Comparison between Arm Selection Policies

HUANG, Hao Yu

A Report Submitted for the Course CMSC5728 Assignment#2 of

Master of Science

in

Computer Science

The Chinese University of Hong Kong

Nov 2024

Contents

Chapter 1

Introduction

1.1 Stochastic MAB

In this section, we simply clarify the concept of stochastic MAB, which is the random arm selection policy.

In this arm selection policy, our algorithm can not learn anything from experiences.

Algorithm 1: Stochastic MAB

Input: N: number of arms, T: total number of time slots

for *time slot* $t \in [T]$ **do**
    The arm selection algorithm selects arm a_t ;
    Learner receives reward $r_t \in \mathbb{R}$ for the selected arm
end

Figure 1.1: Stochastic MAB

1.2 ETE Policy

In this section, we simply clarify the concept of Explore-then-Exploit(ETE) arm selection policy.

We assume that the number of arms is N. In the ETE arm selection policy, we use $m \times N$ time slots to do the explorations and after that we do the exploitations.

Algorithm 2: Explore-then-exploit (ETE) arm selection policy

Input : N: number of arms, T: total number of time slots; m: number of times to explore an arm

Output: $A(T)$: cumulative reward of the ETE arm selection algorithm

for $j = 1, \ldots, N$ **do**
  $\hat{\mu}_i = 0$; /* initialize all empirical estimates */
end
$A(T) = 0.0$ /* initialize cumulative reward to zero */

/* ———— Exploration Phase ———— */
for *time slot* $t \in \{1, 2, \ldots, mN\}$ **do**
  Select arm $a_t = t \mod (N + 1)$;
  reward $= r_t$;
  $\hat{\mu}_{a_t}$ += reward ; /* accumulate rewards for arm a_t */
  $A(T)$ += reward ; /* accumulate arm selection reward */
end
for $i = 1, \ldots, N$ **do**
  $\hat{\mu}_i(mN) = \hat{\mu}_i / m$; /* compute empirical reward for arm i */
end

/* ———— Exploitation Phase ———— */
$\hat{a} = \text{argmax}_i\{\hat{\mu}_1(mN), \hat{\mu}_2(mN), \ldots, \hat{\mu}_N(mN)\}$; /* find the best arm */
for *time slot* $t \in \{mN + 1, \cdots, T\}$ **do**
  select arm $\hat{a}$;
  reward $= r_t$;
  $A(T)$ += reward ; 1/0 * reward, which is r_t.
end

Figure 1.2: ETE Policy

1.3 Basic Epsilon-Greedy Policy

In this section, we simply clarify the concept of basic epsilon-greedy arm selection policy.

In this algorithm, we spread out the exploration for all T time slots. At each

time slot, we may do exploration or exploitation.

1.4 Adaptive Epsilon-Greedy Policy

In this section, we simply clarify the concept of adaptive epsilon-greedy arm selection policy.

In this algorithm, compared with basic epsilon-greedy arm selection policy, the ϵ is a function of the current time slot.

Algorithm 3: ϵ-Greedy Arm Selection Algorithm

Input : N: number of arms, T: total number of time slots
Output: $A(T)$: cumulative reward of the ϵ-Greedy algorithm

```
A(T) = 0 ;
for time slot t ∈ [T] do
    Generate a random number ∈ [0, 1] ;
    if the random number is less than ε_t = (N ln(t)/t)^{1/3} then
        arm a_t is selected uniformly from the set [N] ;   /* perform exploration */
    else
        a_t = argmax_{i∈[N]} μ̂_i(t)                        /* perform exploitation */;
    end
    reward = r_t ;
    A(T) += reward ;
    update μ̂_{a_t}(t);   /* update empirical estimate of the selected arm */
end
```

Figure 1.3: Adaptive Epsilon-Greedy Policy

1.5 UCB Policy

In this section, we simply clarify the concept of Upper Confidence Bound (UCB) arm selection policy.

In this algorithm, we select the arm with the highest upper bound in average reward.

Algorithm 3: ϵ-Greedy Arm Selection Algorithm

Input : N: number of arms, T: total number of time slots
Output: $A(T)$: cumulative reward of the ϵ-Greedy algorithm

$A(T) = 0$;
for *time slot* $t \in [T]$ **do**
  Generate a random number $\in [0, 1]$;
  if *the random number is less than* $\epsilon_t = (N \ln(t)/t)^{1/3}$ **then**
    arm a_t is selected uniformly from the set $[N]$; /* perform exploration */
  else
    $a_t = \text{argmax}_{i \in [N]} \hat{\mu}_i(t)$ /* perform exploitation */;
  end
  reward $= r_t$;
  $A(T)$ += reward ;
  update $\hat{\mu}_{a_t}(t)$; /* update empirical estimate of the selected arm */
end

Figure 1.4: UCB Policy

Chapter 2

Comparison

2.1 Different Policies

In this section, I will compare the performances of different arm selection policies. We set $T = 10000, N = 10$, and each arm follows the uniform probability distributions. For 10 arms, we set the interval of each arm as $[0, 2]$, $[1, 9]$, $[1, 3]$, $[3, 5]$, $[4, 6]$, $[2, 10]$, $[1, 5]$, $[4, 10]$, $[5, 7]$, $[0, 10]$, in which the optimal arm is the eighth arm. And to let the result be more reasonable, in each time slot we do 200 times experiments. And the result of each time slot is the average of results of 200 times experiments.

We illustrate the performance of each policy by the accumulative average rewards and accumulative average regrets.

Firstly, we demonstrate the performance of cumulative average regrets. As shown in 2.1, there are 14 policies. And here we introduce the "Linear" performance of cumulative average regrets, because the distribution of each arm is uniform distribution and the regret in each time slot can be bigger than 1. So the real linear function can not reflect the true performance of not learning anything. So here we define the "Linear" function as shown in 2.1, 2.2, 2.3, 2.4.

$$E[R(T)] = \sum_{i=1}^{T}(\mu^* - \bar{\mu}) \tag{2.1}$$

$$\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N}\bar{\mu_i} \tag{2.2}$$

$$\bar{\mu_i} = E_{reward}(\mu_i) = \frac{a_i + b_i}{2} \tag{2.3}$$

$$\mu^* = argmax_{i=1}^{N}\bar{\mu_i} \tag{2.4}$$

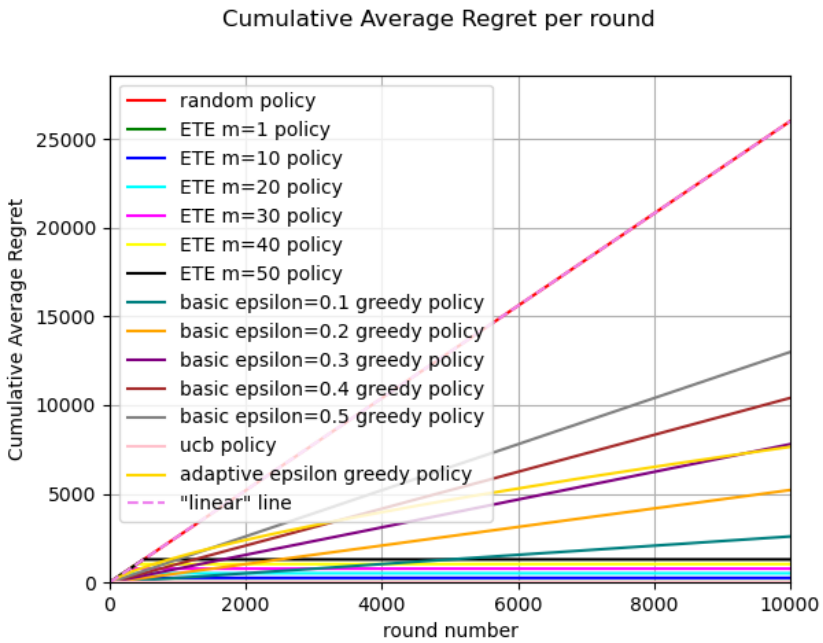

Figure 2.1: Cumulative Average Regret

And a_i, b_i are the lower bound and upper bound of the i-th arm. In the "Linear" function, we do not learn anything from experiences and I take the expectation reward of all arms as the reward I get in each time slot. And it is the same as random selection policy, when the time of experiments conducted is big enough. And as shown in 2.1, the random policy is the same as our defined "linear" line. And when $\epsilon = 0.1$, the performance of epsilon-greedy policy is the

best in our arm distribution. And when $m = 1$, the performance of ETE policy is the best. And the UCB policy have the best performance of all policies we demonstrate here. And we also find that when $m = 1$, the performances of ETE policy and UCB policy are the same in our arm distribution.

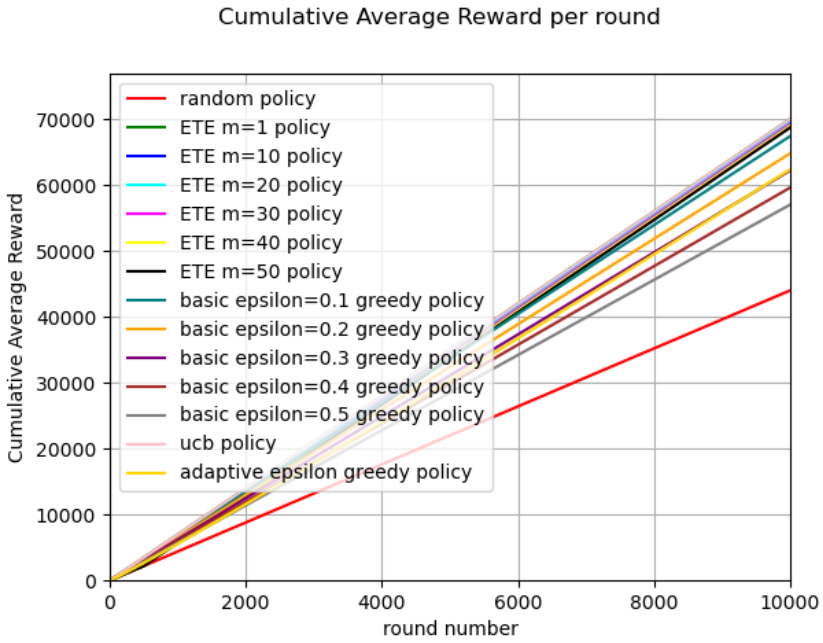

Figure 2.2: Cumulative Average Reward

Fig.2.2 shows the cumulative average reward of each arm selection policy. And the conclusions from this are the same as those from cumulative average regrets. And the worst is random selection, which is reasonable.

2.2 Different T

In this section, we use the same settings we set in the last section. But we set $T = 1000, 1500, 2000$ to do the experiments respectively as shown in 2.3, 2.4, 2.5, 2.6, 2.7, 2.8. And we can see that when the T is small, the performance of $\epsilon = 0.5$

epsilon-greedy policy is better than ETE $m = 10$ policy. But the T is large, the performance of ETE $m = 10$ policy is better than $\epsilon = 0.5$ epsilon-greedy policy.

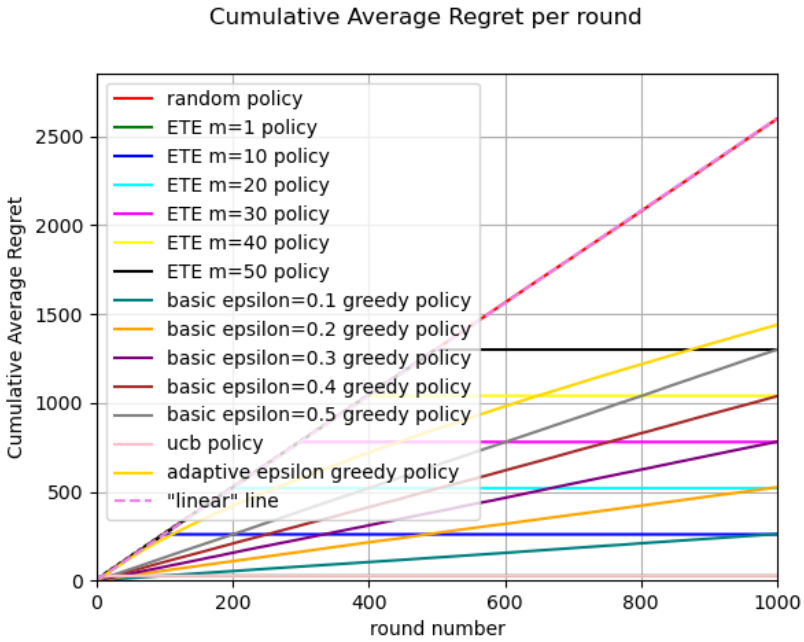

Figure 2.3: Cumulative Average Regret-T=1000

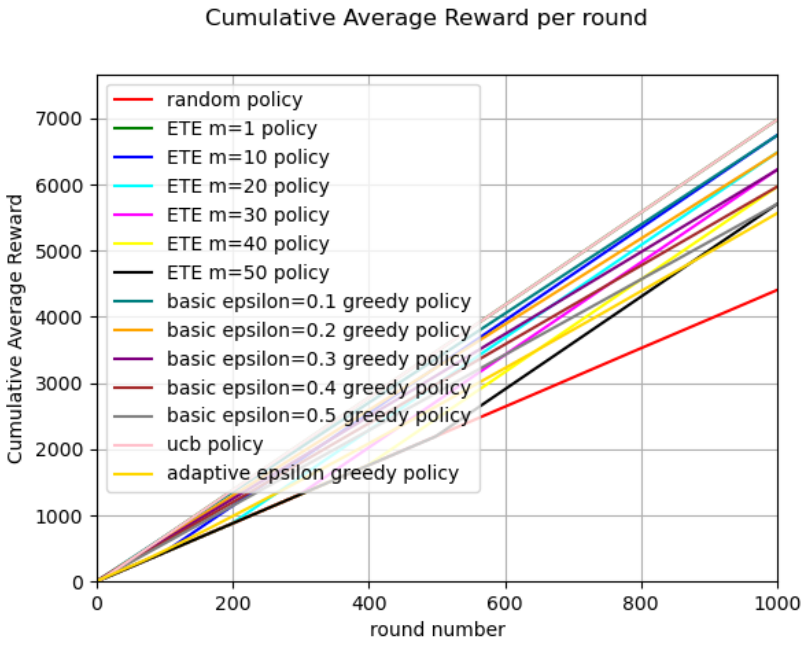

Figure 2.4: Cumulative Average Reward-T=1000

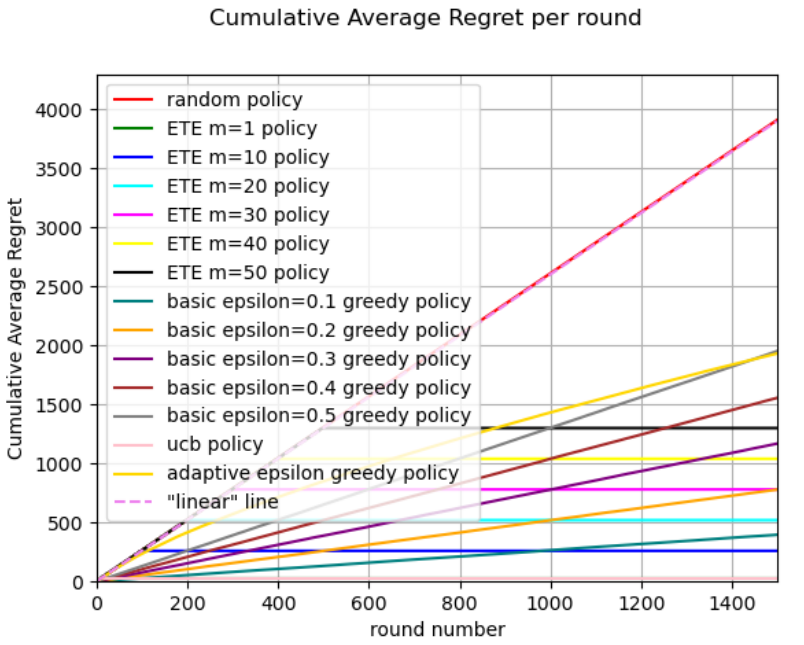

Figure 2.5: Cumulative Average Regret-T=1500

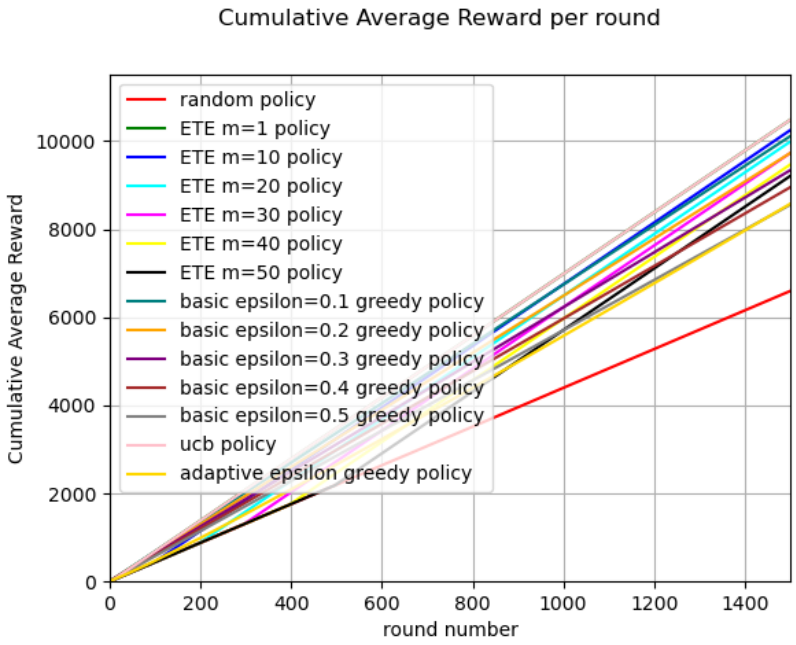

Figure 2.6: Cumulative Average Reward-T=1500

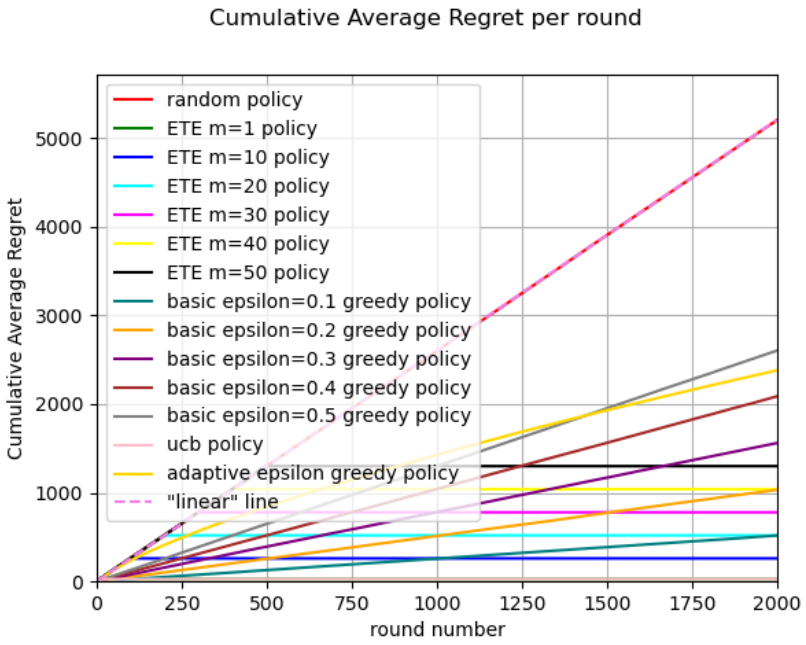

Figure 2.7: Cumulative Average Regret-T=2000

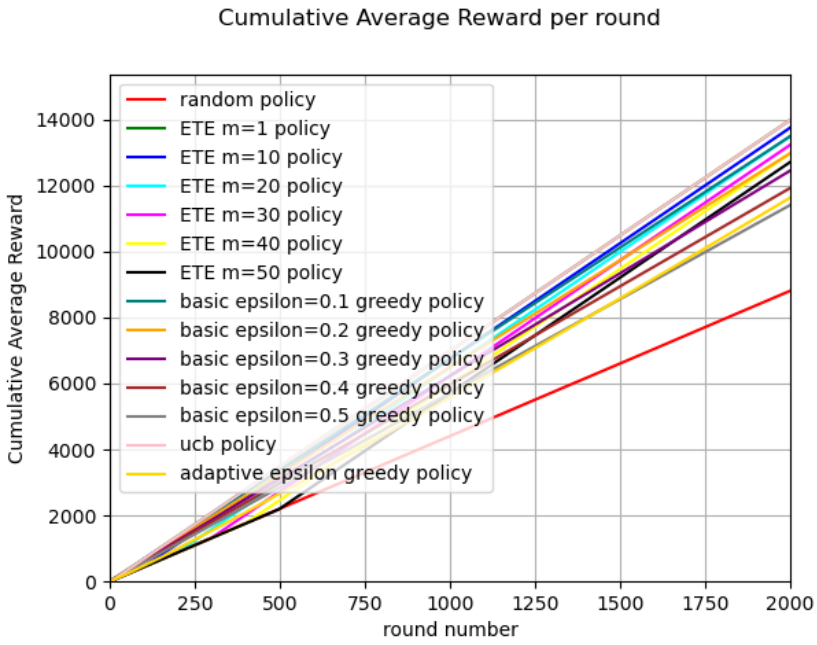

Figure 2.8: Cumulative Average Reward-T=2000