Homework Assignment 5

Zhaoyang Xu

October 6, 2020

Contents

1 Deep Q-Learning (60 points)

1.1

I used function $tf.reduce\_mean$ in python to calculate the loss.

```
self.q_state_action = tf.reduce_sum(tf.multiply(self.q_state, action_one_hot),
                      axis=1)
self.loss = tf.reduce_mean(tf.square(self.q_state_action - self.q_target_in))
```

We know that the error is given by

$$L = \frac{1}{B}\sum_{i=1}^{B}([\hat{Q}(s_i)]_{a_i} - y_i)^2$$

From question, the batch size $B = 100$, it means we take 100 samples from the experience memory. The size of target is $[100, 1]$.
But the size of Q is $[100, |A|]$. We want these 2 terms in same size. By taking action in one-hot type, only the Q where take action is ratained. Then by $tf.reduce\_sum$ we can transfer $\hat{Q}(s_i)$ size to $[100, 1]$. Thus we can calculate the loss.

1.2

In Deep Q-Learning, the input of the neural network are states and the output are actions. From question, we know that the input are 4, the output are 2. And the neural network has two hidden layers with 64 neurons each.
Now, we calculate the weights,

$$Input \times Hide_1 + Hide_1 \times Hide_2 + Hide_2 \times Output$$
$$4 \times 64 + 64 \times 64 + 64 \times 2 = 4480$$

Then we calculate bias,

$$Hide_2 + Hide_2 + Output$$
$$64 + 64 + 2 = 130$$

Hence the number of trainable parameters is weights + bias = 4610.

1.3

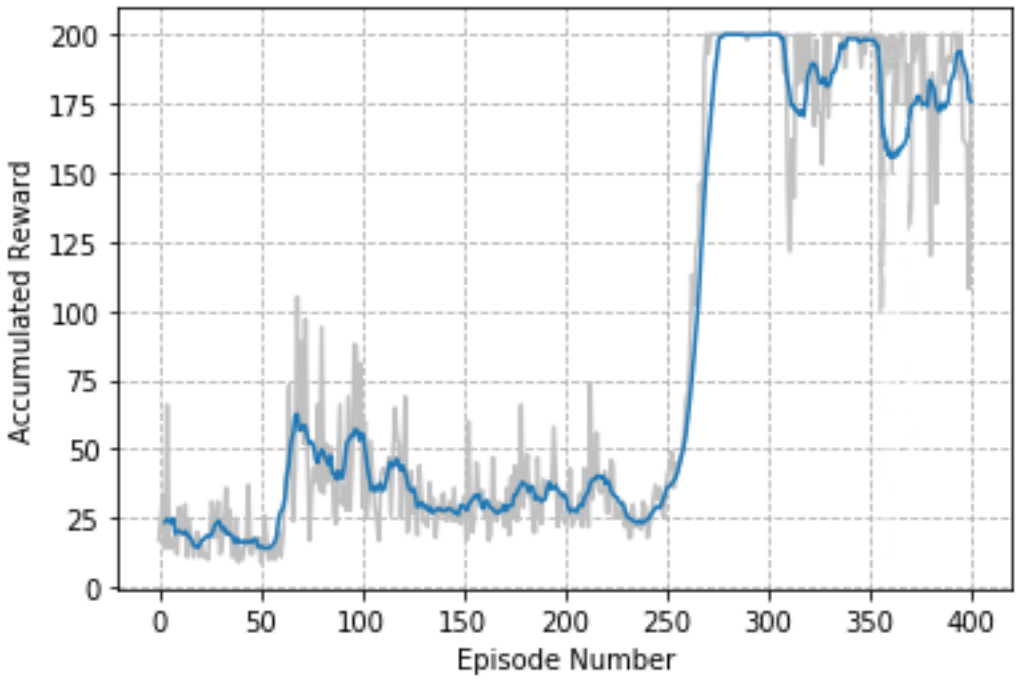


Figure 1: Example of a single training trial. The blue line shows a moving average.

2 Tighter analysis of the Hedge algorithm (15 points)

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.

$$\frac{W_t}{W_{t-1}} = \frac{\sum_a e^{-\eta L_t(a)}}{\sum_a e^{-\eta L_{t-1}(a)}} \tag{1}$$

$$= \frac{\sum_a e^{-\eta X_t^a} e^{-\eta L_{t-1}(a)}}{\sum_a e^{-\eta L_{t-1}(a)}} \tag{2}$$

$$= \sum_a e^{-\eta X_t^a} \frac{e^{-\eta L_{t-1}(a)}}{\sum_{a'} e^{-\eta L_{t-1}(a')}} \tag{3}$$

$$= \sum_a e^{-\eta X_t^a} p_t(a) \tag{4}$$

$$= \mathbb{E}[e^{-\eta X_t^a}] \tag{5}$$

$$\leq e^{-\eta \mathbb{E}[X_t^a] + \frac{\eta^2}{8}} \tag{6}$$

where in (4), $\sum_a p_t^a = 1$ and the equation can be written in terms of expectation. In (5) we use Hoeffding lemma.

Now, we consider the ratio $\frac{W_t}{W_0}$,

$$\begin{aligned}\frac{W_T}{W_0} &= \frac{W_1}{W_0} \times \frac{W_2}{W_1} \times \cdots \times \frac{W_T}{W_{T-1}} \\ &\leq e^{-\eta \sum_{t=1}^{T} \mathbb{E}[X_t^a] + \frac{\eta^2}{8} T} \\ &\leq e^{-\eta \mathbb{E}[\sum_{t=1}^{T} X_t^a] + \frac{\eta^2}{8} T} \\ &= e^{-\eta \mathbb{E}[L_T(a)] + \frac{\eta^2}{8} T}\end{aligned}$$

where the second line used Jensen's inequality.

On the other hand:

$$\frac{W_t}{W_0} = \frac{\sum_a e^{-\eta L_T(a)}}{K} \geq \frac{\max_a e^{-\eta L_T(a)}}{K} = \frac{e^{-\eta \min_a L_T(a)}}{K}$$

where we get the lower bound. By taking the two inequalities togeter and applying logarithm we obtain:

$$-\eta \min_a L_T(a) - \ln K \leq -\eta \hat{L}_T + \frac{\eta^2}{8} T$$

By changing the sides and dividing by η we get:

$$\mathbb{E}[L_T(a)] - \min_a L_T(a) \leq \frac{\eta}{8} T + \frac{\ln K}{\eta}$$

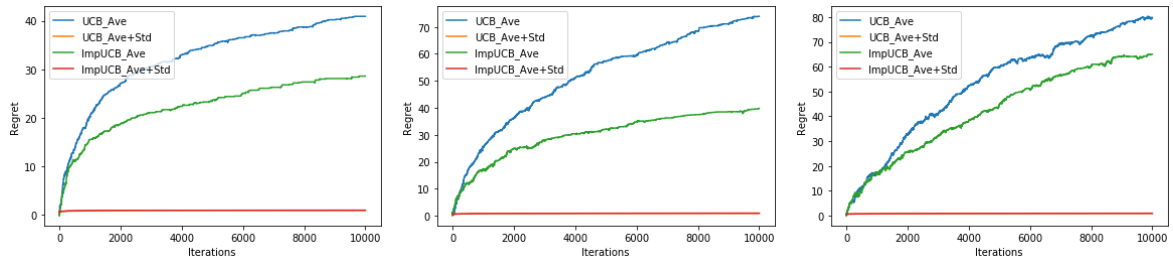
$\mathbb{E}[L_T(a)]$ is the expected cumulative loss of Hedge after T rounds. Thus, the left hand side of the inequality is the expected regret of Hedge. Hence we can get that

$$\mathbb{E}[R_T] \leq \frac{\eta}{8} T + \frac{\ln K}{\eta}$$

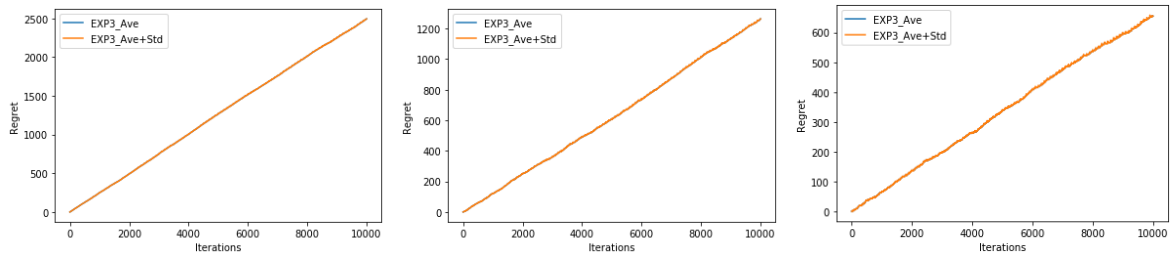
By taking the derivative of the right hand side and equating it to 0, we can get $\frac{T}{8} - \frac{\ln K}{\eta^2} = 0$ and thus $\eta = \sqrt{\frac{8 \ln K}{T}}$. The second derivative also can be proved positive.
We take $\eta = \sqrt{\frac{8 \ln K}{T}}$ back, then we obtain,

$$\mathbb{E}[R_T] \leq \sqrt{\frac{T \ln K}{8}} + \sqrt{\frac{T \ln K}{8}} = \sqrt{\frac{T \ln K}{2}}$$

3 Empirical comparison of UCB1 and EXP3 algorithms (25 points)

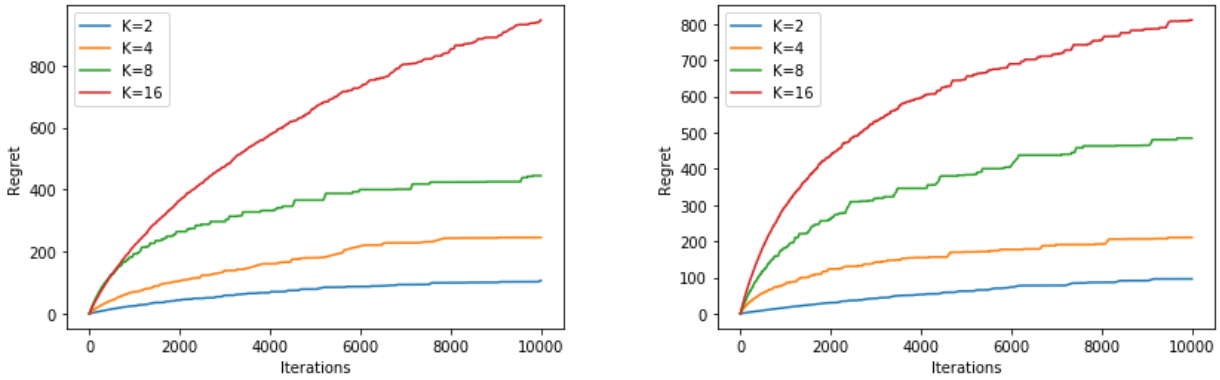


(a) $\mu = \frac{1}{2} - \frac{1}{4}$ (b) $\mu = \frac{1}{2} - \frac{1}{8}$ (c) $\mu = \frac{1}{2} - \frac{1}{16}$



(d) $\mu = \frac{1}{2} - \frac{1}{4}$ (e) $\mu = \frac{1}{2} - \frac{1}{8}$ (f) $\mu = \frac{1}{2} - \frac{1}{16}$

Figure 2: **Pseudo regret with different μ and $K = 2$**



(a) UCB (b) Improved UCB

Figure 3: **Pseudo regret with different K and $\mu = 0.25$**

It is obvious that with μ become greater, the regret become larger. Also, with more K, the regret become larger.