CSC 696H: Topics in Bandits and Reinforcement Learning Theory **Fall 2024**

Lecture 5: Multi-Armed Bandits

Lecturer: Chicheng Zhang *Scribe: Aryan Pathare*

1 Recap: Exponential Weight Algorithm [EWA]

The general steps in the Exponential Weight Algorithm are -

1. Initialize $w_1(f) = 1$ for all $f \in \mathcal{F}$
2. For $t = 1, 2, \ldots, T$, compute the distribution -

$$q_t(f) = \frac{w_t(f)}{W_t}, \quad \text{where} \quad W_t = \sum_{f \in \mathcal{F}} w_t(f),$$

3. Compute the predictor

$$\hat{f}_t(x) = \sum_{f \in \mathcal{F}} q_t(f) f(x).$$

4. Observe (x_t, y_t) and incur loss $\ell(\hat{f}_t(x_t), y_t)$
5. Update weights:

$$w_{t+1}(f) = w_t(f) \cdot e^{-\eta \ell(f(x_t), y_t)}.$$

In EWA, over time, the weights of the models with good performance increase exponentially.

For absolute loss, the regret of EWA is bounded by Reg(F,T) $\leq O\left(\sqrt{T \ln |\mathcal{F}|}\right)$.

For α-exp-concave loss, the regret of EWA is bounded by Reg(F,T) $\leq O\left(\frac{1}{\alpha} \ln |\mathcal{F}|\right)$

2 Recap: KL-Divergence

The Kullback-Leibler divergence is used to measure the difference between two probability distributions.

For distributions p and q, the KL-Divergence is defined as -

$K(p(x), q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$

Some important properties of KL-Divergence are -

1. $K(p(x), q(x)) \geq 0$

2. If $K(p(x), q(x)) = 0$, then the distributions p and q are the same/

3. $K(p(x), q(x)) \neq K(q(x), p(x))$ i.e KL-Divergence is asymmetric.

3 Introduction

This lecture introduces the multi-armed bandits problem. It sheds light on the setup of the problem and also explains three algorithms for solving the problem.

4 Online Learning Remarks

1. V Vovk authored the paper "A Game of Prediction with Expert Advice" on the unified treatment of online learning with general losses.

2. Consider online learning with square loss $l(\hat{y}, y) = (\hat{y} - y)^2$

   with the additional assumption that there exists a $f^* \in f$ such that $y_t = f^*(x_t) + \epsilon_t$ for all t and zero-mean independent noise ϵ_t. This is called realizable regression.

   The EWA guarantee is -

   $\sum_{t=1}^{T}(\hat{f}_t(x_t) - y_t)^2 - (f^*(x_t) - y_t)^2 \leq c.ln|F|$.

   We can make the guarantee more interpretable as follows:

   Taking expectation on both sides the t-th term is,

   $\mathrm{E}[reg_t] = \mathrm{E}[(\hat{f}_t(x_t) - f^*(x_t) - \epsilon_t)^2 - \epsilon_t^2] = \mathrm{E}[(\hat{f}_t(x_t) - f^*(x_t))^2 - 2(\hat{f}_t(x_t) - f^*(x_t)).\epsilon_t]$

   Denote by $\hat{f}_t(x_t) - f^*(x_t) =: z_t$. Therefore, as ϵ_t is zero-mean independent noise, $z_t \perp\!\!\!\perp \epsilon_t$. Therefore, $\mathrm{E}[z_t.\epsilon_t] = \mathrm{E}[z_t].\mathrm{E}[\epsilon_t] = 0{*}0 = 0$.

   Based on the above, $\mathrm{E}[\sum_{t=1}^{T} reg_t] = \mathrm{E}[\sum_{t=1}^{T}(\hat{f}_t(x_t) - f^*(x_t))^2] \leq c.ln|F|$

   Thus, this equation signifies that in the long run, even though we do not ever observe the condition mean $f^*(x_t)$ (since they are always corrupted by noise), the predictions estimate the conditional means well enough.

   The behaviour of f^* can be observed in figure 1

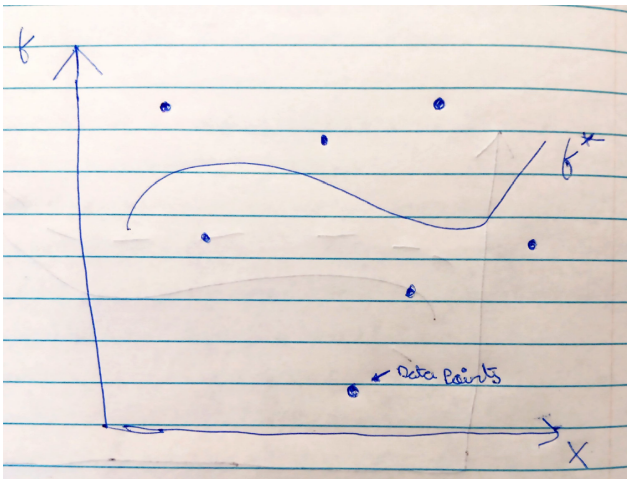

Figure 1: Behaviour of f^*

5 Multi-Armed Bandits [MAB]

In the MAB model, the agent performs an action and based on that receives evaluative feedback. This is in constrast to instructive feedback for supervised learning.

The setup for MAB is as follows -

- Action set $A = \{1, 2, ...A\}$. The action set is also called the arm set.
- For $t = 1, 2, ..., T$,

  Agent takes action a_t and based on this receives the reward $r_t = f^*(a_t) + \Sigma_t \epsilon_t$. Here, $f^*(a_t)$ and the distribution of the zero-mean random noise $\Sigma_t \epsilon_t$ are unknown. We know, though, that ϵ_t is 1-subgaussian
- The agent's performance is evaluated based on the expected total reward i.e $\mathrm{E}[\sum_{t=1}^{T} r_t] = \mathrm{E}[\sum_{t=1}^{T} f^*(a_t)]$

Throughout, we make the assumption that the expected reward function f^* is such that $f^*(a) \in [0, 1], \forall a$.

For example, consider a 2-armed setup as in figure 2. In this case, a good strategy would be to always use slot machine 1 as it has a higher expected reward.

In an ideal world, if f^* was known to the agent, the optimal strategy would be to always take the action $a^* = \mathrm{argmax}_{a \in A} f^*(a)$. The corresponding optimal reward would be $T.f^*(a)$.

In MAB, the performance measure used is regret. In this context, it is given by -
$\mathrm{Reg}(T) = T.f^*(a) - \mathrm{E}[\sum_{t=1}^{T} f^*(a_t)] = \mathrm{E}[\sum_{t=1}^{T} (f^*(a^*) - f^*(a_t))]$.

To design a sublinear regret strategy, we need to do the following -

Figure 2: 2-armed bandit example

1. Need to learn f^* by taking all possible actions. This is known as exploration.

2. Need to take action a_t that we believe to be good i.e has a large f^* value. This is known as exploitation.

Now we will look at some of the algorithms for MAB

5.1 Algorithm 1: Explore then Exploit

This algorithm proceeds in 2 phases. In phase 1, we use the 1st T_o rounds to estimate f^* by taking all actions in a round-robin manner and estimating $\hat{f} = (f(1)....f(A))$.

Phase 2 starts from $(T_o + 1)$ round onwards. In this phase, we only take the action $\hat{a} = \text{argmax}_{a \in A}\, \hat{f}(a)$.

Thus, in phase 1 we perform exploration by performing all possible actions to try and estimate their rewards. In Phase 2, based on the results (reward estimates) obtained from phase 1, we "exploit" the high reward actions.

We can now analyze the regret for each of the phases and thereby the total regret.

5.1.1 Phase 1 regret

Phase 1 lasts for T_o rounds. We know, $Reg(T) = E[\Sigma_{t=1}^{T}(f^*(a^*) - f^*(a_t))]$. Therefore, for T_o rounds, Phase 1 regret $\leq T_o$.

5.1.2 Phase 2 regret

In phase 2, by Hoeffding's inequality, $max_a|\hat{f}(a) - f^*(a)| \leq \sqrt{\frac{A}{T_o}}$.

Thus, based on ERM analysis, $f^*(a) - f^*(\hat{a}) \leq 2.\sqrt{\frac{A}{T_o}}$.

Since, phase 2 has $(T - T_o)$ rounds, Phase 2 regret $\leq (T - T_o).2.\sqrt{\frac{A}{T_o}}$.

Figure 3 provides a visualization of the phase 2 regret and its bound.

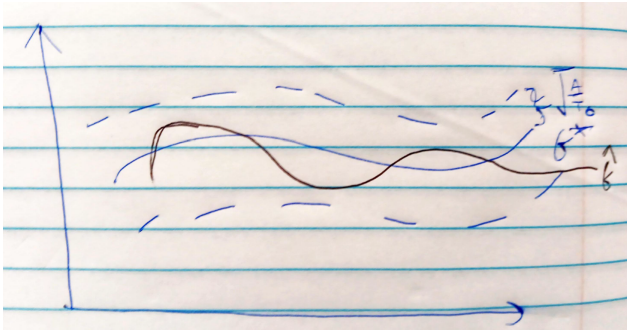

Figure 3: Phase 2 regret visualization

5.1.3 Total regret

Combining the regrets from both the phases, Total regret is Reg(T) $\leq T_o + (T - T_o).2.\sqrt{\frac{A}{T_o}}$.

Further simplifying, Reg(T) $\leq T_o + T.2.\sqrt{\frac{A}{T_o}}$.

The optimal value of T_o is $A^{\frac{1}{3}}.T^{\frac{2}{3}}$ and the corresponding optimal value of regret is Reg(T) $= O(A^{\frac{1}{3}}.T^{\frac{2}{3}})$.

5.2 Algorithm 2: ϵ-Greedy

The ϵ-Greedy algorithm intersperses exploration and exploitation. The algorithm is as follows -

1. For t=1,2,..T, flip a coin $z \sim Bernouli(\epsilon)$. Thus, $z = 1$ with probability ϵ.
2. If $z = 1$, take action $a_t \sim Unif(1, ..., A)$
3. If $z = 0$, take action $a_t = \hat{a_t} = argmax_{a \in A} \hat{f}_t(a)$.

Thus, with probablity ϵ we perform exploration by picking an action uniformly randomly. Conversly, with probability $(1 - \epsilon)$, we "exploit" the action estimated to have the highest reward.

In step 3 of the algorithm, $\hat{f}_t(a)$ is calculated as $\hat{f}_t(a) = \frac{Total\ reward\ of\ a\ at\ (t-1)}{\#\ times\ a\ is\ taken\ upto\ (t-1)}$.

To define this mathematically, we introduce some notations that will be useful for later part of the class. Let I be an indicator variable such that $I(a_j = a) = 1$ if $a_j = a$ and 0 otherwise. Thus, $N_{t-1}(a) = \sum_{i=1}^{t-1} I(a_i = a)$ i.e the variable N_{t-1} indicates how many times an action a is taken in the first $(t-1)$ rounds.

$\therefore \hat{f}_t(a) = \frac{\sum_{i=1}^{t-1} I(a_i=a).r_i}{N_{t-1}(a)}$.

Thereby, Reg(T) $\leq \epsilon.T + \sqrt{\frac{A.T}{\epsilon}}$. Setting ϵ optimally, Reg(T) $= O(A^{\frac{1}{3}}.T^{\frac{2}{3}})$

5.3 Algorithm 3: Optimism Principle

The optimism principal is also know as optimism in face of uncertainty. In this algorithm, the agent acts according to the best pausible world. The best pausible world is also called the optimistic world model.

The effectiveness of the algorithm follows from the the following "win or win" argument -

- If the optimistic world model is correct, all are choices will be optimal and thereby the agent will have no regret.
- If the model is wrong, the agent will learn new things and avoid making the same mistakes in the future.

In order to construct the optimistic world mode, we take samples of rewards for different actions. Based on these samples, we need to find the highest plausible $f^*(a)$. for every action a. We can do this with the help of the confidence interval for each $f^*(a)$.

For each $f^*(a)$, the confidence interval is $[\hat{f}_t(a) \pm b_t(a)]$, where $\hat{f}_t(a)$ is the sample mean for arm a and $b_t(a)$ is the corresponding confidence width i.e $\sqrt{\frac{1}{N_{t-1}(a)}}$. Using this, the best plausible world for $(f^*(1)...f^*(a))$ is defined as $(\hat{f}_t(1) + b_t(1), ..., \hat{f}_t(a) + b_t(a))$. This can be observed in figure 4.

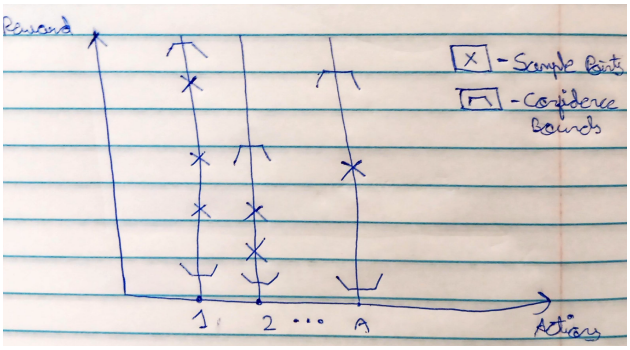

Figure 4: Illustration of the optimism principle

Using the best plausible world model, the Upper Confidence Bound Algorithm [UCB] is devised as -

- At time t, define $UCB_t(a) = \hat{f}_t(a) + b_t(a))$ such that $b_t(a) = \sqrt{\frac{l}{N_{t-1}(a)+1}}$ and $l = 8.ln(2AT)$. Here, the term b_t is the confidence width for arm a.
- Choose the action a_t such that $a_t = argmax_{a \in A} UCB_t(a)$.

We note that the algorithm chooses the arm with the highest UCB index, which is a summation over the historical mean reward $\hat{f}_t(a)$ and a confidence width $b_t(a)$. The role of $\hat{f}_t(a)$ is to promote exploitation:

choosing arms that performs well historically. On the other hand, $b_t(a)$ is called the explorative bonus for action a as it is the confidence interval i.e uncertainty in the possible reward from performing an action. Thus, it incentivizes exploring action by giving the hope of receiving a higher reward than the estimate. As $b_t(a)$ is inversely proportional to the number of times an action is taken, its value decreases with time as we perform the action. Thus, the UCB algorithm can balance between exploration and exploitation by using the b_t term to represent the potential of exploring new actions.

5.3.1 Validity of UCB

Validity of UCB is proven based on the following lemma -

There exists an event E such that $P(E) \geq 1 - \frac{2}{T}$ when E happens $\forall a, t \ |\hat{f}_t(a) - f^*(a)| \leq h_t(a) = \sqrt{\frac{l}{N_{t-1}(a)+1}}$. Thus, $UCB_t(a) \geq f^*(a)$.

An initial trial in proving the lemma is based on the idea $E_{t,a}|\hat{f}_t(a) - f^*(a)| \leq h_t(a) \forall t, a$.

$\therefore E = \cap_{t=1}^{T} \cap_{a=1}^{A} E_{t,a}$. It then suffices to show that each individual $P(E_{t,a})$ is large.

One might use Hoeffding's inequality to prove this, but the key difficulty is that $\hat{f}_t(a)$ is an average over $N_{t-1}(a)$ random variables; However, $N_{t-1}(a)$ is a random number, making Hoeffding's inequality not directly applicable.

This challenge is further addressed in the book "Introduction to Multi-Armed Bandits" by Aleksandrs Slivkins.

5.3.2 Regret analysis of UCB

The regret analysis of UCB is based on the theorem that UCB guarantees Reg $\leq \tilde{O}(\sqrt{AT})$. It can also be shown that UCB has instance-dependent guaranteed Reg $\leq \tilde{O}(ln(T.(\sum_{a \neq a^*} \frac{1}{f^*(a^*)+f^*(a)})))$. The instance-dependent guarantee of UCB is interesting because, unlike the other seen algorithms (i.e Explore then Exploit and ϵ-greedy), the regret of UCB depends on f^*.

The proof is as follows -

Let reg_t be the regret at time t. Thus, we know, Reg(T) = E$[\sum_{t=1}^{T} reg_t]$ = E$[\sum_{t=1}^{T} reg_t(I(E) + I(E^c))]$ = E$[(\sum_{t=1}^{T} reg_t).I(E) + (\sum_{t=1}^{T} reg_t).I(E^c)]$.

Consider E$[(\sum_{t=1}^{T} reg_t).I(E^c)] \leq T.E[I(E^c)] \leq T.\frac{2}{T} \leq 2$

In the next lecture we will come back to handle the more challenging $\mathbb{E}\left[(\sum_{t=1}^{T} reg_t).I(E)\right]$ term.