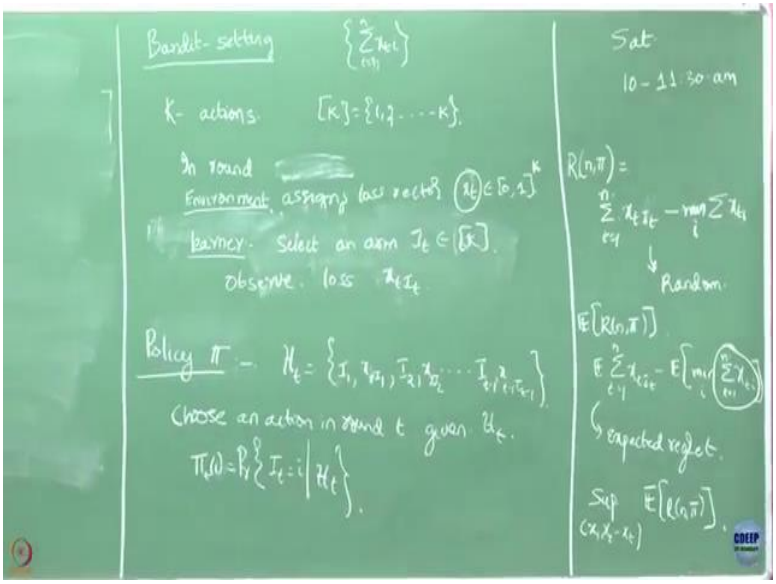
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture - 13
Adversarial Bandit Setting

So, far any doubts on this step here. If you have any doubts on this setup, just ask me now. So, that is it is assigning it could be occur arbitrarily right, you are just assuming that there is some phenomena according to which this losses are generated, which you do not know ok. Your goal is to now select an action here, which would have given the smallest loss.

(Refer Slide Time: 00:53)



So, let to just to rewind this. So, in every round environment is coming up with this loss vector right, at the end of n rounds. So, this is the loss you incur if you happened to play always the ith action right. We are just saying that ok, let I do not have any control over how this environment is generating the losses. But I want to play an action on which the loss assigned by the environment is the smallest.

So, this is the total loss assigned to action I by the environment right, I am just trying to take the smallest one. So, what we are saying in the setup is see I do not have any control over how this losses are being assigned to action. I would pick an action, which gives me the smallest total loss, that will I will take it as my benchmark.

Student: It will not be keep changing?

It can keep changing that is why we have indexing by subscript t right.

If t changes, this vector could change.

Student: But then sir, when we are writing summation xti.

Yeah.

Student: Then, we you are saying that we are setting the ith action on which the losses are minimum.

No, I am looking at you take one I, look at the loss I get on that action.

Student: In all the round?

In all the rounds summed over all the rounds and now, I am looking at an action on which this sum is the smallest.

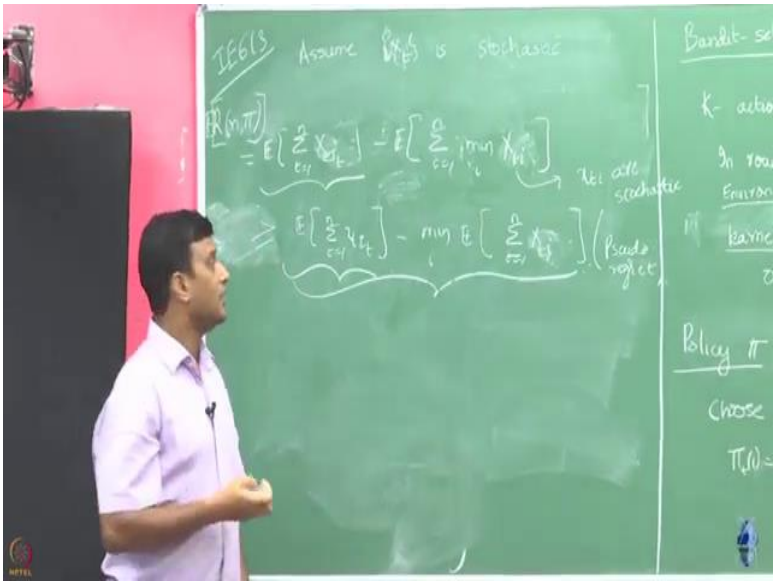
Yeah. So, maybe just to make this concrete let us say let us take you 3 guys sitting in this column like let us say in every class, I will come up with a some numbers for you guys, according to my own criteria. Assign this value, this value 2, 3 and I keep on doing it for let us say 10 days. So, the number assigned for you let say it is the score for you. Now, somebody who do not know how I am assigning scores to you, but he would like to pick one among you to whom possibly I am going to assign the highest score; totals in total sum.

If he has to figure out, how he is going to do that like the guy who is going to get the maximum score right like the total sum, he if you want to identify, he would be your that would be your best choice, that is the best student. Let us say among 3 of you, if I want to identify the best, this is what I want to do. But I do not know a priori right like how I am going to assign this and this assignment process is could be I, it is up to me how I assign and you do not any of you do not know about it. Now, how you are going to select that is what this setup is telling you. Minimum?

I am saying over the entire round I am basically taking about single best action for the entire round right. If you are allowing fine, if you want to take this min inside this summation, what

you are looking is, the smallest score in each round right. So, right now, I have not gone into that; but that situation is harder to handle; that means, you are you could have done that ok.

(Refer Slide Time: 04:54)



So, let me; so, what is saying is ok, let me write this. Why not consider this right? Now, let us just I want to again write that quantity here. You understand the difference between these two criteria's now? All of you? So, what we are doing is, when I am looking into this my criteria's, I am looking at the single best action throughout. But if you take the minimum inside, what I am looking at in each round, I am looking at the smallest value and because of that it is not necessarily that in each round, I would be playing the same guy. So, which criteria is good?

Student: First one.

First one is more stringent right. He want to see in each round, what is the smallest? So, if you could do that it may happen that in one particular round, I may assign him 3, I may assign him 2 and I may assign him 1. In that in this round, he got the smallest value; but in the next round, I may assign him 0, I may assign him 4 and I may assign him 2.

So, now, in the second round, this guy got the smallest value. So, in that way in this criteria is asking go and choose the best action in each round that is that is the benchmark. But often its so happens that this criteria is very hard to compete against; we are always, so this is what we are saying this is what we incur and this is what we are competing. This is our benchmark strategy right.

In this case, playing the single arm always is the benchmark and in this case, playing the best in each round is the benchmark. But this my benchmark always happens to be hard to deal with . Now, let us is that relation between this quantity and this quantity? Right now, we do not putting any constraint on how this xt_i's are generated right as a special case it could be cyclic as he is telling.

So, for example, let us say let us take K equals to 3 and in this case my xt first in x1 is (1 2 3) and x2 is (3 1 2) and x3 is what is that after one more cycle (2 3 1); you understand this? So, the environment is assigning this and then, again x4 becomes equals to x1, x5, it could be in this cyclical fashion. So, with this, if you are going to look at a single best.

Student: First find.

Yeah, so.

Student: (Refer Time: 08:27).

Fine, if you have such a specific sequences if you are saying ok, I am not learning anything by this one. In this case, this criteria is better ok. But as I am telling this is fine, this is when we have such a special sequence, but a priori you do not know anything about how the is being generated right. For a arbitrary sequence, giving any performance guarantees with such is very hard; whereas, if you look at comparing against a single best action we will be able to say something. This is indeed a string this is a stronger benchmark, but it also will is not much tractable, we cannot say much about this; but we will be able to say much about this ok. For time being assume that this xt's are randomly generated.

Student: (Refer Time: 09:41).

It could be some something, but some random values. Can I say something about this and this? Expectation of minimum of random variables and minimum of the expectation, which one is greater?

Student: Smaller guy.

This guy is going to be smaller, this guy right because you are taking expectation of the path wise minimum. This is going to be minimum expectation. So, because this is smaller, but this is coming with a negative sign ok. So, which is going to be larger? What will be the relation?

So, this is going to be like this right and what we have defined our regret to be actually our we have defined this to be our regret. This is the expected regret. What we ideally would have like to give a bound on this right, this is what here we are interested in.

But if this is going to be if I allow this xi to be random, then this relationship holds and we will see that later, when I allow this xt_i's to be random. I will only be interested in bounding this, not this; then and that time, we are going to call it as Pseudo regret. So, right as long as sequence is given to you, its fine if the sequence are random; then, comparing against a single best action will be lower bound on this, in this fashion ok.

So, this we will again revisit when we are going to talk about stochastic bandits, where we allow this loss to be drawn according to some distribution. Here we are saying this x_{ti}'s could be stochastic or it could be arbitrarily generated ok. I mean stochastic in the sense they follow some particular distributions or they did not follow any distributions ok. So, this we call as pseudo regret fine. Now, the question is what is the algorithm? Yeah?

Student: (Refer Time: 12:28).

This one here for a given sequence. This expectation has no meaning right. If you give me the sequence xt, this our fixed quantity; whereas, this quantity is random because your It is random. So, I mean I could as well not write expectation here ok. But now, what we are saying is if this x_{ti}'s are random, are stochastic; you are assuming that this is not a particular sequence, but this is some stochastically generated sequence. It by an according to some arbitrary distribution and not necessarily in IID fashion or anything. It could be in arbitrary fashion, then this expectation here is also valid because the sequence is random.

Now, in which case if I am going to look at the single best action as my benchmark, then this becomes a lower bound on this. Is this point clear? Maybe what this confusing here is this is given sequence.

(Refer Slide Time: 14:06)

So, in that case, maybe I do not need to write an expectation here; whereas, if I allow this x_t's to be a stochastic, then this holds. Is this clear? Then, in this case this expectation involves two level of randomness here right. When I write, when I allow this x_t's to be stochastic these are random and also, the selection of It is random. So, this expectation is over randomness of losses and arm selection.

So, the randomness of losses is by the adversary that is the environment and the arm selection is by the learner. So, you are basically taking average over the distributions of both environment as well as the learner; whereas, in this part, we have removed learner right. We have just trying to see what is the best thing we can get. We only take the expectation with respect to the randomness of this loss values ok.

Student: Is that x_{ti} (Refer Time: 15:40).

And this should be x_{ti} here.

Student: Quantity that environment is giving logic right.

Yeah.

Student: i is also random?

No.

Student: Which i we are taking?

Because we are taking minimizing it on all possible i's right.

Student: So, in one round, we was like 1i, in second round and other i.

It could be, depending on what is the value of this x_t ok. Again, let me revisit this things. We are saying the environment in each round is going to generate a loss vector and then, the learner is going to pick an action and the learner is going to pick an action according to policy, that policy will be based on the past history that he has observed ok. This is the observation the learner has made.

Now, this policy maybe I should also write it as a function of x_t because now, this regret is defined for a given sequence. Is that clear? And now, if for a given sequence, this is the loss you have incurred for playing It and this is what the best you could have got.

So, that is why I have said, may be again I will write may be, so this is the regret I got. But still this is even though this is for a given sequence, this regret is still random right. Because the learner is randomising his strategy ok. So, I would be I may wanted to look at the expected value of this, but still for this given sequence. So, that is exactly this value that is only this part is random. So, I am going to take the expectation of this; but this part there is no randomness. Once you give me the sequence, there is no randomness in this quantity and this is the expected regret.

So, because of this having expectation here has no meaning right because this is for a given sequence, I could get rid of this expectation as well. So, this things are clear here. All this confusion araised, when I wrote expectation here and the question was the whether this expectation and minimization were interchangeable here. Now, as you instead of xt sequence is given, this Xt sequence is stochastic ok. So, because stochastic I am going to write it as Xt because now the loss vector assigned in each round is a random quantity.

Now, instead of looking this regret for a given sequence xt, now I am allowing this xt's to be arbitrary, then what we will get this is going to be $E[R(n,\pi)]$ and notice that, I am now no more writing it as xt because that sequence is I am not considering a fix sequence, but I am considering stochastic sequence now. Yeah?

Ok, it is the same thing what I have written here; but instead of a fix sequence, I am allowing this taking this xt's to be stochastic. So, fine

Fine, I am just saying this is what when from this when I went here, when I allowed my Xt's to be random, this is the quantity right finally. So, this is expectation. I have also expectation of this because my Xt's are random, but here this sequence Xt is.

Student: Yeah, this is fine.

Yeah.

Student: So, if we are go over there.

Yeah.

Student: So, if they are just X_{tlt} are stochastic.

Yeah.

Student: So, on the basis remain the same?

Here it is remained same.

Student: No, second one?

In the second same. So, I am just saying this happens because it is not the same, they are not equal.

Student: No, they are not equal.

What is the relation?

Student: (Refer Time: 21:32) stochastic.

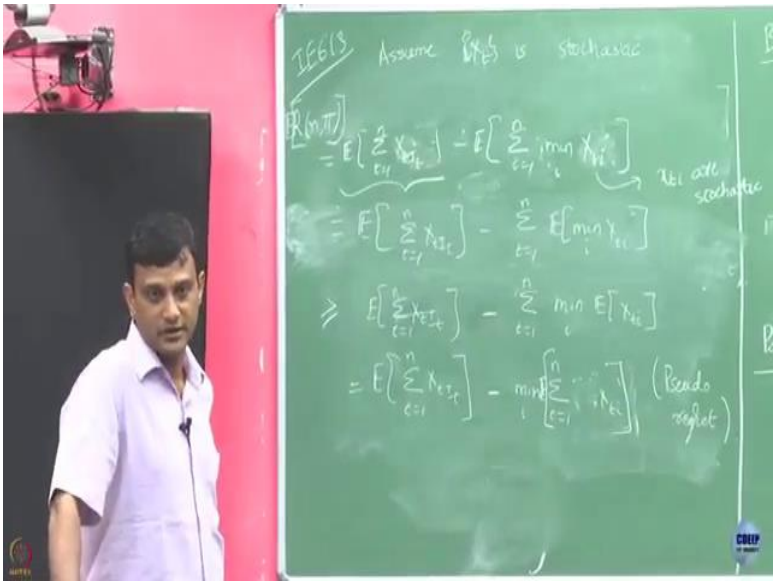
Here it is stochastic.

Student: Fine. Here it is stochastic, there will be (Refer Time: 21:36).

Right, because of this you have expectation of this minimum right. Now, if you are going to. So, what is this I could write this as this quantity here as also expectation of sorry summation of expectation of minimum, if I take expectation inside. Now, expectation of minimum of xi,

it is going to be greater than this ok. So, maybe you guys were still confused. Let me rewrite this.

(Refer Slide Time: 22:17)



Let me rewrite this expectation of t equals to 1 to n X_{tIt} minus t equals to 1 to n expectation of minimum of i Xti. This is fine. Now, I will just leave it like this. Now, minus this quantity instead of expectation of minimum of this, if I write it as minimum of i expectation of Xti what it will be?

Student: Greater than 0.

This is going to be ok. So, this is still. Now, can I go back to this step, this step is what? I mean this I can always re write it as minimum of i.

Student: Bracket t.

So, we have just doing this manipulation here. Now, so this is equal. Now, this was looking at the minimum quantity in every rounds, but now what it is looking at? It is looking at the single action which is giving me minimum over the all the rounds. So, and this is going to be actually lower bound on this ok.

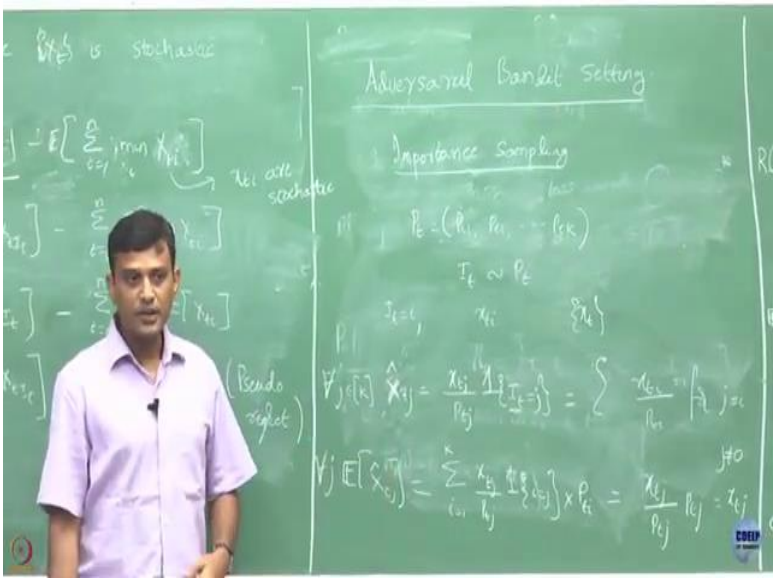
What we actually wanted is we defined this to be our regret ok; but this regret is often too demanding, the benchmark here is to be demanding. In each round, I want to find the minimum loss; but here if I do some manipulation, I will end up with this. Here my benchmark is to look

at the single best action for the entire n rounds. But this is a lower bound is fine, but we will we will take this as our regret definition.

And we call this as I wrote earlier pseudo regret. Yeah, we call it pseudo regret. I mean say like what we just the whole business of getting confused here happened because we first defined this regret for a given sequence and then, we wanted to take the expected value. So, this expected regret is solved like this. But had we allowed instead of a sequence particular sequence if they are allowed for any probabilistic sequence, we I will end up with this definition. But this definition requires a this stringent benchmark instead of that we can go and consider this benchmark which is like a weaker question of this ok..

I would have like to come to this ideally bit later, but now you have ask this we kind of clarified this. Is there any confusion now on what is an expected regret and what is a pseudo regret here? That is fine right. In the expected regret, we have this benchmark which is bit stringent ok.

(Refer Slide Time: 26:40)



Now, let us come back to by the way we should call this adversarial bandit setting, not just bandit setting. So, we will just discuss the main idea that we are going to use in this setup and then, talk the algorithm and its proof in the next class. Ok fine. In the weight majority, we got to know the loss for all the actions and what we did? We updated the weights for each of the actions right, according to some exponential factors.

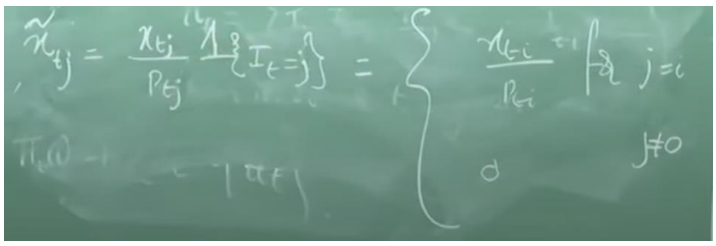
So, they we were able to update the weights for all the actions in each round because we have observation for all the actions. But now in bandit setting, we have observation for only the action we played; but not for the other actions. Now, how I am going to update? Is it that in each round, I am only going to update the action that I played and not update anything about the other actions or that is there a mechanism that I update all the weights in each round ok?

So, we will see that it even though we do not observe a losses for the actions which we do not play, but we can pretend to come up with a mechanism, where we will say that we have some information about the other arms even though we did not observe them and accordingly update their weights. So, what is that mechanism? That mechanism is called as Importance sampling.

Let us say, so Pt is the distribution with which you select the arms in each round. So, we said arms are selected according to some distribution right and the policy governs what is that distribution. Let us say accord in round t, this is the distribution according to which you are going to select an action.

Now, that is we are going to select It according to distribution Pt. So, we are going to select one of this. Suppose, let us say use happen to select It equals to i in round t and then, you observed xti as your loss for this action and you do not observe anything for the other actions ok. Now, how to update the weights for all the actions? Let us say, I am going to define this quantity called the estimates for the loss and going to define it maybe I should write. I am going to define the estimates in round t, the environment chooses this vector xt.

What I am going to do is, I am going to estimate the value that environment chooses in round t by some mechanism and I am writing that values to be in this fashion ok. So, this is my estimator for the loss observed on round on j in round t ok. So, how I am defining this? $\frac{x_{tj}}{P_{tj}} 1_{\{I_t=j\}}$. So, suppose now let us say if you have played It equals to i in the t'th round, for what is this quantity is going to be? So, let us say so this quantity, how this is going to be?



So, this is how I am estimating the values of the loss in round t. I am saying that if I observed, if I have played action i, I know the value xti because I observed that and I define its estimate to be xti by Pti. For the other guys, I just define it to be 0 ok. this is what this definition is saying that ok. Now, the question is and this is an estimator, estimator for loss values. The question is why this estimator? Ok. So, now, let us try to see a this estimator here is a random quantity?

Because it depends on It which is selected randomly in that round right ok. Now, let us take expected value of maybe because this is a random quantity maybe I will write it as Xt ok. So, this is an estimator which is, now this is what this quantity. What is the expected value of this? If you want to compute the expected value of this, this is going to be. So, this is the random quantity right. I am going to looking expectation with respect to It.

So, this is going to be $E\left[\widehat{x_{tJ}}\right] = \sum_{\{i=1\}}^{\{k\}} \frac{x_{tj}}{P_{tj}} 1_{\{i=j\}} * P_{t_i}$. Is this expected value correct? Now, simplify this. Only when i is equals to j, this indicator remains; for all others, it is going to be 0 right.

So, this estimator is such that the expected its expected value is exactly the sample value in that round, even though you may only I have observed xti in that round, but if you are going to define your estimators like this, even for every this is true for all j. So, for every components, you have an unbiased estimator. So, you understand what I mean by unbiased estimator? What is an unbiased estimator?

The expectation of the estimator is the same as the of what? The value which were estimating ok. So, here this guy what it was whatever its quantity this is if you are assuming that this quantity is trying to estimate the true value that is Xtj in round j, this is exactly its in expectation it is doing that job. So, this estimator here is that unbiased estimator.

Now, you see that if you forget about this estimator, in every round you kind of having values for each of you have predicted or predicted each component which are a good estimators for the true values, even though you do not know the true values, but you have a mechanism in which you are able to estimate those unknown quantities for which you are estimations are pretty much they are unbiased.

So, can you now use them in your leaning as you did in the full information case? So, in the full information case, you had information about you have observed the loss for all the actions.

So, you are using them to update all the actions. But here you only observed loss for one action; but you for everybody else, you have this estimator; but these estimators in expectation as good as the true values which you did not observe. So, may be yes, I do not have true observe, but I have this value observes which could act as a proxy for the true losses and I can use them to update all my weights, weights for all the actions right.

So, we will use this idea to come up with an algorithm for adversarial bandit algorithm called EXP 3 that we will do it in the next class.