Homework Assignment 6

Zhaoyang Xu

October 20, 2020

Contents

1 Offline Evaluation of Bandit Algorithms (60 points)

1.1

Algorithm UCB1

Initialization : Play each action once.
for t=N+1,N+2,... **do**
  Play A_t=arg max$\hat{\mu}_{t-1}(a) + \sqrt{\frac{3\ln t}{2N_{t-1}(a)}}$
  Set $\hat{\mu}_t = \frac{\hat{L}_t(a)}{N_t(a)}$
  $\hat{L}_t(a) = \hat{L}_{t-1}(a) + l_t^a * K$
end for

Algorithm EXP3

Input : Learning rate η
$\forall a$:$\hat{L}_0(a) = 0$
for t=1,2,... **do**
  $\forall a : p_t(a) = \frac{e^{-\eta_t \hat{L}_{t-1}(a)}}{\sum_{a'} e^{-\eta_t \hat{L}_{t-1}(a')}}$
  Sample A_t according to p_t and play it
  Observe and suffer $l_t^{A_t}$
  Set $\hat{l}_t^a = \frac{l_t^a \mathbb{1}(A_t = a) * K}{p_t(a)}$
  $\forall a : \hat{L}_t(a) = \hat{L}_{t-1}(a) + \hat{l}_t^a$ **end for**

where K is the number of actions.

From question, we can bound the expected regret of EXP3 with $\eta = \sqrt{\frac{\ln K}{tK}}$

$$\mathbb{E}[R_T] \leq 2\sqrt{tK \ln K}$$

The proof is based on lemma 5.2 on page 63 in notes.
Let $X_1^a, X_2^a, ...$ be K sequences of non-negative numbers. Let $L_t(a) = \sum_{s=1}^{t} X_s^a$. We define $W_t = \sum_a e^{-\eta L_t(a)}$. Start with an upper bound.

$$\begin{aligned}
\frac{W_t}{W_{t-1}} &= \frac{\sum_a e^{-\eta L_t(a)}}{\sum_a e^{-\eta L_{t-1}(a)}} && (1)\\
&= \sum_a e^{-\eta X_t^a} \frac{e^{-\eta L_{t-1}(a)}}{\sum_{a'} e^{-\eta L_{t-1}(a')}} && (2)\\
&= \sum_a e^{-\eta X_t^a} p_t(a) && (3)\\
&\leq \sum_a (1 - \eta X_t^a p_t(a) + \eta^2 (X_a^t)^2) p_t(a) && (4)\\
&\leq e^{-\eta X_t^a p_t(a) + \eta^2 (X_a^t)^2 p_t(a)} && (5)
\end{aligned}$$

in (3), we used the inequality $e^x \leq 1 + x + x^2$, in (4) we used the inequality $1 + x \leq e^x$. Thus, we can get

$$\textstyle\sum_{t=1}^{T} \sum_a p_t(a) \hat{l}_t^a - \min \hat{L}_T(a) \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_a p_t(a) (\hat{l}_t^a)^2$$

By taking expectation,

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T} \sum_a p_t(a) \hat{l}_t^a] - \mathbb{E}[\min \hat{L}_T(a)] \leq \frac{\ln K}{\eta} + \eta \mathbb{E}[\sum_{t=1}^{T} \sum_a p_t(a) (\hat{l}_t^a)^2]\\
\Longrightarrow &\mathbb{E}[\sum_{t=1}^{T} \sum_a p_t(a) \hat{l}_t^a] - \min \mathbb{E}[\hat{L}_T(a)] \leq \frac{\ln K}{\eta} + \eta \mathbb{E}[\sum_{t=1}^{T} \sum_a p_t(a) (\hat{l}_t^a)^2]
\end{aligned}$$

We consider the expectation terms,

$$\textstyle\mathbb{E}[\sum_{t=1}^{T} \sum_a p_t(a) \hat{l}_t^a] = \mathbb{E}[\sum_{t=1}^{T} \sum_a \mathbb{E}[p_t(a) \hat{l}_t^a | A_1, A_2, ..., A_{t-1}]] = \mathbb{E}[\sum_{t=1}^{T} \sum_a p_t(a) l_t^a]$$

which is the expected loss of EXP3

$$\mathbb{E}[\hat{L}_T)(a)] = \mathbb{E}[\textstyle\sum \hat{l}_t^a] = \sum l_t^a$$

which is the cummulative loss of time T.

$$\textstyle\mathbb{E}[\sum_{t=1}^{T} \sum_a p_t(a) (\hat{l}_t^a)^2] = \mathbb{E}[\sum (\hat{l}_t^a)^2] \leq KT$$

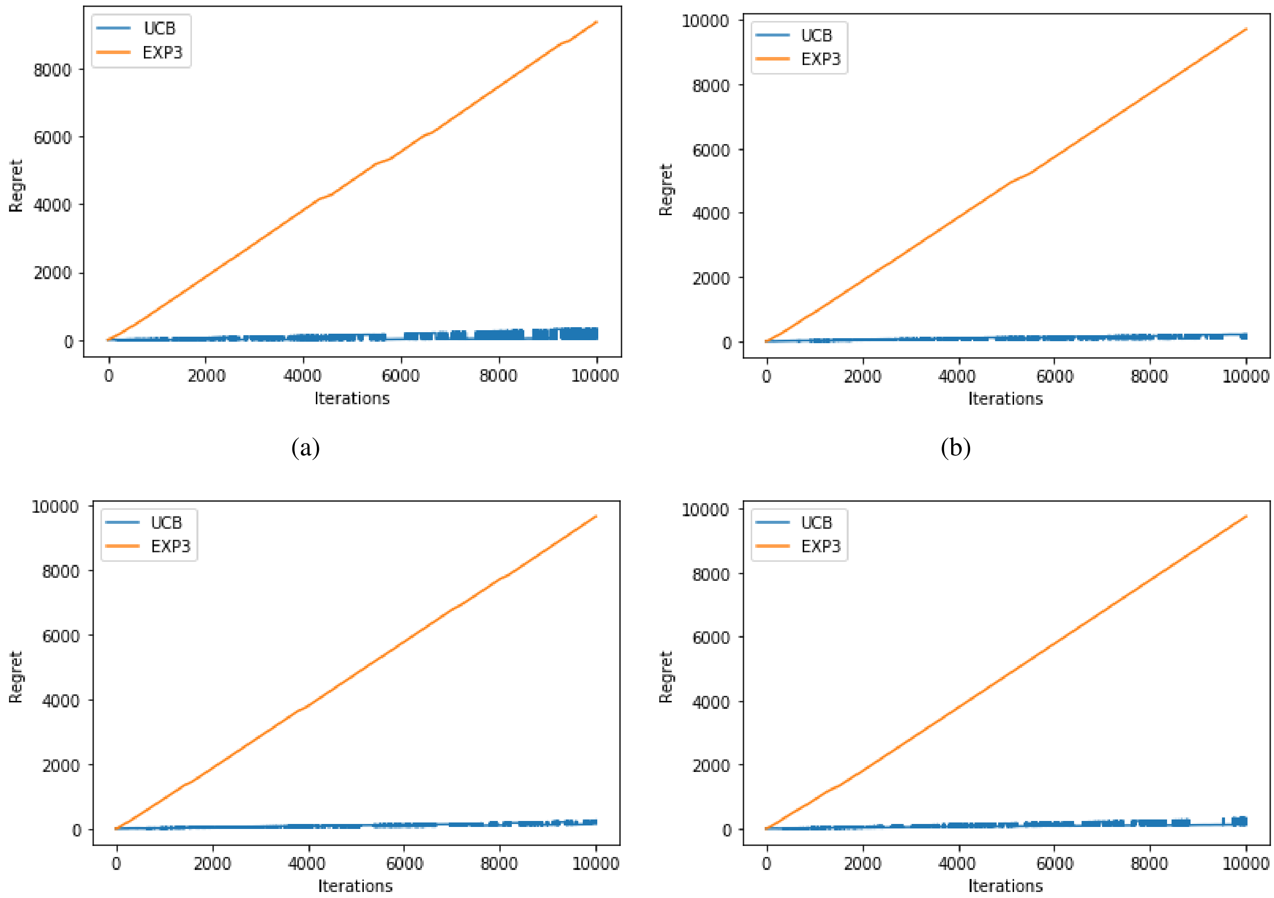
Hence, we can get

$$\mathbb{E}[R_T] \leq 2\sqrt{tK \ln K}$$

with $\eta = \sqrt{\frac{\ln K}{tK}}$, $\mathbb{E}[R_T] \leq 2\sqrt{tK \ln(K)}$

1.2

(a)The best arm is 14 and the worst is 16.
(b)The best and two worst arms is 14, 2, 16.
(c)The best and three worst arms is 14, 13, 2, 16.
(d)The best, the median, and the worst arm is 14, 1, 16.



(a)

(b)

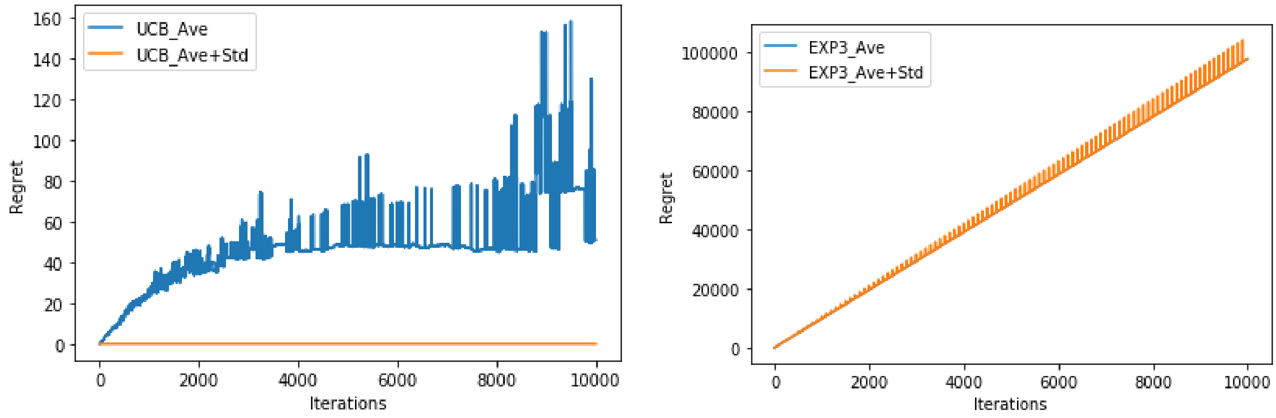
(c)

(d)

Figure 1: **Regret with different arms**

(But I think the plots are not correct, the line shouldn't be linear.)

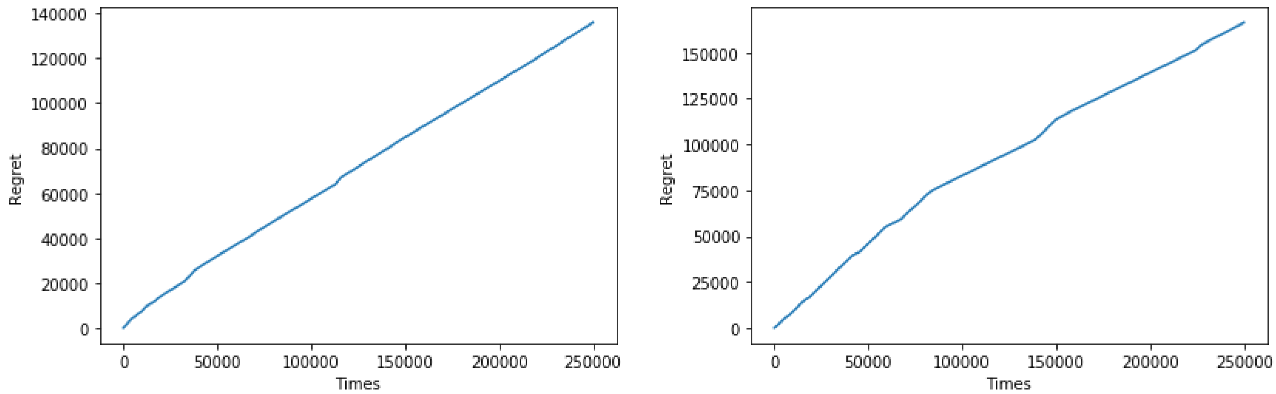
1.3



(a) UCB1 (b) EXP3

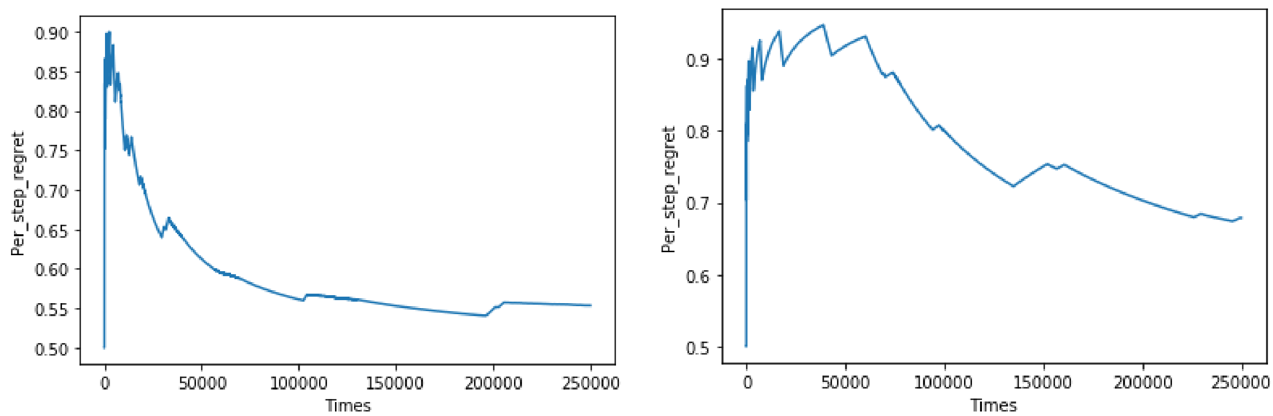
Figure 2: **Regret of two algorithms**

2 Empirical Evaluation of UCRL2 (25 points)



(a) L=6 (b) L=15

Figure 3: **Regret with different states' numbers**



(a) L=6 (b) L=15

Figure 4: **Pre_step_Regret with different states' numbers**

With more states, the accumulated regret become more. Also each step's regret become bigger. L=15 take much more times to decrease the average regrets.

3 Computing Diameter (15 points)

3.1

For all optimal action a^*, let $p(s|s_0, a) = \delta$, whereas $p(s|s_0, a^*) = \delta + \epsilon$ for $\epsilon \in (0, \delta)$. Further, let $p(s_0|s, a) = \delta$ for all a. The diameter is $\frac{1}{\delta}$.