IE613: Online Learning - Assignment 1

Subhadeep Chaudhuri | 19i190010

Question 1 (Full Information Setting)

Consider the problem of prediction with expert advice with $d = 10$. Assume that the losses assigned to each expert are generated according to independent Bernoulli distributions. The adversary/environment generates loss for experts 1 to 8 according to $Ber(0.5)$ in each round. For the 9^{th} expert, loss is generated according to $Ber(0.5-\Delta)$ in each round. The losses for the 10^{th} expert are generated according to different Bernoulli random variable in each round - for the first $T/2$ rounds, they are generated according to $Ber(0.5+\Delta)$ and the remaining $T/2$ rounds they are generated according to Bernoulli random variable $Ber(0.5-2\Delta), \Delta = 0.1, T = 10^5$. Generate (pseudo) regret values for different learning rates (η) for Weighted Majority algorithm. The averages should be taken over at least 20 sample paths (more is better). Display 95% confidence intervals for each plot. Vary c in the interval [0.1, 2.1] in steps of size 0.2 to get different learning rates. Implement Weighted Majority algorithm with $\eta = c\sqrt{\frac{2log(d)}{T}}$

Answer: The required plot from the implementation of the Weighted Majority algorithm for the given parameters is as follows (we averaged it out over 100 runs):

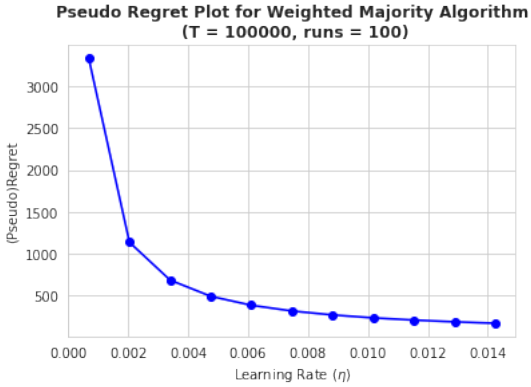

Figure 1: Plot showing (Pseudo)Regret for Weighted Majority algorithm

From the figure, we can observe the following:

- The average regret suffered decreases with increase in the learning rate η.
- The rate of change of regret keeps decreasing with increase in the learning rate η.
- The regrets suffered in each run are very close to one another, in spite of the losses being randomly generated from the underlying distribution. In turn, the variation in the regrets suffered over multiple runs is very low, and hence the 95% confidence intervals are not visible.

Question 2 (Bandit Setting)

Consider the problem of multi-armed bandit with $K = 10$ arms. Assume that the losses are generated as in Question 1. For each of the following algorithms generate (pseudo)regret for different learning rates (η) for each of the following algorithms. The averages should be taken over atleast 50 sample paths (more is better). Display 95% confidence intervals for each plot. Vary c in the interval [0.1, 2.1] in steps of size 0.2 to get different learning rates.

- EXP3: Set $\eta = c\sqrt{\frac{2log(k)}{kT}}$
- EXP3.P: Set $\eta = c\sqrt{\frac{2log(k)}{kT}}, \beta = \eta, \gamma = k\eta$
- EXP3-IX: Set $\eta = c\sqrt{\frac{2log(k)}{kT}}, \gamma = \eta/2$

Answer: The required plot from the implementation of the EXP3, EXP3.P and EXP3-IX algorithms for the given set of parameters are as follows (we averaged it out over 60 runs):

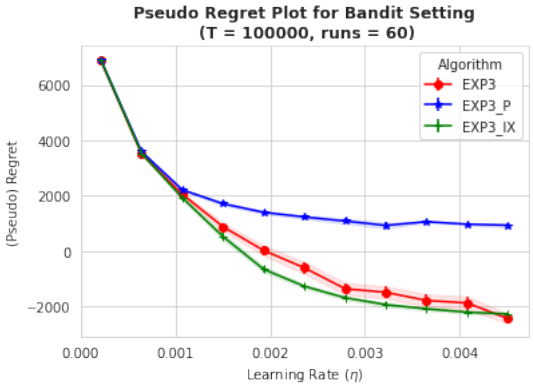

Figure 2: Plot showing (Pseudo)Regret for EXP3, EXP3.P and EXP3-IX algorithms

---

Question 3

In Question 2, which one of EXP3, EXP3.P and EXP3-IX performs better and why?

Answer: Clearly, EXP3-IX has the best performance with lower expected regret and lower deviation. The good performance of EXP3-IX can be attributed to the fact that it explores and detects the new best advisor after the change in odds at T/2.

The bad performance of EXP3.P is maybe due to high exploration rates leading to low exploitation of the current best adviser.
The negative regret is justified by the way of defining the pseudo-regret, which makes the algorithm work smarter than the oracle policy, which in this case is being considered as a fixed arm being pulled, but in reality is not the best possible policy.

Question 4

Show that for any deterministic policy π there exists an environment ν such that $R_T(\pi, \nu) \geq T(1 - 1/K)$ for T rounds and K arms.

Answer: The regret over t rounds for the environment ν corresponding to a policy π is defined as follows:

$$R_T(\pi, \nu) = \sum_{t=1}^{T} \max_{i \in [k]} x_{ti} - \sum_{t=1}^{T} x_{t,I_t}$$

with I_t being the arm chosen by the policy π in the t^{th} round, x_{tI_t} the reward obtained in round t, and each of the rewards being binary i.e. $x_{ti} \in \{0, 1\}\ \forall i \in [K]$ In the adversarial setting, the environment tries to maximise the regret of the learner. Now, given that the policy π is deterministic, the environment already knows the action of the learner in the t^{th} round i.e. I_t. Thus, the environment would set the reward for the different arms as follows:

$$x_{ti} = \begin{cases} 0, & \text{if } i = I_t \\ 1, & \text{otherwise} \end{cases}$$

For such an environment (let's call it ν), the regret over t rounds for the deterministic policy π would then be as follows:

$$\begin{aligned}
R_T(\pi, \nu) &= \sum_{t=1}^{T} \max_{i \in [k]} x_{ti} - \sum_{t=1}^{T} x_{t,I_t} \\
&\geq \max_{i \in [k]} \sum_{t=1}^{T} x_{ti} - \sum_{t=1}^{T} x_{t,I_t} \\
&= \max_{i \in [k]} \sum_{t=1}^{T} x_{ti} \quad \text{[the policy collects zero reward i.e. } \sum_{t=1}^{T} x_{t,I_t} = 0\text{]} \\
&\geq \mathbb{E}\Big(\sum_{t=1}^{T} x_{ti}\Big) \quad \text{[since } max(X) \geq \mathbb{E}(X) \text{ for any random variable } X\text{]} \\
&= \sum_{t=1}^{T} \mathbb{E}(x_{ti}) \quad \text{[Sum Law of Expectation.]} \\
&= \sum_{t=1}^{T} \frac{1}{K} \sum_{i=1}^{K} x_{ti} \\
&= T\Big(\frac{K-1}{K}\Big) \quad \text{[Numerator has } K-1 \text{ since } x_{ti} = 0 \text{ when } i = I_t\text{]} \\
&= T\Big(1 - \frac{1}{K}\Big)
\end{aligned}$$

Thus, we get the required lower bound on regret as follows:

$$R_T(\pi, \nu) \geq T(1 - 1/K) \qquad \text{(Proved)}$$

Question 5

Suppose we had defined the regret by

$$R_T^{track}(\pi, \nu) = \mathbb{E}\big[\sum_{t=1}^{T} \max_{i \in [k]} x_{ti} - \sum_{t=1}^{T} x_{t,I_t}\big]$$

where I_t is the arm chosen by the policy π and x_{tI_t} is the reward observed in the round t. At first sight this definition seems like the right thing because it measures what you actually care about. Unfortunately, however, it gives the adversary too much power. Show that for any policy π (randomized or not) there exists a $\nu \in [0, 1]^{K \times T}$ such that:

$$R_T^{track}(\pi, \nu) \geq T(1 - \frac{1}{K})$$

Answer: For a deterministic policy π, we have already shown the existence of a ν such that the given inequality holds. Let us now show the validity of the inequality for a non-deterministic setting.
We define the (expected) regret as follows: $R_T^{track}(\pi,\nu) = \mathbb{E}\big[\sum_{t=1}^{T} \max_{i\in[k]} x_{ti} - \sum_{t=1}^{T} x_{t,I_t}\big]$

The learner tries to decide on the selection of arms based on its past actions chosen and their corresponding rewards. If an arm gives high reward over multiple rounds, the likelihood of that arm being pulled again is higher. Let us denote these likelihoods by π_{ti}, for arm i in round t. So, the environment being adversarial would try to assign the reward to that arm which performs worst i.e. which has the lowest likelihood π_{ti}. Formally,

$$x_{ti} = \begin{cases} 1, & \text{if } i = \arg\min_i \pi_{ti} \\ 0, & \text{otherwise} \end{cases}$$

We can write the regret as follows:

$$\begin{aligned} R_T^{track}(\pi,\nu) &= \mathbb{E}\big[\sum_{t=1}^{T} \max_{i\in[k]} x_{ti} - \sum_{t=1}^{T} x_{t,I_t}\big] \\ &= \mathbb{E}\big[T - \sum_{t=1}^{T} x_{t,I_t}\big] \\ &= T - \sum_{t=1}^{T} \mathbb{E}[x_{t,I_t}] \quad \text{[the maximum reward obtainable from an arm in a round is 1.]} \end{aligned}$$

Now, we can write the second term as follows:

$$\begin{aligned} \mathbb{E}[x_{t,I_t}] &= 1 \times \Pr(i = I_t = \arg\min_i \pi_{ti}) + 0 \times \Pr(I_t \neq \arg\min_i \pi_{ti}) \\ &= \Pr(I_t = \arg\min_i \pi_{ti}) \end{aligned}$$

Since the learner pulls an arm based on its history of actions and corresponding rewards, it chooses the arm with maximum likelihood. Therefore, as time progresses, the learner would not pull the sub-optimal arms, implying that the initial uniform probability $\frac{1}{K}$ of selection for each arm gets changed and approaches degeneracy, where all the probability mass would be assigned to the arm with the maximum likelihood after a fixed number of rounds.
Hence, $\Pr(I_t = \arg\min_i \pi_{ti}) <= \frac{1}{K}$. Using this inequality, we get:

$$\begin{aligned} R_T^{track}(\pi,\nu) &= T - \sum_{t=1}^{T} \mathbb{E}[x_{t,I_t}] \\ &\geq T - \sum_{t=1}^{T} \frac{1}{K} \\ &\geq T - \frac{T}{K} = T\Big(1 - \frac{1}{K}\Big) \end{aligned}$$

Hence, we prove the validity of the inequality.

Question 6

Let $p \in P_k$ be a probability vector and suppose $\hat{X} : [k] \times \mathbb{R} \to \mathbb{R}$ is a function such that for all $x \in \mathbb{R}^k$, if $\mathbf{A} \sim p$,

$$\mathbb{E}[\hat{X}(A, x_A)] = \sum_{i=1}^{k} p_i \hat{X}(i, x_i) = x_1$$

Show there exists an $a \in \mathbb{R}^k$ such that $\sum_{j=1}^{k} a_j p_j = 0$ and $\hat{X}(i,x) = a_i + \frac{\mathbb{I}\{i=1\}x_1}{p_1}$

Answer: Let us assume $a_i = \hat{X}(i, x_i) - \frac{\mathbb{I}\{i=1\}x_1}{p_1}\ \forall i \in [k]$. We can then simplify $\sum_{j=1}^{k} a_j p_j$ as follows:

$$\begin{aligned}
\sum_{j=1}^{k} a_j p_j &= \sum_{j=1}^{k} p_j \left(\hat{X}(j, x_j) - \frac{\mathbb{I}\{j=1\}x_1}{p_1} \right) \\
&= \sum_{j=1}^{k} p_j \hat{X}(j, x_j) - \sum_{j=1}^{k} p_j \frac{\mathbb{I}\{j=1\}x_1}{p_1} \\
&= x_1 - x_1 \quad [\text{since we are given} \sum_{j=1}^{k} p_j \hat{X}(j, x_j) = x_1] \\
&= 0
\end{aligned}$$

Also, from the definition of a_i, we see that the following condition holds:

$$\hat{X}(i, x_i) = a_i + \frac{\mathbb{I}\{i=1\}x_1}{p_1}$$

Thus, we are able to show the existence of $a \in \mathbb{R}^k$ which satisfied the required conditions, completing the proof.

Question 7

Suppose we have a two-armed stochastic Bernoulli bandit with $\mu_1 = 0.5$ and $\mu_2 = 0.55$. Test your implementation of EXP3 from the Question 2. What happens when $T = 10^5$ and the sequence of rewards is $x_{t1} = \mathbb{I}\{t \leq T/4\}$ and $x_{t2} = \mathbb{I}\{t > T/4\}$?

Answer: The plot of (pseudo)regret obtained from the stochastic setting and the deterministic setting are being given below:

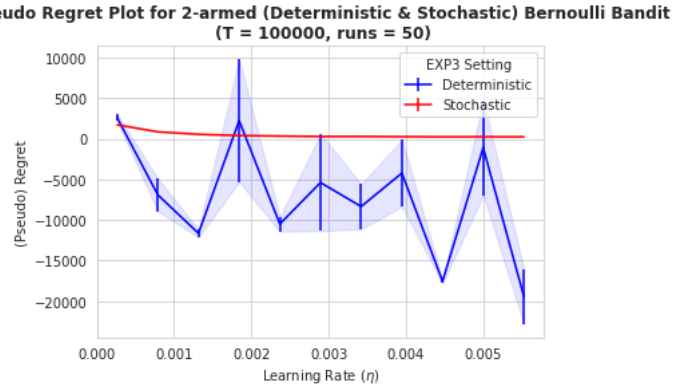

Figure 3: (Pseudo)Regret for EXP3 algorithm under stochastic and deterministic settings

The confidence interval for the stochastic setting are not clearly visible because of their small variation between different runs, whereas the deterministic setting shows considerably large variation over runs.

We also conclude that the observed regrets are in line with the intuition that the deterministic setting should generally yield lower regret in comparison to the stochastic setting.
