Bandit Algorithm (Online Machine Learning)

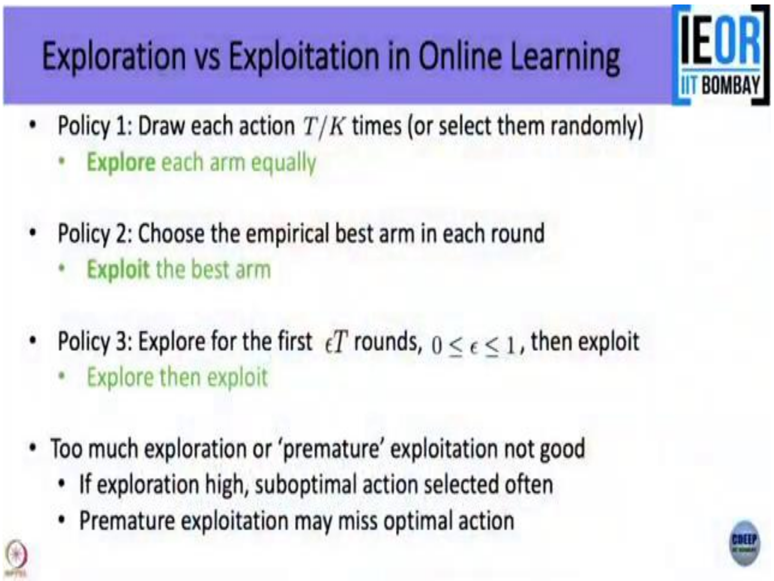
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture – 02
Introduction to Online Learning - II

You do not know you are not an oracle, you want to your goal has been set like you want to maximize the mean reward, how you are going to do? So, one thing is naive strategy, I do not care about what goal you set, I just do beating in the dark. I, there are I have been asked to play t rounds, what I will do is I will play each one of them equal proportion of the rounds. So, let say there are K actions or K arms and there are t rounds, you play each one of them equal number of time that is T/K number of rounds, you play one each one of them. Is that fine?

You can do this. But do you think this is going to be good policy? You just tried all of them. Or what you can do is instead of playing all of them equal number of times, you may just randomly uniformly select a arm in each round. Let say there are 10 arms, you give 1/10 probability to each one of them. And in each round you just pull one of this arm with this probability like this is uniform probability and just play that. Do you think this is going to be anything good?

(Refer Slide Time: 01:46)



Exploration vs Exploitation in Online Learning

IEOR
IIT BOMBAY

- Policy 1: Draw each action T/K times (or select them randomly)
  - Explore each arm equally
- Policy 2: Choose the empirical best arm in each round
  - Exploit the best arm
- Policy 3: Explore for the first ϵT rounds, $0 \leq \epsilon \leq 1$, then exploit
  - Explore then exploit
- Too much exploration or 'premature' exploitation not good
  - If exploration high, suboptimal action selected often
  - Premature exploitation may miss optimal action

So, in both of these cases, when you played each of the arm T/K number of rounds or played them according to the uniform sampling, you are basically not learning like you are not taking in to the feedback that you are going to see. You are just like whatever happens, I am just going to do this that is never going to get you to the optimal strategy, that is never going to improve. So, you are just basically ignoring what happens?

Student: What does it mean by empirical (Refer Time: 02:18)?

That is we are going to talk now. So, if you just ignore whatever you observed in the policy first like you just of played all of them equal number of times or just sampled them uniformly according to uniform distribution, you are not adapting or you are not even thinking of learning here. So, you are not going to do any better.

So, policy 2 what you can do better? I have listed already, but think through what you what do you think you could have done better. So, I am interacting with the environment. I apply; environment told whether that applied action was good or bad. I have this information. So, from this I can do any better. So, how many of you know what is empirical mean? So, you have observed how good a action is through your past actions, past observations.

So, let us focus on one. Let say you out of this 100 rounds let say you have out of this 100 you have applied one particular action let say 10 times and for this 10 times you would have observed its feedback. Can you take this 10 observations and find out what is the empirical mean of that action?

Student: Yes.

Can you?

Student: Yes (Refer Time: 03:48).

What I am doing, how I am computing the empirical mean? That is why let say Bernoulli case, and let say I am every time I am going to play this particular action, I am going to get independent observations. So, all of you know what is independent IID. So, how many of you do not know what is IID, all of you know it?

So if you can if let say these observations are independent let say and but they are Bernoulli like as I said like you can just count how many times I have observed one, take average and that is going to give me the empirical mean value.

So, let say out of this 100 rounds whatever I can based on my observation I can compute empirical mean for each of these arm. Once I have empirical mean, what I can do take the?

Student: Arm which have highest (Refer Time: 04:44).

Take the arm with the highest empirical mean. So, is that good? So, do you think anybody if sees that this is not really a good strategy?

Student: Better than (Refer Time: 04:58).

How many of you agree at least that it is better than the first policy?

Student: (Refer Time: 05:03).

Yeah.

Student: (Refer Time: 05:04) just one round how will we get the empirical mean (Refer Time: 05:08).

Not one round, I said, let say you have already, first round. First round, you do not have any sample, you cannot do anything. First let say you have K actions, you do not know like which one is better, because you do not have one sample. So, at least as a bare minimum let say we start with playing each one of them once. Now, I have got one, one sample for each one of them.

So, you lost K rounds in this already, because you have played each one of them. And in the (K+1)th, I can compute empirical mean because at least like I have one simple, but in the trivial case that one sample itself is the empirical mean.

Let say you played something and some particular action, you have played some particular number of times, and based on that n number you can compute the empirical mean. So, once I have this empirical mean you can order them and play which is the best one. So, again so all of many of you said it is better than first, but any of you feel this is not really better than first or at least in some cases first one have been better than second one?

Student: (Refer Time: 06:27).

Student: (Refer Time: 06:34) sampled each (Refer Time: 06:35).

No, goal is not I said the sum we have to maximize the reward.

Student: Second one.

So, anybody has any feeling that still first one is better when I wanted to maximize the sum of the rewards?

Student: (Refer Time: 06:50) Second order itself we are starting to pick up the best (Refer Time: 06:58), what we are doing or we are going to wait for some rounds (Refer Time: 07:01).

So, at least we are going to wait K rounds where we you are trying to get one sample from each one of them, first K round that is what like that. After that we just compute the empirical mean. Now, I has one sample for in each one of them, I will just compute the empirical mean; and from K+1 round onwards, I will pick the one which has the highest empirical mean.

And let say in K+1, I played some arm it gave me some reward, then I updated its empirical value, rest of them remain the same. And in the (K+2)th round, I again choose the one which has the highest empirical mean. So, fine like it looks like second policy is going to be better than the first, but what is bad about this?

Student: Like in cases where like in second policy if we have two rewards A, B, and one of them is somewhat good, but second one is very good. If we had chosen the action like randomly, then we would have been somewhere in the mid of mid value like p1, p2, we would have (p1+p2)/2 if we have been doing randomly (Refer Time: 08:18).

Student: At a point of time like initially p2 is bad, so we started picking p1 and we ended finally having p1 at the end. So, this one is, there might be cases where policy 1 performs better than (Refer Time: 08:30).

But here, fine like let say let us for just for understanding purpose let say there are only two actions. And let say action 1 has higher mean than the second action. So, in this case, action 1 is optimal, but I just do not know this, I just know that there are two actions.

So, can you imagine a scenario in which second policy is going to be bad in this case? What first policy said? First policy said like if I set T is equals to 100, it said that 50 times of the time play first action and 50 time of the time you said second action. But what is the second policy is saying? It is saying each time you play just in first two round pick one of them; and after that pick the one which has the highest empirical mean that time.

Student: p 1 for the first time.

Yeah.

Student: Then it will be never (Refer Time: 09:44).

Yes.

Student: Because 1, 0 whatever the value (Refer Time: 09:48) 0 will be always greater than 0 (Refer Time: 09:50).

So, let say in the first two samples K=2, first round I played arm 1; in that it so happened that I got a 0 sample in that. And in the second round, I played arm action 2, and I happened to get sample 1 there. I could still get sample 1 from second arm, it is not necessary that I have to get 0 only even though it is smaller mean it may still through up 1. So, what happens in the third round?

Student: (Refer Time: 10:19) arm 2.

Arm 2 you are going to pick?

Student: Forever.

So, will you be stuck in a bad arm in this case?

Student: Yes.

Right, because whatever you are going to observed from arm 2, its empirical mean is going to be larger than 0.

Student: Yes.

Because it has at least already one the observation one included there. So, you could end up stuck in a bad arm in policy 2. So, how to avoid this trap, getting stuck in a bad action?

Student: (Refer Time: 11:00).

Choose.

Student: (Refer Time: 11:02)

So, he suggesting that do not take just one, one sample and then start doing empirical. So, focus initially dedicate some rounds to get the samples, and then take the empirical means. So, in this case the probability that you get trapped in this bad scenario is going to be small. Is that all of you agree with that? We may still end up getting trapped in a bad value, but it is slightly better than policy 2 that scenario arising is slightly better than policy 2.

So, how to go about this? What we are saying is so one way of characterizing what he said is let us choose an ϵ which is between 0, 1. So, for the initial $\epsilon*T$ rounds, that is a some fraction of my overall T rounds, let us do collect samples from all of them.

So, collecting samples from all of them is going we are going to call this as simply exploration. So, I have collected samples from all of them, it is up to you. You can either go and do a random selection in the initial $\epsilon*T$ rounds or you do $(\epsilon*T)/K$ rounds per each of the arms collected.

And then after you doing that, in the after $\epsilon*T$ rounds, you go and select the one which is empirically best. So, when I say go and select the one which is empirically best that is what I call it as exploitation. So, what are the current situation I have, I am from that I am just going to try and choose the best one. So, do you think this should work better than policy 1 and policy 2?

Student: (Refer Time: 13:01).

Depends, then depends on what?

Student: (Refer Time: 13:05).

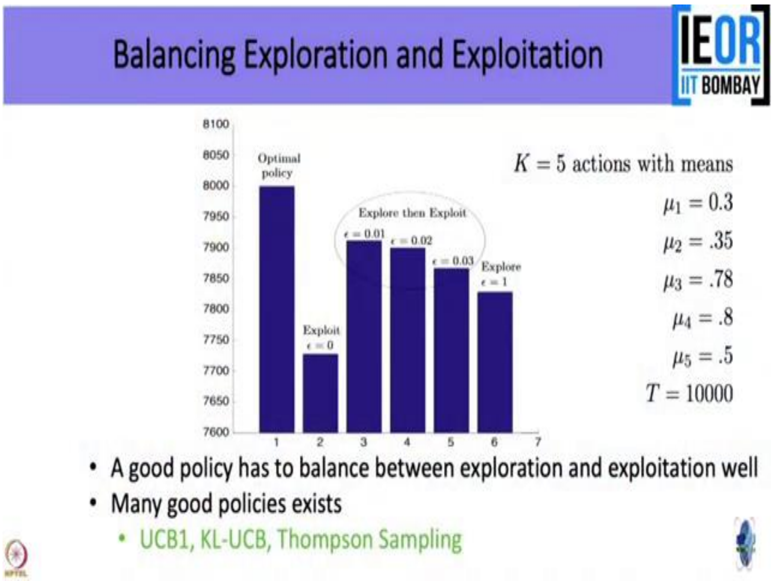
Where is c here? So, then now your problem one headache has transferred to another headache. Now, how to choose this ϵ? So you can choose. So, now, you have a good strategy choose this ϵ. So, what happens if ϵ is 0, and the ϵ is 0, but I still collect one sample from each one of them, then this is policy 2. No, sorry, if ϵ 0 basically I am not doing any exploration I am just (Refer Time: 13:38) exploitation, so it is policy 2. So, if what if ϵ equals to 1.

Student: Policy 1.

It is going to be policy 1, I am exploring like I am just may be selecting all of them uniformly random or just like taking playing each one of them equal number of times. But then here a problem is like ϵ how to choose and it has to between 0, 1. So, if ϵ is going to be smaller value, then you are exploring less, and then you are just jumping for exploitation not good. And if ϵ is 1, you are exploring too much and exploiting less, so that is also not good. So, you have to hit a balance between that.

So, too much exploration or even too much exploitation is not good. If exploration is high what is going to be happen, if exploration is high at the end of this exploration, you might be you might start playing the good one, but for a lot number of rounds you have played bad actions that is not good. And similarly if you start prematurely exploiting, you would have not collected enough information, so you may start exploiting bad arm, and that is also not good. So, you have to find balance between these two.

(Refer Slide Time: 15:05)



Balancing Exploration and Exploitation

$K = 5$ actions with means

$\mu_1 = 0.3$

$\mu_2 = .35$

$\mu_3 = .78$

$\mu_4 = .8$

$\mu_5 = .5$

$T = 10000$

- A good policy has to balance between exploration and exploitation well
- Many good policies exists
  - UCB1, KL-UCB, Thompson Sampling

So, this is just one example I have here just to demonstrate this. So, you see that in all this when you are trying to make a decision in this uncertain environment, this natural phenomena of how much to explore and how much to exploits comes into picture. So, the whole crucks of this course will be about understanding how much to explore and how much to exploit, or is there a way like I can do this together – exploration and exploitation.

So, just let say and this is some quick experiment here let say I have set 5 arms here with the means like this 0.3, 0.35, 0.78, 0.8, 0.5. Which is the best arm here? So, action arm 4 is the best one. So, if you know this, you would always like to play this one, but you do not know this, but let say I am going to I have been given, what is this is?

Student: (Refer Time: 16:03).

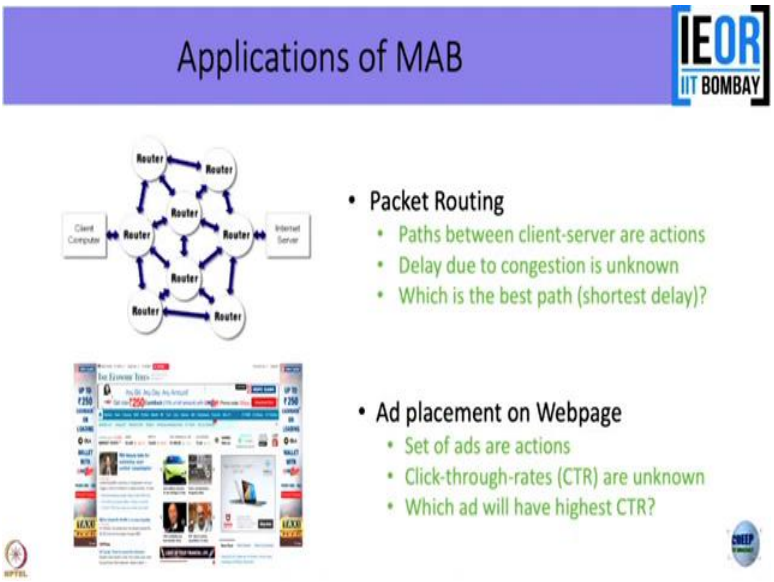
Some 10000 number of rounds. Suppose, in the previous policy I have discussed let say I set ϵ equals to 0, that means, I am just going to do exploitation this is the mean reward I am going to get 10 round 10000 rounds. So, had I always pulled this 0.8 this is the total mean I expect like almost like out of 10000 rounds 0.8 of that like around 8000 if what I expect the mean reward to be if I have played always action 4.

But let say I am started playing only exploitation this much I get; but as I increase ϵ to 0.01, little bit of exploration my mean reward actually increase this much.

But if I increase further this ϵ like did more exploration, my total mean actually decreased, and it decrease and I get the again bad. So, you see that if I increase ϵ 0 to 1, I am kind of getting a bell shaped curve here.

So, it says that there is a some ϵ that is good that is between 0 and 1, but I have to figure out what is that. So, a good policy has to balance between this exploration and exploitation. So, throughout our course that will be our focus how to do this. So, there are some good policies that we are going to revisit in this course.

(Refer Slide Time: 17:26)



Applications of MAB

- Packet Routing
  - Paths between client-server are actions
  - Delay due to congestion is unknown
  - Which is the best path (shortest delay)?
- Ad placement on Webpage
  - Set of ads are actions
  - Click-through-rates (CTR) are unknown
  - Which ad will have highest CTR?

A quick applications we just discussed some, but some what are the application from other filed, it is not just about recommendation systems. There are this multi arm bandits has applications in many things. The one I did not put it here the one that is of very importance in the medical field and that is why some of the studies of this multi arm bandits started.

So, suppose let say you have a drug, you want to identify a drug which is most effective on the population. So, when a initially drug is prepared, nobody knows like how effective it is right, maybe it is it has to be lab tested. So, what how usually we do, we like either we do testing on some rats or something or maybe like if you are brave enough, you may directly go and try on humans.

But whatever it is we want to make sure that the damages is the minimums. So, we want to initially we do not know which is the best, so initially may be we want to keep trying this different one, but obviously because the penalty is too much here it is life involved here what we want to do is quickly identify the drug which is the most effective. So, what do you feel like this, does this model fits here?

Student: Yes.

So, here think of patients as the environment like how so that are coming in each round, and environment is like we do not know the effectiveness of this drugs, and your actions are number of drugs you have or number of treatments you have. You apply on them.

You see how effective that treatment was like may be when you applied that drug that the that the patient got cured immediately and he started jumping. So, you are very happy about that drug.

But it may happen that the same drug if you applying the something that guy immediately died. So, it may not this as effects effectiveness may not be the same, but what we are interested is the one which is most effective on this population that is the best we can guarantee, how to figure it out.

So, but what we from the modeling point of all with this thinks that at least in the beginning, we will assume that that effectiveness is going to be the same on each one of you. So, let say in the beginning, I will assume that all of you are at the same capability.

It may be possible that all of may have different, different capability, and based on that the action I have to or attention I have to give is going to be the same for each one of you but we will not make. Both are possible and at least in the drug example it is assumed that all the patients have the same effective on the drugs.

We will study actually both the models what you are saying. So, the effectiveness of a drug on a person need not be the same, it could be depend, it could depend on that person in that way environment is not the same.

So, there are many other possibilities like in the if you are from communication (Refer Time: 20:44), it could be like packet routing. So, and also like not just packet routing it is like about you want to find the best path from a source to destination. So, let say you are at your home and you are going to office some day, and you have some 3-4 route options available.

So, the time it takes you to traverse on this rout depends on multiple things the congestion, weather, environment whatever other things may. So, you have to figure out which is the best route for you, best here is the one which may gives you the smallest travel time.

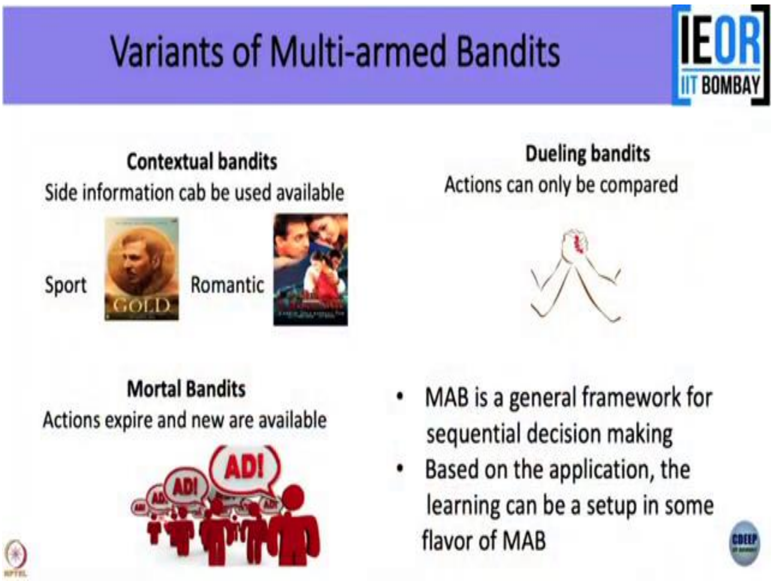
So, can you think of modeling this as this multi arm bandit problem? So, what are the actions here? So, the number of routes available to you. And who is the learner here? Learner is the you like who is taking paths to reach your office. So, first thing and I say like this one ad placement like when we visit this webpages we will say that we are bombarded with so many advertisements.

So, there is a whole big business behind this. So, people would each slot there in the advertisement, people want to put an ad in which you are most interested and most likely to click. So, if you click on that somebody is going to make money out of that. So, they would like to identifying the given slot which is the most appropriate ad to put. Can you think of it has a multi arm bandit problem something like?

Let say I have slot and there are some 10 ads that are available to put in the slot, and I have to decide which is the advertisement I am going to put in this slot so that it gets the most clicks. So, then what are the actions here? There is only one slot one advertisement slot, I mean there are many, but let us think consider only one slot in the top. So, just like here the advertisements are the number of arms, you want to decide which is the one you want to put there that has the highest number of click probability.

Student: What about the case where we (Refer Time: 23:04).

(Refer Slide Time: 23:05)



Variants of Multi-armed Bandits

Contextual bandits

Side information cab be used available

Sport Romantic

Dueling bandits

Actions can only be compared

Mortal Bandits

Actions expire and new are available

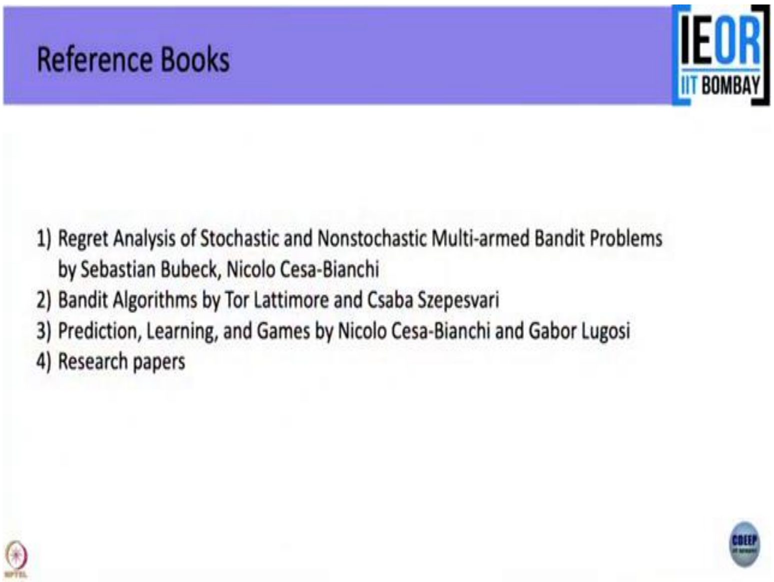
- MAB is a general framework for sequential decision making
- Based on the application, the learning can be a setup in some flavor of MAB

I mean you each one you can think of modeling differently or maybe you can come up with some variants of this, so that like if you are just not one location you are interested in, but all the possibilities you can now think of the arms as all possible combinations, and then you have to expand your problem accordingly. So, this is bare minimum with discussed like there are many many variants that are possible.

As I said this the active research area, you will see some of them will cover basics, but as for the project I would ask you to explore different variants of this bandits that we are that have been studied like, some of the things could be like contextual bandits where you are not interested in showing one particular action, but the action is curated in the sense you want to play action that is the instant specific.

So, for example, when you put an ad and you know that the guy who is visiting that webpage is a very young person, may be you would like to put a some sport shoe or something related to sports rather than some putting some insurance related ad there. So, it depends on the ad you want to put there depends on what is the context here. So, and you are the context here. Like if websites know this is the young guy who is just below 20-22 years, so accordingly he would like to show ads to you.

(Refer Slide Time: 24:39)



Reference Books

1) Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems by Sebastian Bubeck, Nicolo Cesa-Bianchi
2) Bandit Algorithms by Tor Lattimore and Csaba Szepesvari
3) Prediction, Learning, and Games by Nicolo Cesa-Bianchi and Gabor Lugosi
4) Research papers

So, there are different versions of this. So, we are going to follow largely these three books and some other research material for this. These books are available online and you can download them.