t = \frac{\log(t\Delta^2)}{\Delta}$. Then, by 4,

$$\begin{aligned}
\sum_{t=\lfloor\tau\rfloor+1}^{T} P\left(a_t = a, \hat{\mu}_{t-1}(a) - \hat{\mu}_{t-1}(a^*) < -\frac{\Delta_a}{2}\right) &\leq \sum_{t=\lfloor\tau\rfloor+1}^{T} e^{-\eta_t \cdot \frac{\Delta_a}{2}} \\
&\leq \sum_{t=\lfloor\tau\rfloor+1}^{T} e^{-\frac{\log(t\Delta^2)}{\Delta} \cdot \frac{\Delta}{2}} \quad (\text{since } \Delta_a \geq \Delta) \\
&= \sum_{t=\lfloor\tau\rfloor+1}^{T} \frac{1}{t\Delta^2} \\
&\leq \frac{\log(T+1)}{\Delta^2}
\end{aligned} \tag{7}$$

Now, we control the third term of 6 in the same way. For the third term, we have that,

$$\mathbf{1}_{\{\hat{\mu}_{t-1}(a^*) \leq \mu(a^*) - \frac{\Delta_a}{4}\}} \leq \mathbf{1}_{\{n_{t-1}(a^*) \leq t_1\}} + \mathbf{1}_{\{\hat{\mu}_{t-1}(a^*) \leq \mu(a^*) - \frac{\Delta_a}{4}, n_{t-1}(a^*) > t_1\}} \tag{8}$$

holds for any fixed t and for any $t_1 \leq t-1$. Hence,

$$\begin{aligned}
\sum_{t=\lfloor\tau\rfloor+1}^{T} P\left(\hat{\mu}_{t-1}(a^*) \leq \mu(a^*) - \frac{\Delta_a}{4}\right) \leq \sum_{t=\lfloor\tau\rfloor+1}^{T} P\left(n_{t-1}(a^*) \leq t_1\right) + \\
\sum_{t=\lfloor\tau\rfloor+1}^{T} P\left(\hat{\mu}_{t-1}(a^*) \leq \mu(a^*) - \frac{\Delta_a}{4}, n_{t-1}(a^*) > t_1\right)
\end{aligned} \tag{9}$$

Consider the first term, for a_t as the arm chosen at time t, we apply 3 during the first $\lfloor\tau\rfloor$ rounds. In this case, $\eta_t = 1$, and we have,

$$P(a_t = a^*) \geq \frac{1}{Ke^{\eta_t}}$$

Hence,

$$\mathrm{E}[n_{t-1}(a^*)] \geq \frac{e^{\eta_t}}{K}.\tau = \frac{\tau}{eK}$$

holds for all $t < \tau$. By setting $t_1 = \frac{1}{2}\mathrm{E}[n_{t-1}(a^*)] \geq \frac{\tau}{2eK}$, Chernoff bounds (in multiplicative form) give,

$$\begin{aligned}
P\left(n_{t-1}(a^*) \leq t_1\right) &= P\left(n_{t-1}(a^*) \leq \left(1 - \frac{1}{2}\right)\mathrm{E}[n_{t-1}(a^*)]\right) \\
&\leq e^{-\mathrm{E}[n_{t-1}(a^*)]/8} \\
&\leq e^{-\frac{\tau}{8eK}}
\end{aligned} \tag{10}$$

Now, move to the second term, using Hoeffding inequality,

$$
\begin{aligned}
P\left(\hat{\mu}_{t-1}(a^*) \leq \mu(a^*) - \frac{\Delta_a}{4}, n_{t-1}(a^*) > t_1\right) &= \sum_{s=t_1+1}^{t-1} P\left(\hat{\mu}_{t-1}(a^*) \leq \mu(a^*) - \frac{\Delta_a}{4}, n_{t-1}(a^*) = s\right) \\
&\leq \sum_{s=t_1+1}^{t-1} e^{-\frac{s\Delta_a^2}{8}} \quad \text{(Hoeffding inequality)} \\
&\leq \sum_{s=t_1}^{\infty} e^{-\frac{s\Delta^2}{8}} \\
&\leq \int_{t_1}^{\infty} e^{-\frac{s\Delta^2}{8}}\, ds \\
&= \frac{8}{\Delta^2} e^{-\frac{t_1\Delta^2}{8}} \\
&\leq \frac{8}{\Delta^2} e^{-\frac{\tau\Delta^2}{16eK}}
\end{aligned} \tag{11}
$$

Therefore, combine 10 and 11,

$$
\begin{aligned}
\sum_{t=\lfloor\tau\rfloor+1}^{T} P\left(\hat{\mu}_{t-1}(a^*) \leq \mu(a^*) - \frac{\Delta_a}{4}\right) &\leq T\left(e^{-\frac{\tau}{8eK}} + \frac{8}{\Delta^2} e^{-\frac{\tau\Delta^2}{16eK}}\right) \\
&= T\left(e^{-\frac{16eK\log T}{\Delta^2 eK}} + \frac{8}{\Delta^2} e^{-\frac{16eK\log T\Delta^2}{\Delta^2 16eK}}\right) \\
&\leq 1 + \frac{8}{\Delta^2}
\end{aligned} \tag{12}
$$

The second term in 6 can be bounded exactly the same way. Putting together, we have thus obtained, for all arms a that are not the optimal arm a^*,

$$
\begin{aligned}
\mathrm{E}[R(T)] = \sum_a \Delta_a \mathrm{E}[n_T(a)] &\leq \sum_a \mathrm{E}[n_T(a)] \\
&\leq \frac{K\log(T+1)}{\Delta^2} + K + \frac{8K}{\Delta^2} \\
&\leq \tau + \frac{9K}{\Delta^2} \\
&\leq \frac{16eK\log T}{\Delta^2} + \frac{9K}{\Delta^2} \\
&= O\left(\frac{K\log T}{\Delta^2}\right)
\end{aligned}
$$

□

The proof is tedious, but the main idea is to bound $\Delta_a := \mu(a^*) - \mu(a)$ by 1 for simplicity. The

remaining task is then to control the expected number of times the non-optimal arms are selected.

4 Experiment

In this section, we conduct empirical experiments on the expected regret and mean reward over time. The main objective is to examine specific scenarios in which one algorithm may outperform the others. We are also interested in evaluating whether the sublinear regret bounds are satisfied, as predicted by the theoretical results. To ensure meaningful comparisons, we maintain consistency in the choice of the probability schedule ϵ_t and the learning rate η_t as specified in the corresponding theorems.

We run experiments with varying numbers of arms K and different number of rounds T. We choose the Bernoulli distribution for simplicity, as its support is $\{0, 1\}$. Specifically, each arm follows a Bernoulli distribution with a mean drawn uniformly from the interval $[0, 1]$. In addition to comparing the three main algorithms, we also examine the performance differences between the constant and decayed versions of the ϵ-greedy algorithm. The final results, based on expected regret and mean reward, are illustrated in the plots below.

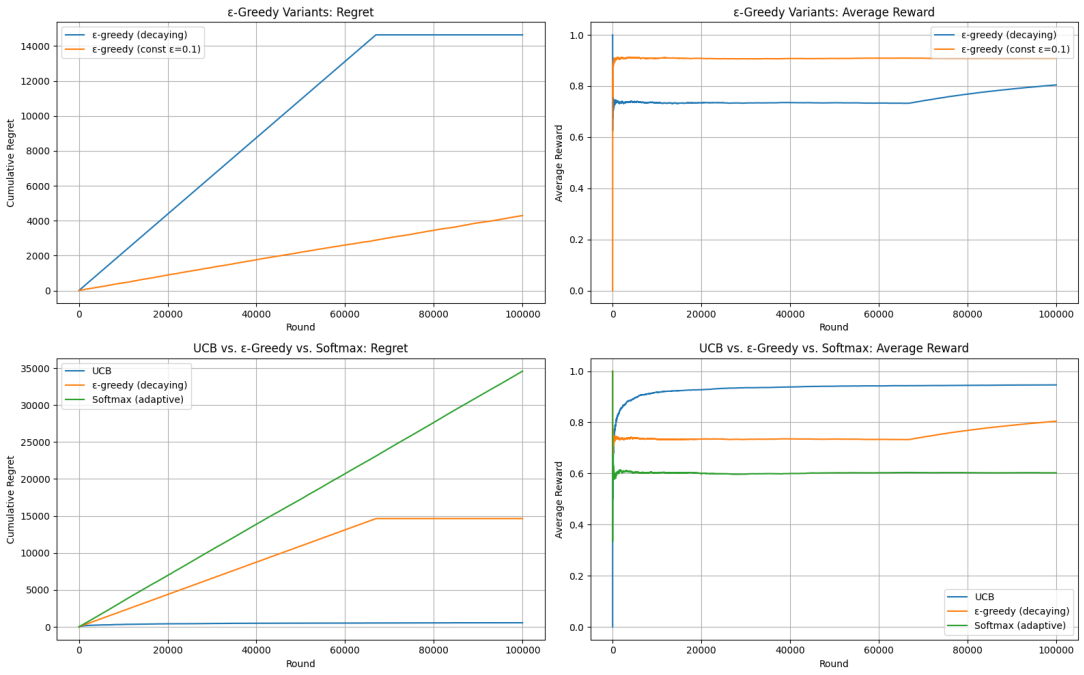

Figure 1: K=10, T=100000

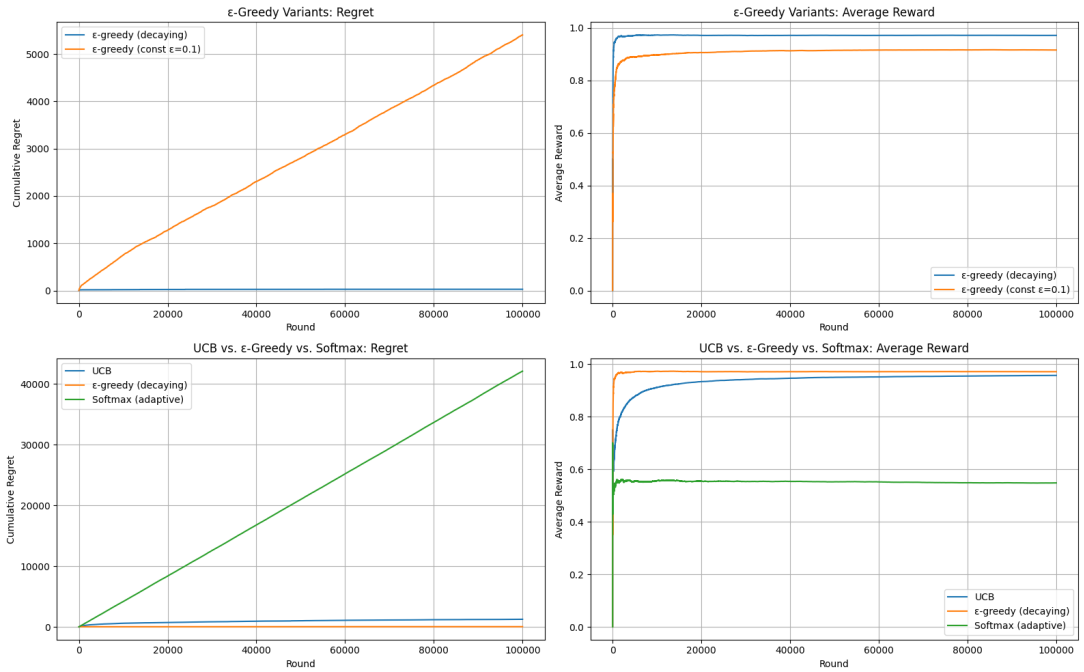

Figure 2: K=20, T=100000

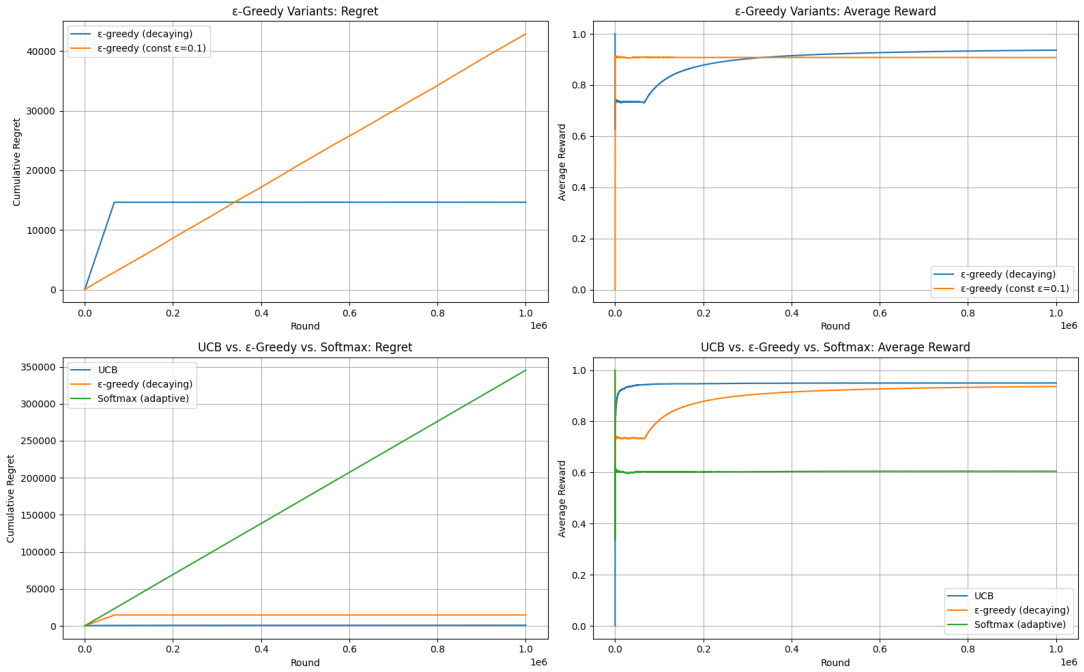

Figure 3: K=10, T=1000000

From the plot, it is clear that both the decayed ϵ-greedy and UCB algorithms consistently achieve sublinear regret, which aligns with the theoretical guarantees. As expected, the regret of the constant ϵ-greedy algorithm grows linearly. Interestingly, for small values of K and T, the constant ϵ-greedy variant outperforms the decayed version. However, for larger values of T, as ϵ_t decreases toward zero, the algorithm has typically already identified the optimal arm and continues to select it for the remaining rounds - an outcome that is quite intuitive.

One notable observation is that the Softmax algorithm performs poorly across all tested values of K and T. This contradicts the bound given in Theorem 3, which suggests that Softmax should, in theory, achieve strong performance. However, as noted in [4], Theorem 1, when a monotone learning rate is used, the resulting reward sequence can induce suboptimal behavior. This may explain the problem observed in our experiments.

5 Conclusions

In this project, we studied the multi-armed bandit (MAB) problem and conducted a comparative analysis of three well-known algorithms: ϵ-greedy, Upper Confidence Bound (UCB), and Softmax, with the objective of minimizing the expected regret over T rounds. To support this, we analyzed the asymptotic regret bounds, aiming for sublinear growth. Through theoretical analysis, we found that appropriately designing the exploration probability ϵ_t and learning rate η_t as functions of time t can lead to sublinear regret.

Our experiments confirmed that both the ϵ-greedy and UCB algorithms perform in line with theoretical expectations. However, the Softmax algorithm showed poor empirical performance, deviating from the theoretical guarantees, which may be explained by the monotone learning rate, as noted in [4].

The choice of algorithm can depend on the number of arms K and the number of rounds T. In particular, when K is small or during the early rounds, the constant ϵ-greedy variant performs well and may be a practical choice in real-world applications.

This project only introduced some basic variants of the studied algorithms. In the literature, there are many advanced modifications that achieve improved regret bounds. Additionally, we did not consider the role of the variance of the reward distributions D_a, which is another important factor influencing algorithm performance. Those can be considered as recommendations for future research.

References

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[2] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*, 2019.

[3] Nicolo Cesa-Bianchi and Paul Fischer. Finite-time regret bounds for the multiarmed bandit problem. In *ICML*, volume 98, pages 100–108. Citeseer, 1998.

[4] Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. *Advances in neural information processing systems*, 30, 2017.

[5] Robert Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17, 2004.

[6] Robert Kleinberg. Anytime algorithms for multi-armed bandit problems. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 928–936, 2006.

[7] Volodymyr Kuleshov and Doina Precup. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028*, 2014.

[8] Ioannis Panageas. Lecture 9: Introduction to multi-armed bandits, 2025. Lecture notes for Optimization for Machine Learning 50.579, Northeastern University. Scribed by Deepanway Ghosal and Wayne Lin.

[9] Herbert Robbins. Some aspects of the sequential design of experiments. 1952.

[10] Aleksandrs Slivkins. Introduction to multi-armed bandits. *CoRR*, abs/1904.07272, 2019.

[11] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[12] Joannes Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *European conference on machine learning*, pages 437–448. Springer, 2005.

[13] Christopher John Cornish Hellaby Watkins et al. Learning from delayed rewards. 1989.

[10, 7, 12, 5, 6, 9, 2, 11, 1, 3, 13, 8, 4]