Bandit Algorithm (Online Machine Learning)

Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

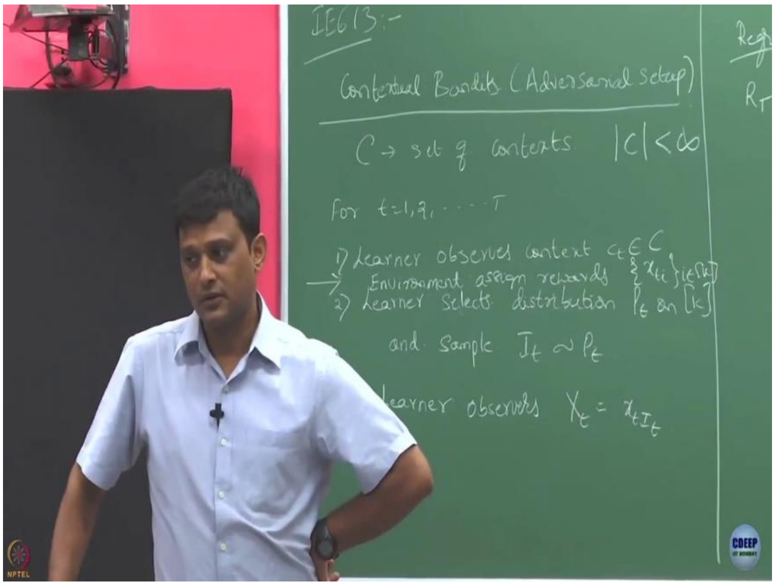
Lecture - 50
Adversarial Contextual Bandits – I

So, so far we have been studying stochastic contextual bandits. We just observed that this stochastic contextual bandits can be considered as a special case of stochastic linear bandits when we assume the main rewards are linear. In the last class, we discussed some algorithms and we just discussed the broad sketch of how the proof goes about and there are some assumptions and then, we said using some special techniques we should be able to relax those assumptions and prove the regret bonds whatever we claimed.

So, we will just left it there. Now, that was the part for the stochastic case and in the assignments you are going to look into some problems on that and also, you are going to look into some algorithms for this stochastic bandit case.

So, now we are going to move to the case when my setup can be adversarial. So, so far my rewards were all stochastic, but I could as well consider the case when my rewards are adversarial, right. So, how does my contextual bandits work for the adversarial case?

(Refer Slide Time: 01:50)



IE613:-

Contextual Bandits (Adversarial Setup)

$C \rightarrow$ set of contexts $|C| < \infty$

For $t = 1, 2, \ldots T$

1) Learner observes context $c_t \in C$

$\rightarrow$ Environment assign rewards $\{x_{ti}\}_{i \in [k]}$

2) Learner selects distribution P_t on $[k]$

and sample $I_t \sim P_t$

Learner observes $X_t = x_{tI_t}$

So, the setup is same as earlier except that when you have a context, the reward associated with that context from an arm need not be coming down from a fixed distribution. It could be chosen in an arbitrary fashion and in particular that could be selected by an adversary.

So, we will consider the following setup. I am going to denote C to be the set of context, for t equals to 1, 2, 3,… following happens. So, C is set of context is assumed to be fixed. It is some set from which the contexts are drawn in every round ok. So, the learner observe context from that. So, let us say there are k arms.

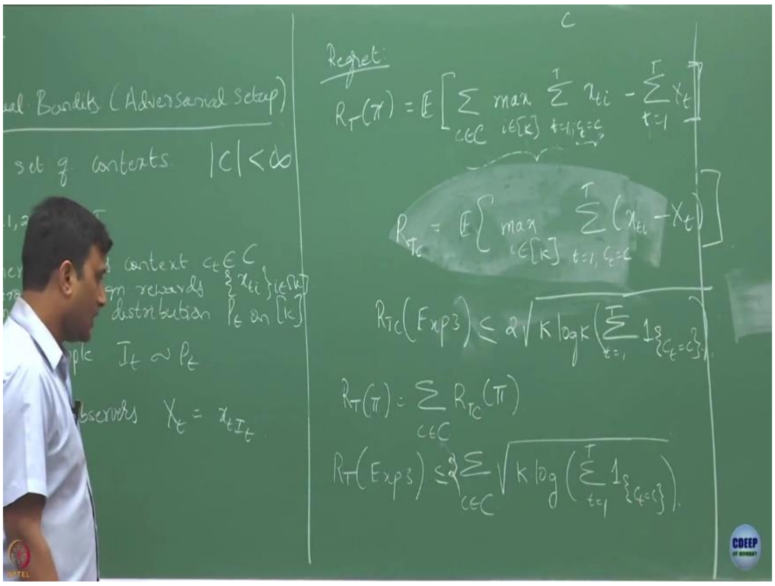
So, the learner observe observing this c_t context he is going to select a distribution on this k arms and let us say samples I_t from P_t, he is going to play an arm and for that he correspondingly learner observes X_t which is nothing, but X_{tIt}.

So, let us go through this. We are saying that in every round t learner observes the context and the environment assigns the reward vector in that round to each of my arms and then learner select a distribution P_t on this arms. So, what does learner, how does he select P_t?

He observes the context in that round t and based on that he is going to come up with a distribution, but he does not see the rewards assigned by the environment in that round and then whatever he going to sample an arm according to the next distribution and he is going to get a reward that he actually played in that round. So, X_{tIt} is the reward that is assigned to arm it in round t and this is what his reward in that round t, ok.

Now, so this is how the interaction between the learner and the environment is happening here. Now, our goal is as usual to compare the performance of a learner against the best reward I would have got if I knew all the reward assignments.

(Refer Slide Time: 06:54)


al Bandits (Adversarial Setup)
set of contexts $|C| < \infty$
1,2
context $c_t \in C$
rewards $\{x_{ti}\}_{i \in [k]}$
distribution P_t on $[k]$
ple $I_t \sim P_t$
bservers $X_t = x_{tI_t}$
Regret:
$R_T(\pi) = \mathbb{E}\left[\sum_{c \in C} \max_{i \in [k]} \sum_{t=1, c_t=c}^{T} x_{ti} - \sum_{t=1}^{T} X_t\right]$
$R_{T_c} = \mathbb{E}\left[\max_{i \in [k]} \sum_{t=1, c_t=c}^{T} (x_{ti} - X_t)\right]$
$R_{T_c}(Exp3) \leq 2\sqrt{k \log k \left(\sum_{t=1}^{T} 1_{\{c_t=c\}}\right)}$
$R_T(\pi) = \sum_{c \in C} R_{T_c}(\pi)$
$R_T(Exp3) \leq 2\sum_{c \in C} \sqrt{k \log \left(\sum_{t=1}^{T} 1_{\{c_t=c\}}\right)}$

So, then we are going to define regret or the expected regret as of a policy whatever policy π we have expected value of now notice that. So, ok; what we are doing here? What is this quantity here? This quantity says that suppose you fix a context. So, for time being assume that this is finite.

Now, take a context and look which is the arm that gives the maximum reward for that context, and then sum it our all possible context, right. So, notice that when I am saying this I am only looking at all counting, adding the rewards for which the contextual round t is c. So, this is for a particular c. This is giving me the total reward accumulated when I saw particular context c and now this is the total reward I would have accumulated over t periods.

And now this I am comparing against the reward I would have collected by playing my policy π. So, how does this π effect this X_t? So, according to that policy π that policy π is going to tell how I am going to choose my distribution in every round and accordingly I would have drawn some arm, I apply that arm and for that particular arm I am going to receive the corresponding reward.

So, all these X_t's are governed by my policy π and that is why this is the total reward I would have accumulated over a time period t and this is what I would have gotten. So, here we are looking. So, this is what we would have got. Now, what we are comparing this against is a benchmark; this is my benchmark in this I am saying I am looking for the best reward I would have got for each possible context, right. So, I am going to take a context and I am going to see what is the best reward I would have got here. Yeah?

Student: There is the summation over c on which sum it the context (Refer Time: 10:19).

No, I do not care it. I why should I have write it about like you whatever you do, you are saying it c t, it is maybe up to you want to use it or ignore it.

Student: You has the (Refer Time: 10:39) so.

Let it be in whatever way it is. I am just want to this is my total reward I would have got and this is what my cumulative reward is.

Student: But the total reward is (Refer Time: 10:56) context in that case the (Refer Time: 10:59).

Where do I am saying it is for single context? It is you take c, but you only look those rewards in which the context is c out of like this 100 that particular c may have occurred only 10 times only. For those 10 times, you would have now looking for which is the best arm you would have like to pick. We have a vector assigned for each arm.

If this is the total reward I am going to get over time period t and here I am looking at for every possible context, what is the best reward I would have gotten. Is this fine or if this ok fine.

Now, suppose there is only one context. So, there is only one context means this summation there is no, this summation I do not need to worry about, right in every time c t equals to c. Is this the standard adversarial setup I have?

Student: Yes.

So, there I am looking at the best single best arm against what I would have got, but now that I have different possible context now I am looking at for every context what is the best I would have obtained.

So, now, so what it is saying is basically see like I as of now I know that like my optimal arm could depend on my context whenever that context I observe, I want to see over the duration wherever I have observed this context what is the best action I would have taken in hindsight.

So, what is this part doing? Wherever I saw context c, it is give me the total reward you accumulated whenever I you saw that c. Now, max over i_k is telling that in hindsight if you know all those X_{ti}'s for that corresponding c, what is the best action you should have played and what is the best corresponding reward you have gotten? And now this is for all possible contexts and now you are trying to compare it with what is the cumulative reward you would have got, fine. So, now obviously you want to minimize this.

Now, let us start thinking about suppose for a particular context what is my regret? My this is our all possible context right for a particular context I am going to write it as and then c_t equals to this particular context and then I am going to write x_{ti}. Now I am going to write it inside this.

So, now I am only focusing on those instances where I have observed my context c. I have fixed my context c now. Now, I am seeing what is the regret I am going to incur for that particular context. The best I could have got for that context is this sum and minus this is the reward I would have accumulated whenever I saw that particular context.

Student: So $R_T(\pi)$ is the sum of all these term.

Yes. So, in that case what is $R_T(\pi)$ is going to be simply going to be the sum of all these term as long as I assume that my contexts are finite, ok. Now, let us focus on this how you will going to deal with each context? Yes there are multiple contexts, but if you focus only on particular context, you already in a known territory right, you are what is that? That is the standard adversarial bandit setting you already you know, ok.

So, do you like whenever now you can what you can do? For each particular context you can run an EXP3 algorithm. Yes like EXP3 algorithm was maybe like the that is the. So, the standard adversarial setting was a special case of this where it is assumed there is only one context.

Now, that we have multiple contexts, so I could run different EXP algorithms for each of these context. Anyway I know this context, I know how many contexts are there. For each context I would like to run a different EXP algorithm. Now, if I do that what is the regret

I am going to get for this particular context c? Can I write a bound on this? So, let us say this I am now want to write R_{TC} where I am going to use my policies EXP3.

What could be the bound I am going to get? So, my context particular C is appearing only at some points. It is not like it is appearing in every round, at some point it is appearing. Let us say out of 100 rounds I ran, only 10 times this particular context C appeared. Now, if I have that information can I write what is the regret bound I could get on this if I apply EXP3 algorithm? What is that?

So, what is the general. So, what is the for the standard setup, what is the regret bound we get by applying EXP3? So, we get something like a 2 times square root k t log k right. What was t there?

Student: Number of time of they have played.

Number of times you have played, but that t is the same t you are going to get here.

Student: No.

No, right. What is that t is going to be?

Student: Number of times the (Refer Time: 18:33).

Number of times you have observed the context, right. So, in that case can I write it as $R_{T_C}(EXP3) \leq 2\sqrt{k\log k\ \ (\sum_{\{t=1\}}^{T} 1_{\{C_t=c\}})}$; can I write it like this? So, what is this? This is basically telling you the number of times my context c has been observed right under which this is the bound, but notice that we got this bound in EXP3 by assuming that we knew the time horizon, right. So, in EXP3 we needed to tune a parameter η right that parameter η dependent on what?

Student: t.

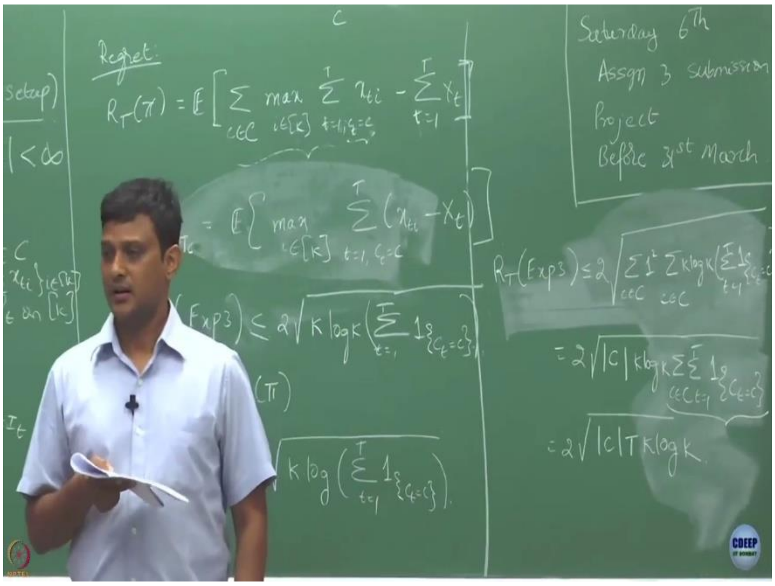
The time horizon, but in when I going to apply EXP3 on this setup, I apriori do not know how many times I am going to see this particular context c, right. It may appear on all the rounds or it may never appear or it may appear only on few of the instances. Now, how you are going to apply EXP3? Yeah.

So, you could use the case where we do not know the time horizon, then what we do? We are usually do the doubling trick, but with a doubling trick we know that we almost get the same regret bound with some constant factor loss in the regret bound, right. So, fine we can still do this, then I will just keep doing the doubling trick and with that I am going to get this as my regret bound.

So, then my R$_T$(π) is nothing, but summation $R_T(\pi) = \sum_{c \in C} R_{T_C}(\pi)$ and if I am going to use my EXP3 kind of algorithm. Let us say I am going to get is. So, we have this regret bound if I get it. So, now but somehow this upper bound here appears like it depends on what is the sequence of context I observed, right.

So, let us try to get rid of that what is the sequence of context I observed. Let us try to get an bound here which is independent of the sequence of context observed. So, can you apply a Cauchy Schwarz inequality on this and see what is the bound you are going to get? We have been doing this trick many times, right. So, treat this as. So, let us say there is a 1 here, this is 1 into some quantity treat it as a product of A$_i$ into B$_i$, where A$_i$ are all 1 B$_i$ are all changing for each context.

(Refer Slide Time: 22:55)


Regret:
$R_T(\pi) = \mathbb{E}\left[\sum_{c\in C} \max_{i\in[k]} \sum_{t=1, c_t=c}^{T} x_{ti} - \sum_{t=1}^{T} X_t\right]$
$= \mathbb{E}\left\{\max_{i\in[k]} \sum_{t=1, c_t=c} (x_{ti} - X_t)\right\}$
$(Exp3) \le 2\sqrt{k \log k \left(\sum_{t=1} 1_{\{c_t=c\}}\right)}$
$\sqrt{k \log\left(\sum_{t=1}^{T} 1_{\{c_t=c\}}\right)}$
$R_T(Exp3) \le 2\sqrt{\sum_{c\in C} 1^2 \sum_{c\in C} k \log k \left(\sum_{t=1} 1_{\{c_t=c\}}\right)}$
$= 2\sqrt{|C| k \log k \sum_{c\in C} \sum_{t=1} 1_{\{c_t=c\}}}$
$\le 2\sqrt{|C| T k \log k}$
Saturday 6th
Assgn 3 submission
Project
Before 31st March
CDEEP

So, now if you apply on this Cauchy Schwarz inequality, what is the bound you are going to get? So, keep the constant outside if I do that. So, I am going to get summation i is equals to summation $R_{T_C}(EXP3) \le 2\sqrt{\sum_{c\in C} 1^2\, k \log k \ \left(\sum_{\{t=1\}}^{T} 1_{\{C_t=c\}}\right)}$.

So, what I will do now is? So, this is cardinality of C (|C|), the first term is simply cardinality of C, ok. So, is this correct if I just apply Cauchy Schwarz inequality on this? So, now just what is this term is going to be?

This term is going to be simply T, right. So, this is, so what has just happened is if you are going to simply apply each context as a separate for each context you are going to think it as a separate adversarial bandit problem and apply maintain an EXP algorithm for each context, then this is the regret you are going to get and how this regret has scaled compared to the single contextual bandit? By $\sqrt{|C|}$ square root of cardinality of c. This much of.

So, as long as your contact size is finite that is cardinality is finite for C, you could a apply C EXP3 for maintaining it separately for each context and we are going to still get a sub linear regret like this. It is this regret is still sublinear, right.

Student: If this cardinality of C is around capital T, then?

But as long as fix it I mean this is about fix.

Student: Right.

You are not going to vary your variables here are T other quantities you have to fix whatever C you are going to choose and that is fix for cardinality C, then it is sub linear in t,

Student: Yes sir.

Ok fine. Then the question is what if this cardinality of C is large, then this regret bounds can be very bad right.