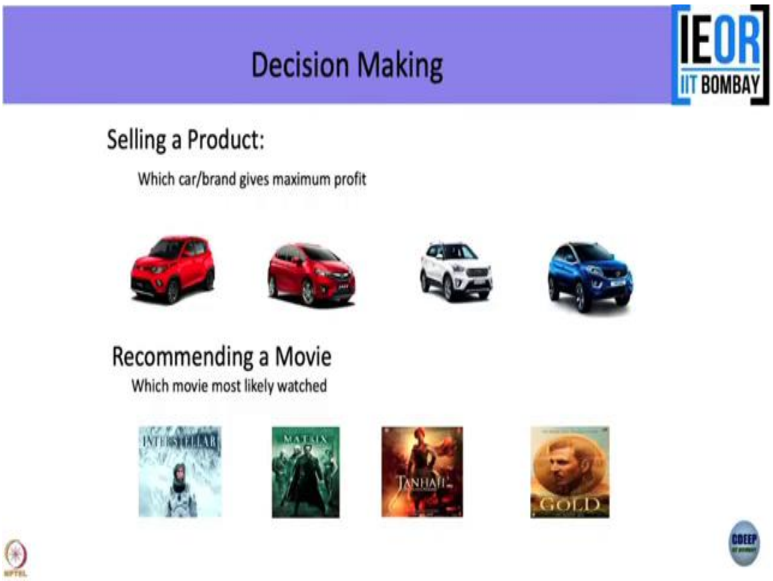
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture – 01
Introduction to Online Learning – I

So, what this course is all about? So, you might have heard about already so many machine learning courses that have been floated what is about this online machine learning right. So, just let us I have so today we will going to keep a short course, I mean a short class I have just about 10 slides I will just give a brief overview of what I mean by this course and give a flavor of what we what will be in this course and we will get into the main things from the next class ok.

(Refer Slide Time: 00:50)



Decision Making

Selling a Product:

Which car/brand gives maximum profit

Recommending a Movie

Which movie most likely watched

INTERSTELLAR
MATRIX
TANHAJI
GOLD

So, this is all about decision making ok. So, when you and you we guys do a lot of surfing over the internet this is just like a motivation, where this online aspects comes into this course. So, when let us say you are a seller of a car, let us say you are not from a particular brand let us say Tata or some Mahindra or something, you are like a you are some dealer or like you are you have a some webpage where you are trying to sell some car models or at least show people get interested in some car models.

So, what you would like you are basic goal in this will be like instead of randomly showing some car to somebody you may want to show it to somebody who will be more interested in this right or the car which is like most likely suited for that person or like may be like the one which that that guy who is looking at the who will potentially buy. So, if that guy clicks on this and buys it is a profit for you. So, many people who will be visiting your webpage and all what would be your goal, you want to basically may be show a car which will be bought in maximum numbers.

If that happens may be it is a great profit for you right. But you are in an online platform many people who are getting into this online platform you may not know anything about them. If you know already something about that person may be you would have made a proper recommendation, but you do not know how you are going to do in these things. So, you have to kind of figure out our learn what is that guy would be interested in that right. If you can somehow figure out may be you can make an appropriate recommendations.

So, here you see like there is already some kind of decision is making is happening. But it is happening in an online fashion like in the sense that something is coming you do a recommendation and may be that guy clicked or bought it or not you can kind of get a feedback from him, whether that guy was really interested in what you showed right. So, it is in a sense feedback to the action you took.

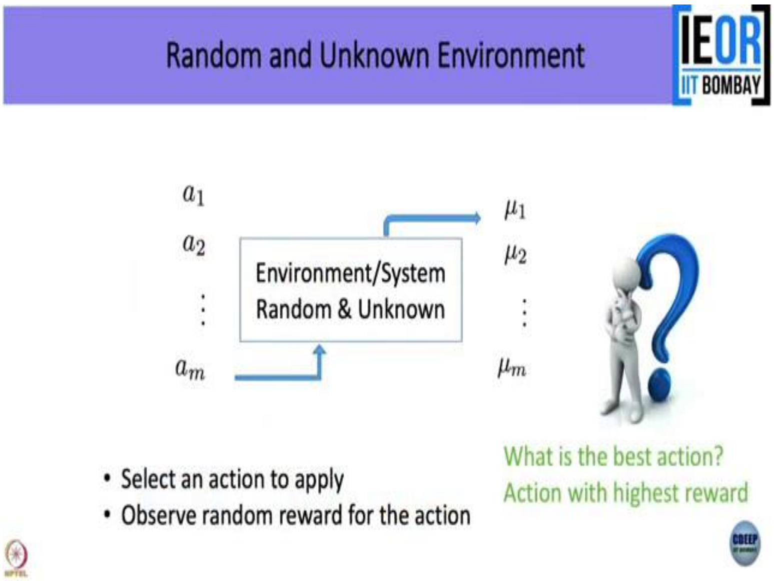
So, here action could be like this car models and you keep showing him, you get to know whether that guy is interested or not and based on this feedback you are going to get you may want to refine what is that you want to show ok. And you want to keep on this doing this iteration like you show some item that the guy who is on that platform at that time will click on that and you get a response and if he does not click on that then also you got a response, then you know that that guy is not interested in that.

Then next slot or may be the next person you do the same thing. So, you have kind of interacting, somebody comes you show him something he clicks you got some information if he does not click you still got some information. Using this using this information can you refine and try to show him the one which he is most likely to get interested or buy. So, if you could do that may be you have maximizing the number of units you are going to sell ok.

Similarly, you can think of many many other examples, may be like you might have been more familiar with this recommendation systems especially the movie recommendation system, like you might have seen like if you login to this hotstar Amazon prime or even what is that Netflix. It is fine Netflix first time when you have logged in first time you created an account Netflix kinds of asks you right tell me what are give me three examples that you would like.

So, from that it is starts try to infer you what could be a potential interest and then as you watch more and more movies it kinds of start figuring out what would be of your interest, may be you are a guy who is interested a more in scientific fiction or action or whatever some romantic movie or something. So, by the way this is all my old slides are used, the only thing I change in this is this third movie Tanhaji which one is latest otherwise these are all just old movie ok.

(Refer Slide Time: 05:08)



Random and Unknown Environment

- Select an action to apply
- Observe random reward for the action

What is the best action?
Action with highest reward

So, little bit abstracting all these things what is this all about, this online decision making is not just standard decision making, but you have to make a decision under uncertainity here. Because the environment under which you are taking decision you do not know a priori right. For example, when you are in a movie recommendation system let us say in the Netflix platform Netflix does not know like who you are already. So, it is a kind of you is like a random thing for Netflix, you do not know a priori what is your interest are. So, it has to figure out those interest and it is not like you right like Netflix or Amazon will have thousands and thousands of may be millions of users like you. And for all of them it may not know a priori what is their interest, but it would like to show a or recommend a movie that is most appropriate for that guy. So, if you could do that may be a Netflix or

Amazon keep you more engaged and you like to continue their subscription, if you like to continue their subscription you they will earn more money and this is how they want to sustain the business.

So, they basically want to figure out that is you for them is like a random thing and they want to make an appropriate decision in this random environment. So, here decision for them is showing you a right thing ok. So, we can in the simplest things we can now we want to a model and analyse these things how we go about this. So, as a learner you are one entity in this whole set up and there is an environment and now this is like a interaction between you as a learner and the environment.

For example think of in the Netflix example, Netflix is a kind of learner the platform whole is a learner and the users there they are like the environment there whose preference you may not know a priori right. Now, I am going to put in this box as that random environment which I a priori I do not know and now I have to take set of actions in this random environment and what is my interest here I want to take an action here which gives me maximum benefit in this random environment.

So, the movies I said let us think of them as a different action of the cars I said think of them as different action, here a 1 a 2 I am going to denote these are actions. Is the abstraction clear to you? What I am taking talking as a learner and what is the environment here. So, you as the learner you can decide which action to apply, let us say when you say a user or if you are taking making decision in a sequential fashion one by one.

You have this options like to take an action like let us say let us denote them like m actions denoted a_1 , $a_{2,}$ all the way to a_m. And let us say each one of this action when you play they will give an output which the value of which is stochastic in nature ok. So for example, so all of you know Bernoulli random variables right. So, Bernoulli random variable with a parameter p means what it is going to show me 1, the probability p and 0 with probability 1 - p.

So, p is the actual value the p is the mean there right. So, if between two Bernoulli random variables let us say one with parameter p1 and another with parameter p2 and let us say p1 is going to be larger than p2. Let us call p1 as process 1 and p2 as process 2, where p1 is larger than p2. So, in which process we are going to see more ones?

Student: (Refer Time: 9:19).

In the first process right because that has. So, anytime you are going to play any of these process you are going to observe a realization which is 0 or 1 which is governed by this underlying means p1 and p2 right. So, 1 and 0 these are like a stochastic observations associated with parameter p1 and p2. So, here we can think of that like these actions $a_{1,}$ a_2 .. a_m they will have their associated mean values, which are denoted as μ_1, μ_2 all the way up to μ_m ok.

These are the true mean that are that they are going to give under these environment. But you will not get to observe mu one right away like let us say when you observe when you apply this action a_1 the value generated would be some value which has the associated mean μ_1 and when you apply a_2 the value generated will be associated mean μ_2, is that part clear to you.

So, this these are all the like true mean values rewards associated with this like. So, we are talking it things in expectation right, it is not like I am interested in observing the expected behavior of the system not about the individual ones. So, when I apply action a_1, I will get a mean value of μ_1, but each time I apply it I only see the stochastic value some random value of this with mean μ_1, so like this.

So, I have m actions, there associated means or $\mu_1, \mu_2 \ldots, \mu_m$, we do not know a priori all these values. So, this value of μ_1, μ_2 depends on these random environment ok, which is unknown to you. So, that is what we I mean or like here the interaction is like you apply an action and observe the random reward associated with that action.

Now, what is the best action? So, in this case what you would like to do let us say you are doing this for hundred rounds ok. Like let me put it in this way let us say you have this m actions, their associated values are μ_1 all the way to μ_m and each round you are going to take one of these action and you observe the associated stochastic value. And let us say I say you are going to do this hundred times and what I will be interested in over this entire 100 rounds I want you to generate maximum reward.

What would you do in that case? Ok. Let us come to movie recommendation system ok. Let us say you have 10 movies to recommend and there are 100 peoples. So, when you show a movie to one of these. So, for time being let us say let us idealize this idealize the scenario, let us say all these 100 people have the same preference for these 10 movies.

I mean 10 movies it means all of them kind of have maintain the same preference list for these movies. So, let us say movie 1 may be liked most by all of them, movie 2 may be the next preference for all of them movie 3 is next preference for all of them, but you do not know their preference list ok. Now, in that case what is the best thing like if you want to make sure that maximum number of movies that you have shown are observed or watched by these people 100 people which is the movie you want to recommend to all of them.

Is that the case so let us let us say let us write now bit abstract a bit again. Let us say movie 1 movie 2 movie m the probability of that being observed are let us say μ_1, μ_2 all the way up to μ_m, these value I said you do not know μ_1, μ_2, μ_m ok. Let us presume that we know we know μ_1, μ_2, μ_mvalues I know and when I know them let say I figured out that $\mu_1$is the has the highest value.

That is movie 1 has the highest probability of being watched, here is the set up clear. Now, you are going to you have been asked 100 people are going to come you have to show movie recommend a movie to them to each one of them which is the movie you are going to recommend?

Student: (Refer Time: 14:16).

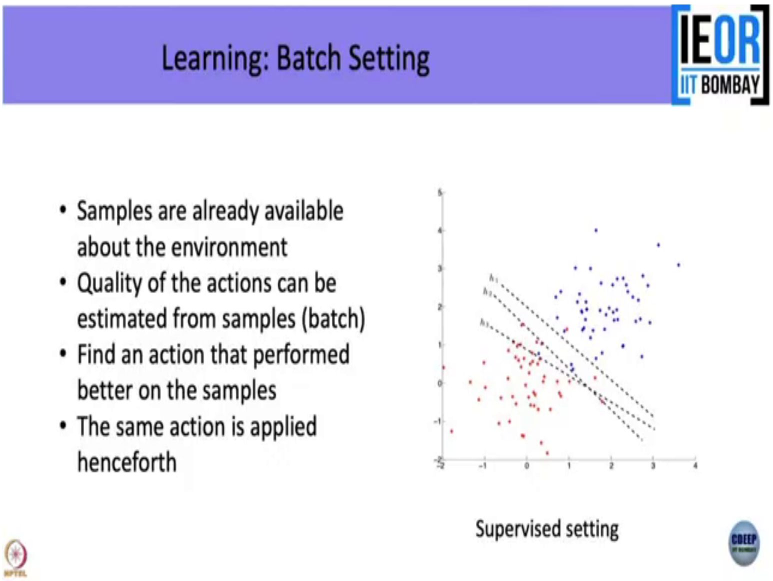
You are going to recommend one right because that has the most likely you would have being watched. And if guy watches that movie, that means let say a reward 1 for you and that he do not watches that thus reward 0 for you. In that case naturally your preference would be to show recommend movie 1 right ok. Now, this is the what I called as ideal case where you know the environment. But I am saying that possibly you do not know the environment right a priori you may not know this. So, then what you are going to do? What is the what you think you should be doing then?

Student: (Refer Time: 15:00).

Yeah first I mean whatever the way of I am going to do is, but first thing I have to do is I have to figure out these values. Once I can figure out then only I can see which is the best one.

How I am going to figure out may be you do sampling or just what whatever way you want to sample or just like do some magic and try to get this number whatever way you are going to do ok. So, this is where the online part of this machine learning comes in, you have this environment you have this set of actions you are going to interact learner is going to interact with the this environment and from that you want to figure out which is the best action ok.

(Refer Slide Time: 15:46)



Learning: Batch Setting

- Samples are already available about the environment
- Quality of the actions can be estimated from samples (batch)
- Find an action that performed better on the samples
- The same action is applied henceforth

Supervised setting

Now, how does this contradicts with the standard machine learning? So, how many of you done any machine learning course here? Many of you. So, can you one of can any of you contradicts this what is this way of machine learning is different from what you have learnt?

Student: (Refer Time: 16:03).

Decision points can change with data right. Anything else?

Student: Normally machine learning traditional machine learning you have all the possible data with you.

Yeah.

Student: Here you have to somehow figure out sample the data or work anywhere.

Let us revisit what is the classical unsupervised learning method you guys might have learned ok. What is the classical supervised learning method? You have a bunch of hypothesis right and you have a bunch of data ok. Now, you want to figure out a hypothesis which does the best classification on your test point ok.

So, how you are going to do that? So, best point is going to come from the same distribution or as your batch data, right. So, let us say somebody has already collected you this data points from a distribution given to you and also given your bunch of hypothesis, they ask you do wherever this and from thus give me a hypothesis which does good job of classification when I generate a new sample from the same distribution.

So, how do you do in that? You basically take away a bunch of hypothesis train it on your data points, what we mean by training. So, training could be anything like you want to basically find a hypothesis which does the best job of classification on your data points already collected right. Now, on this so here you can think of in this has your hypothesis are the actions.

Does this analogy make sense to you? Ok. I have a set of hypothesis I have my bunch of data that are given to me. I would say that so what I have set my goal as. So, you hypothesis are there to you give me a hypothesis from that which does a good classification job on a new data point. That means, so here the environment is what the environment is something which is generating this data point right.

So, what we usually doing this? We have bunch of data points, so what are this data points you can think of data points are something the feature vector and associated label rights in the classical terminology. You want to find the hypothesis, when I give my feature vector as input it gives me a label which is the value which is associated with that feature vector.

So, in this example let us say you have this simple case, where all these blue and red points are your data points and you have some bunch of hypothesis. Let us say h1, h2 and h3. So, h1, h2, h3 are like let us say your actions and I want you to find a hypothesis which classifies them well. So, which one is the hypothesis does that the good job here. So, which is a well separating the blue point from the red points. Which one is that?

Student: (Refer Time: 19:5).

h2 or h1. So, which one has the smallest error here?

Student: h1.

Is which hypothesis here is putting all the blue points on one side and all the red points on the other side and which one is not mixing them at any point. So, what point is this?

Student: (Refer Time: 20:19).

Is this red here?

Student: (Refer Time: 20:20).

Ok.

Student: There is no classifier.

Yeah.

Student: There is no true classifier.

The true classifier is not there if I mean each one of them as making some errors, but if your goal is to take the one which makes the smallest error, which one is good here?

Student: h2.

So, let us say h2 is the one which is making minimum number of classification errors right. So, on this data points the bunch of data point that have been given to you notice that let us say h2 is doing a good job. Now, if I give a new data points which is drawn from the same distribution as this points are drawn, which one you want to apply on that among this three.

Student: (Refer Time: 21:05).

You would like to apply h2 on that right, because it done a good job on this.

Student: Training data.

This is the training data, so in this classic method you have already bunch of data generated, you trained yourself and find the best action in this case the best hypothesis which does job you which is going to do a good job in the future points right. So, what you do usually you train on this and you keep on applying this whatever hypothesis did a good job on this henceforth. Whatever the new points I give you keep on using the same hypothesis ok.

Now, can you contrast this with what I call as online setup. Now, in the online setup let us say my actions are this hypothesis h1, h2, h3, but I have been not given this batch bunch of data a priori. What I will give you one at a time in incremental fashion, in round 1 I gave you one point you decide what hypothesis you want to apply. You applied and you got to know you did a good or bad job of classification on that, then I will give you second data point. So, here when I say I am giving you think me as an environment which is generating this data point and you are the learner who is using this hypothesis.

I mean who is who has access to this hypothesis as a learner your job is to find out which among these hypothesis is the good one. So, I give you the second data point you, you selected one of these hypothesis again applied on that point you got to know whether you did a good job or not. Again I generated a third one you give me you again applied once, so this kind of interaction goes on.

But here in the online setup I would like to figure out which one is the best hypothesis among this as quickly as possible. If I can do that, so over if I do if that means, I start I start applying the right hypothesis as quickly as possible right yeah.

Student: Sir in the setup we have one best (Refer Time: 23:24).

Yeah.

Student: (Refer Time: 23:25) common mode from all error suppose this is (Refer Time: 23:34).

Yeah.

Student: In case (Refer Time: 23:39).

So, you are saying instead of applying one hypothesis you want to apply multiple hypothesis and then ensemble their values here what I am saying is I am interested in applying one best action. So, here in that logic I am interested in finding which is the best hypothesis I should be applying.

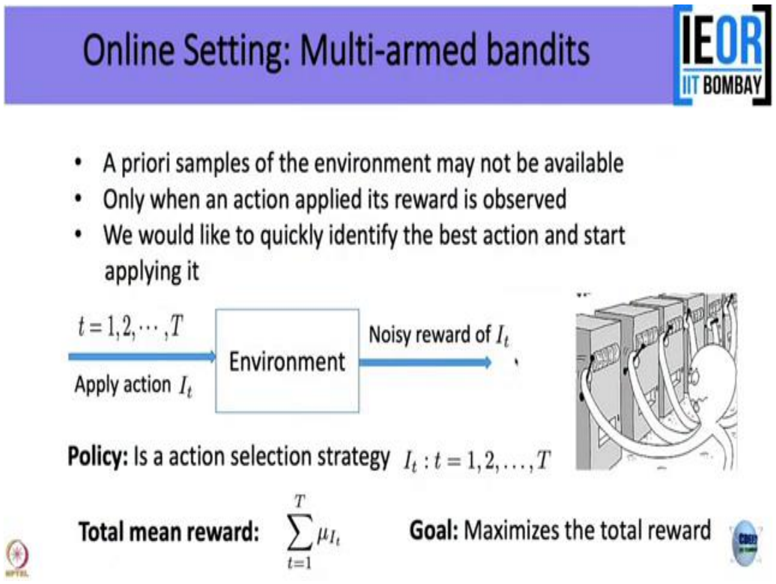
Student: (Refer Time: 23:55).

Yeah I mean this setup like if you are saying that which is the subset of this hypothesis I should select, so that what are the values they give that is like if I aggregate them that will gives me the good value right yeah. So, some way of aggregation them by either weighing them or just adding them whatever, that is fine that we can define as a we can setup the same goal here right. We can say among this hypothesis which of them I should say and how I should weigh them, so that the classification is the best.

So, you so this is how this is what we are going to, see throughout this course what are the different criteria we can set and how we are going to evaluate against this criterias. I am just giving you one simple example here ok. So, if you are going to think of this as different actions this hypothesis as different action and if you are aim is to find which is the best one in terms of the classification error. How you are going to do that? So, environment generate in a points you are applying you got to know whether you did a good job or not. So, you quickly try to figure out which is the best one ok.

So, the difference here between the online setup what we are talking about and what is this batch setup is like in batch we have already bunch of data I will train on that and after that I will fix one hypothesis and keep applying that again and again. So, in the online points are coming to me in an incremental fashion and I keep on updating what is the best hypothesis for me ok.

(Refer Slide Time: 25:31)



Online Setting: Multi-armed bandits

- A priori samples of the environment may not be available
- Only when an action applied its reward is observed
- We would like to quickly identify the best action and start applying it

$t = 1, 2, \cdots, T$

Apply action I_t

Environment

Noisy reward of I_t

Policy: Is a action selection strategy $I_t : t = 1, 2, \ldots, T$

Total mean reward: $\sum_{t=1}^{T} \mu_{I_t}$ **Goal:** Maximizes the total reward

So, in this course we are going to talk lot about this one particular online setting called multi armed bandits. So, there are different flavors of this which we are going to visit throughout this course. So, what happens here as I said unlike in the batch setting we will not have a set of data points already available and when you when you take apply a particular action you will get to see its reward.

For example in the hypothesis class, example we previously discussed when you pick up particular hypothesis and apply it on the sample you get to know whether that guy did a good job or not. So, that is what I mean only when an action is applied you get to know whether that guy is did a good job or not. But here we would like to like quickly identify. So, I mean you can you are aim less one error point comes you can apply whatever hypothesis come is of that strikes to that time. But our goal is to what our goal is to identify the best action right.

So, that something we are we are trying to optimize here. So, broadly this is kind of setup we are going to revisit in this course, as a learner you are on this left side. So, you are going to take an action in each round, if I am going to denote that as I_t as a action you are going to take in round t. When you applied this action on this environment, environment gives you some noisy reward associated with that action ok.

So, suppose let us say your actions are not noisy, they just like everything is deterministic ok. So, then what happens? You apply an action you got to know it is reward then nothing noise about it, that is the reward you are going to get each time that you apply action. So, in that case how you are going to determine what is the optimal action.

Student: (Refer Time: 27:40) take every action and (Refer Time: 27:42),

Yeah, I will just give each action one time you got to know their values, at the end you know these are the values just choose the best one, but things are noisy here right. So, this is how we should be modeling it also right, it is not like somebody liked when you are in Netflix you showed some movie, somebody clicked on it that does not mean that everybody is going to click on that right.

So, everybody has preference for it with some bias on this and that bias is what you want to figure out. And what you want to do in this also this is the classical image you see like when people talk about this bandit problem. So, this is like some bandit trying to pull this slot machine arms ok, this is I do not know like this as stuck to this bandit literature like somehow this name comes from this. Like you if you visit a casino there are different slot machines.

You put a money on that on the slot machine you choose you are going to win or lose on that, based on that you got some information about that slot machines. Now, you have let us say a few slot machines your goal is to identify which slot machine gives you the best winning probability, ok. Now, the goal in this thing is you are interacting with the environment over t rounds and I am interacting with the environment that is unknown to me.

But somehow you still want to make sure that over your t rounds, the total mean reward you are going to accumulate is the maximum. Does it make sense to like that is all like suppose as I said when you are in a Netflix platform, platform do not know you, but it want to show you a movies that. So, that if it shows the 100 people over this 100 people it wants the maximum number of people watched it. So, that is how I that is why I am interested in maximize in this angle, you see that you will have different notions of what we that what is that we are interested to maximize.

So, this is so this is the simplest thing you just over the t rounds you are doing something and you want to maximize that. When you could have maximize this if you know already the best arm you could have just played that again and again you would have done that. But I do not know that right, so how you are going to set your goal.

All I want to do is if you are going to running toward t rounds, whatever I am going to get in that I want it to be maximized and this is going to be the largest, if you have been playing the best term in every round. So, you see that but this number of rounds may not enough for you to figure out that. So, you will just set like I want to maximize my reward over this t period ok.

So, anybody has concerns or doubt like why maximize I am why you are saying maximize the total reward here or like say our goal we said identify the max best action right. But now I am rephrasing it as maximize the mean reward are these two goals aligned with each other or they are not at all aligned with each other.

Student: Sir even we take an or is an out that. So, that true goal which we had that we were identify the best arm.

Yeah.

Student: If it is possible we can run it forever and find out the best eventually.

Yeah.

Student: But that would take us long time. So, we are at sometime putting some limitation that they want to identify the best arm as soon as possible.

Right.

Student: So, that later on we will use that arm itself and enjoy the reward (Refer Time: 31:26).

So let us say this is the It is what It I have said that μ_{I_t} the subscript It there, I said this is the action you played in round It right. So, if μ_{I_t} are the mean values. So, when is this mean reward maximized; when is this mean reward maximized?

Student: (Refer Time: 31:47) when It is (Refer Time: 31:50).

Yeah so when I t is the best action right. So, at least this is in the same direction that I want to go. So, if I am going to pull the best action in each round I am going to maximize this, but I do not know that right. So, my goal would be to then I do not know what is that I am going to still set this as my goal and we will see that how best we can or how close we can come to our goal yeah. So, bandit that guy is a learner ok. So, he has to deal with some k machines. So, these are those are his actions ok.

Student: that is what the parameter that they (Refer Time: 32:45).

So, right so each of these machines on which you are going to win is going to be different value ok. So, let us say this has some 5 machines like at least some 5 machines are captured in the picture right. Let us say these are 5 machines and on each one of them you have different probability of winning. So, what you are doing in each round? You are going to play one of this machine and you are going to see whether you are going to win or lose that ok. So now, what I am going to say is I want and this μ_{I_t} is the expected value of you winning on that action whatever you are you took, I want to maximize this.

So, is this part clear this mapping It is going to change in each round, right. So, if you so t equals to 1, 2 like let us say it running up to capital T. It is not necessary that you going to play same action in every time, you may figure out that on that particular machine I am losing again and again. So, you may ditch it and you may want to play something else. So, this It is changing.

But whatever you played I am saying that you got a reward of μ_{I_t} and over t period this is your total reward, I want to maximize this ok. So, if you are the oracle or let us say you are the one casino owner who actually designed how winner lose happens on this machine, you kind of know already which is the best slot to which machine to play.

But as a learner you do not know, so you are playing different different actions and somehow trying to make sure that the total reward you are going to get or t rounds is going to be maximized ok. Let us discuss this how to do possibly. So, why is this multi time coming here?

Student: To multiple actions to take.

So ok.

Student: Each round (Refer Time: 34:53).

You are doing one but your options are multiple ok. So, again multi arm right like the terminology arm is like an action here. So, this think of I mean from the picture may be you can associate like they have this levers right, like think of there is a arm or something in each round you have to pull one of these levers or like play one of these arms or like play one of these actions. So, we have multiple actions or arms to choose from this, so that is what like we are in the multi armed bandit setting here yeah.

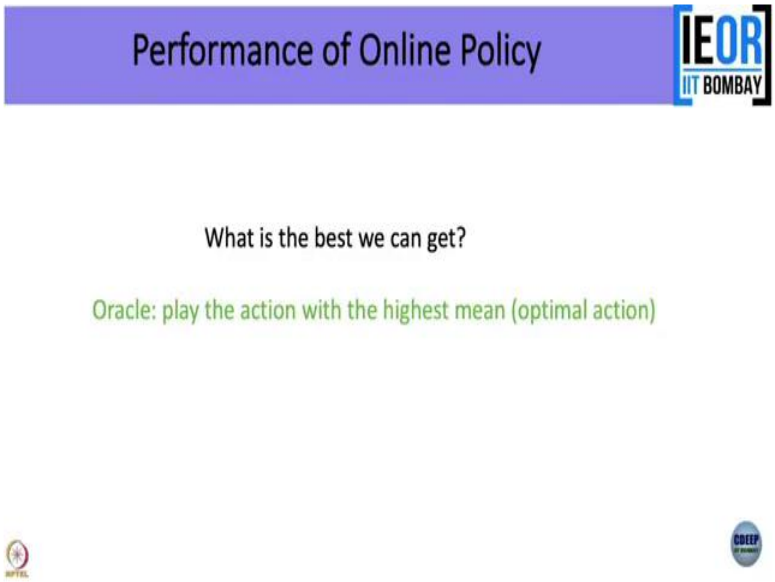
Student: So, what do we mean by (Refer Time: 35:30).

So, we said that right like let us say this casino devices you have to make them the winning probability bit. Let us say you are a casino owner you will if you have some 5 machines, are you going to design in such a way that some machines always win and some machines always lose you are not going to do that way right.

But you will set them with may be some stochastic probability on which their winning probability small, but whatever but you want to make money. So, you probably you will design it in such a way that any player using that he is going to be with a small probability. But let us say for you as a player here among these itself like whatever the probability there of winning you want to still go with the one which maximizes yours.

So, may be the winning probability on all of them need not be the same may be it is different, may be it depends on your talent how you are going to put up your strategy based on the winning probability could depend. Now, based on that you want to identify the one on which your winning probability is the highest. So, that is what stochastic I am talking about. So, when you play an arm here or action you may win or lose with the associated probability. So, win 1 0 that is the noisy observation of that parameter right ok.

(Refer Slide Time: 37:00)



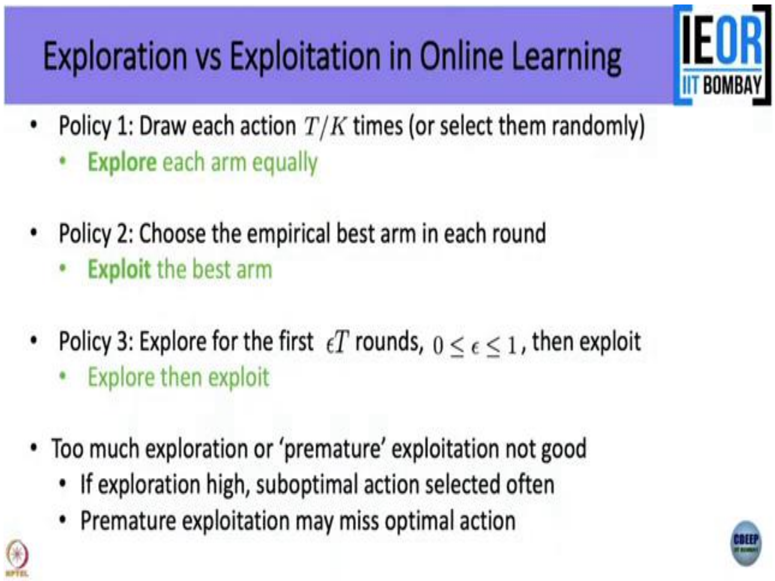
Performance of Online Policy

What is the best we can get?

Oracle: play the action with the highest mean (optimal action)

So, we said that like if we have been a oracle who knows everything about the winning probabilities, what would have basically done is you go and play the one with the highest mean, but you are not a oracle. So, you do not know this you have to learn it ok.

(Refer Slide Time: 37:18)



Exploration vs Exploitation in Online Learning

- 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

Now comes the question, if you are not a oracle you want to figure out which is the best action what are the options with you? So, how you will go about this? So, What could be policy. So, I am going to call this as a policy right, the action you are going to choose in each round it is like your policy, based on what the observation you have made.